

Motif formation and emergence of mesoscopic structure in complex networks

Jacopo Iacovacci

A thesis submitted to the University of London for the degree of
Doctor of Philosophy

School of Mathematical Sciences
Queen Mary, University of London
United Kingdom

March 2017

Al capitano che doppiò il Capo, all'alba, e al maestro della pietra.

6. Iacovacci J., Bianconi G., Extracting information from multiplex networks, *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26 (6), 065306 (2016).
7. Iacovacci J., Rahmede C., Arenas A., Bianconi G., Functional Multiplex PageRank, *EPL (Europhysics Letters)* 116(2), 28004 (2016).
8. Lacasa L, Iacovacci J., Visibility graphs of random scalar fields and spatial data, arXiv preprint arXiv:1702.07813 (2017).
9. Rahmede C, Iacovacci J, Arenas A, Bianconi G., Centralities of Nodes and Influences of Layers in Large Multiplex Network, arXiv preprint arXiv:1703.05833 (2017).

Abstract

Network structures can encode information from datasets that have a natural representation in terms of networks, for example datasets describing collaborations or social relations among individuals in science or society, as well as from data that can be mapped into graphs due to their intrinsic correlations, such as time series or images. Developing models and algorithms to characterise the structure of complex networks at the micro and mesoscale is thus of fundamental importance to extract relevant information from and to understand real world complex data and systems. In this thesis we will investigate how modularity, a mesoscopic feature observed almost universally in real world complex networks can emerge, and how this phenomenon is related to the appearance of a particular type of network motif, the triad. We will shed light on the role that motifs play in shaping the mesoscale structure of complex networks by considering two special classes of networks, multiplex networks, that describe complex systems where interactions of different nature are involved, and visibility graphs, a family of graphs that can be extracted from the time series of dynamical processes. This thesis is based on the research papers listed below, in particular on the first five, published between 2014 and 2016:

1. Bianconi, G., Darst R. K., Iacovacci J., Fortunato S., Triadic closure as a basic generating mechanism of communities in complex networks, *Phys. Rev. E* 90 (4), 042806 (2014).
2. Iacovacci J., Wu Z., Bianconi G., Mesoscopic structures reveal the network between the layers of multiplex data sets, *Phys. Rev. E.* 92 (4), 042806 (2015).
3. Battiston F., Iacovacci J., Nicosia V., Bianconi G., Latora V., Emergence of multiplex communities in collaboration networks, *PloS one* 11 (1), e0147451 (2016).
4. Iacovacci J., Lacasa L., Sequential visibility-graph motifs, *Phys. Rev. E.* 93 (4), 042309 (2016).
5. Iacovacci J., Lacasa L., Sequential motif profile of natural visibility-graphs, *Phys. Rev. E.* 94 (5), 052309 (2016).
6. Iacovacci J., Bianconi G., Extracting information from multiplex networks, *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26 (6), 065306 (2016).
7. Iacovacci J., Rahmede C., Arenas A., Bianconi G., Functional Multiplex PageRank, *EPL (Europhysics Letters)* 116(2), 28004 (2016).
8. Lacasa L, Iacovacci J., Visibility graphs of random scalar fields and spatial data, *arXiv preprint arXiv:1702.07813* (2017).
9. Rahmede C, Iacovacci J, Arenas A, Bianconi G., Centralities of Nodes and Influences

of Layers in Large Multiplex Network, arXiv preprint arXiv:1703.05833 (2017).

Acknowledgments

This thesis would not have been possible without the complex network of interactions I have experienced in the stimulating environment of the Mathematics Department of Queen Mary University during the last three and half years of my Ph.D. course.

The first thanks go to my supervisors Lucas Lacasa and Ginestra Bianconi, who guided me in my discovery of complex networks and systems theory, and were able to pass on to me their dedication and passion in doing scientific research.

A special thanks to Vito Latora, Vincenzo Nicosia and my longtime friend Federico Battiston, of the Complex Systems and Networks group, with whom I have collaborated on the work regarding the emergence of multiplex communities, for the many interesting conversations, and precious advices.

Then a thanks to all the external collaborators with whom I had the pleasure of working with to produce the papers that became the core of the thesis, in particular Santo Fortunato and Richard Darst for the work on triadic closure and emergence of communities, Zihao Wu for the paper on the characterisation of mesoscopic structures in multiplex networks, and Christoph Rahmede and Alex Arenas for our recent works regarding the Functional Multiplex PageRank and the centrality of nodes and influences of layers in large multiplex networks.

I have to thank very much David Arrowsmith who, together with Vincenzo, gave me important feedback and suggestions regarding the structure and the organization of this thesis.

Of course nothing would have been the same without all my other Ph.D friends, network scientists and non, Valerio Ciotti, Moreno Bonaventura, Andrea Cairoli, Massimo Cavallaro, Nils Haug, Sol Gil Gallegos, Andre Nock, Thomas Coyne, Louise Sutton, Sune Jakobsen, Ryan Flanagan, Paul Geffert, Andrea Santoro and Iacopo Iacopini for all the coffee and tea breaks, post-office beers and time spent together inside and outside the department.

Finally these last lines are for Elif, who has shared with me, in the most beautiful way, all aspects of my life during this journey.

Table of Contents

Abstract	iii
Acknowledgments	v
Table of Contents	vi
1 Introduction	1
1.1 Complex network structures	1
1.2 Multiplex architectures	3
1.3 Visibility graphs	4
2 Triads' formation in networks and emergence of modular structure	8
2.1 Defining and detecting modules	8
2.2 The triadic closure mechanism	10
2.3 Emergence of community structure in an evolving network model	19
2.3.1 The basic model including triadic closure	19
2.3.2 The triadic closure model including fitness of the nodes	29
3 Characterisation of mesoscopic structures in multiplex networks	39
3.1 Communities in multiplex networks	39
3.2 Information theory to characterise multiplex mesoscopic structures: the $\tilde{\Theta}^S$ indicator	41
3.2.1 Testing $\tilde{\Theta}^S$ on Benchmark Models	45
3.2.2 The network between the layers of the APS Collaboration Multiplex Networks	50
3.2.3 Comparison of the results obtained with $\tilde{\Theta}^S$ respect to other similarity measures	59
3.3 Emergence of multiplex communities in social collaboration networks	61
3.3.1 Empirical analysis of multiplex collaboration network datasets	62
3.3.2 A model for evolving multiplex communities	64

3.3.3	Model calibration for generic multiplex networks	70
4	Extracting network motifs from time series	74
4.1	Motifs in time series	74
4.2	Sequential visibility graph motifs	76
4.3	Theory for sequential HVG motifs	81
4.3.1	Random dynamics: i.i.d.	84
4.3.2	Deterministic chaos: fully chaotic logistic map	87
4.3.3	Convergence of finite series	90
4.3.4	Stochastic processes with correlations	91
4.4	Theory for sequential VG motifs	93
4.4.1	<i>Unbounded</i> variable $x \in (-\infty, \infty)$	94
4.4.2	<i>Bounded</i> variables $x \in [a, b]$; $a, b \in \mathbb{R}$; $a, b < \infty$	95
4.4.3	Fully chaotic logistic map	97
4.4.4	Uniform white noise	100
4.4.5	Gaussian white noise	100
4.4.6	Gaussian red noise	101
4.4.7	Noise characterisation	102
4.5	Time series classification via visibility graphs motifs	104
4.5.1	Robustness	104
4.5.2	Principal component analysis	106
4.6	Unsupervised learning: disentangling meditative from other relaxation states using HVG motif profiles from heart rate time series	109
4.6.1	Data	109
4.6.2	Unsupervised clustering based on HVG motif profiles	110
4.7	Comparison of the performance in classification tasks between HVG and VG motifs	115
4.7.1	Convergence properties for finite size series	115
4.7.2	Robustness against measurement noise	117
5	Conclusion	121
	Appendix A Methods for the analysis of multiplex social collaboration networks	125
	Appendix B Explicit computation of the HVG motif profile for the fully chaotic logistic map	128
	Appendix C Motif profiles for all subjects in the empirical study	131

Chapter 1

Introduction

1.1 Complex network structures

Network Science [1–3] aims to study complex systems observed in nature, society and technology with the idea of representing the interactions and the interacting entities involved as a network structure of nodes and links and analysing the structure to extract relevant information on the system.

Far from being confused with applied graph theory this field of science has gathered and integrated over the last 20 years scientists and ideas from mathematics, physics, computer science, biology, sociology, economy and many other disciplines and has become unique and highly interdisciplinary. One of the reasons of its success is that during the development of Network Science, thanks to the advances in technology, information and computing, there has been a parallel progress in the volume of data collection, storage and availability in all different areas of science and society, and scientists soon realised that many of these datasets have a natural representation in terms of networks.

For example we can think at the World Wide Web, at social networks such as Twitter or Facebook, at collaboration networks between co-authors of scientific papers from Science or Nature or between actors co-starring in the movies we can find online on the IMDb database; at the connectome, the complete map of synaptic connections between neurons, of the *C. elegans* worm; at the air transportation network and at urban systems; at the money transfers between banks; at the interactions of the genes and proteins in the cell. All these networks have something in common: they all show a structure that is far from being similar to the one of a regular lattice or to the one of a random graph, a graph where the links between nodes are put at random, and we call these network structures ‘complex’.

Developing models and algorithms to characterise complex networks has been and is still a fundamental challenge to analyse and extract relevant information from network datasets and to gain insights for understanding real world complex systems. It becomes even more relevant when we think that networks can be used as a general mathematical framework to analyse any kind of data that can be mapped into a graph due to some intrinsic correlations, such as, for example, complex signals, time series and multi-dimensional data.

Much of the information from a network can be gained by looking at the local properties of the network such as the node degree (number of connections) or the node clustering coefficient, the number of triangle each node closes, or the average degree of the node's neighbours [1, 3]. Nevertheless some of the properties that mostly characterise complex network structures are mesoscopic, that means that are properties observed not at the scale of the single nodes and not at the scale of the overall network. One of these features are called *network motifs* [4, 5]. They are patterns of interconnections which appear in complex network structures with a probability much higher than in random networks. They are found in networks from biochemistry, neurobiology, ecology and technology and they represent small repeated and conserved evolutionary devices, thus building blocks of complex network structures. The concept of motifs was first introduced in 2002 [4] by Milo and collaborators who studied the presence of motifs in different typology of biological and engineering networks and realised that motifs can be used to define universal classes of networks.

The other important structural feature observed in almost all real world networks in nature is their inherent simple modular structure that means they can be separated into units or 'modules' that are coherent subgraph associated to distinct functions in the case of biological and technological networks or to organisational structures in the case of social systems. Modularity appears to be one of the fundamental organising principles for all complex networks we find in nature, going from metabolic [6, 7] and cellular networks [8, 9] to neural networks [10] to ecological and social networks [11–13], and modular structure and community structure have become synonymous of mesoscale structure in network literature.

It's interesting to notice that from a network perspective motifs and modules are defined in similar way: clustered in a graph theory sense. We can distinguish the two only by emphasising small size and recurrence for motifs and larger size for modules, and modules can be considered overlapping and functionally significant subnetworks with a structure probably dominated by interconnected motifs [14, 15].

In what follows we will discuss how these two mesoscale properties are related to each other and in particular the role that motifs play in characterising complex network struc-

tures at the mesoscale for different families of networks: standard networks, multiplex networks and visibility graphs.

1.2 Multiplex architectures

Multiplex networks [16, 17] are a family of networks that describe a large number of complex systems where interactions of different nature are involved. Multiplex networks are formed by a set of N nodes, the interacting elements of the system, connected through M different networks (layers), each one describing the connections for a specific type of interaction (see Figure 1.1). Multiplex networks have been first proposed for modelling social networks [18], where the same set of individuals are connected by different types of social ties (friendship, collaboration, family tie, etc.), or can communicate via different means of communication (email, mobile phone, chat, Facebook, etc.). Nevertheless the limitation of understanding complex biological and technological systems through single-layer networks has been gradually and increasingly pointed out through a series of works in which multiplex networks have been used to model a larger set of data including transportation networks [19, 20], scientific and actor collaboration networks [20, 21], biological networks in the cell or in the brain [22–25], complex infrastructures, and economical networks. Multiplex networks encode significant more information than the network which includes all the interactions of the multilayer network compressed in one single layer without distinguishing the nature of the links. It is intuitively clear that if we want to understand traffic congestion phenomena in the London public transportation system we have to consider simultaneously many network layers which describe tube lines, bus lines and railways lines, or if we want to gain knowledge of the cell functions we need to integrate information from different layers of interaction, such as, for example, the protein-protein interaction network, the signalling network, the metabolic network and the transcription network.

As a consequence of this, one the most pressing challenge in multiplex network theory is devising algorithms and numerical methods to extract relevant information from these network structures, and one way to do that is by exploiting the ubiquitous structural correlations that are found in multiplex networks.

For example the study of the overlap of the links [26] in the different layers of a multiplex network has been conducted for systems as different as in-silico societies [27], multilayer airport networks [28] or citation-collaboration networks [21], and provides information that cannot be extracted if the single layers are taken in isolation and that is fundamental for the understanding of the dynamical processes taking place at the level

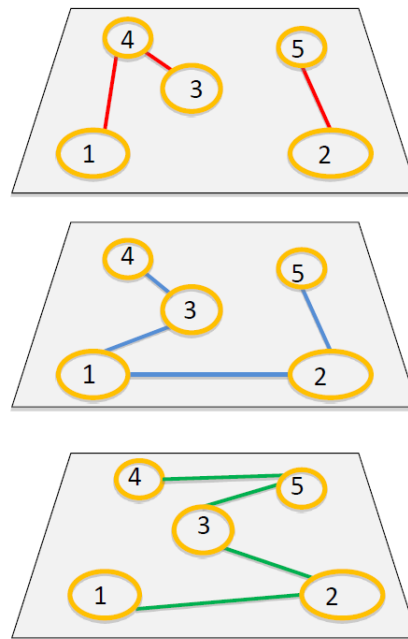


Figure 1.1: A multiplex network structure formed by a set of four nodes and three layers. Multiplex networks are often used in the case in which the same set of nodes is connected by links indicating the different types of interactions. The links of the same type form the layers of the multiplex network, for example a social network can be represented by a multiplex in which the same set of individuals can interact via phone, email or post.

of the aggregated network [19, 21].

If we consider the nodes in a multiplex we can measure in how many of the layers they are connected or ‘active’ and study the the heterogeneous of the nodes’ activity in different layers [20, 22], and in general we can extract relevant information by measuring the correlations between the local properties of the same node in different layers.

We will see in Chapter 2 that it is possible to gain important and unique information from multiplex networks from their mesoscopic structure by looking at the structural correlations between the communities in different layers [29, 30].

1.3 Visibility graphs

In recent years different methods [31–35] have been introduced to map the structure and underlying dynamics of a given time series into an associated graph representation, with the scope of exploiting tools and concepts proper of the modern network science [2, 3, 36] in the traditional task of time series analysis [37, 38], thus building a bridge between the

two fields.

The family of visibility graphs comprehends special graphs that can be extracted from time series by means of the so called *visibility algorithms* and allow to analyse the structure of the series through the tools developed in the graph/complex network theory.

Let $\{x(t_i)\}_{i=1..N}$ be a time series of N data. The natural visibility algorithm [39] assigns each datum of the series to a node in the natural visibility graph (NVG). Two nodes i and j in the graph are connected if it is possible to draw a straight line from the point $x(t_i)$ to the point $x(t_j)$ that stays above any intermediate data point $x(t_k)$ (see Figure 1.2). Hence, node i and node j are connected nodes if the following geometrical criterion is fulfilled within the time series:

$$x(t_k) < x(t_i) + (x(t_j) - x(t_i)) \frac{t_k - t_i}{t_j - t_i}. \quad (1.1)$$

It is easy to check that the associated graph extracted from a time series through this algorithm, is always:

- (i) connected: each node sees at least its nearest neighbours (left-hand side and right-hand side).
- (ii) undirected: a criterium for the directionality of the links is not provided by the algorithm.
- (iii) invariant under affine transformations of the series data: the visibility criterium is invariant under rescaling of both horizontal and vertical axis, as well as under horizontal and vertical translations.
- (iv) “lossy”: partial information regarding the time series is inevitably lost in the mapping-compression process. Indeed the network structure is completely determined by the binary adjacency matrix and, for example, two periodic series with the same period would give the same visibility graph, despite being quantitatively different.

In order to understand the geometric interpretation of the visibility graph, let us focus on a periodic series. It is straightforward that its visibility graph is a concatenation of a motif: a repetition of a pattern (see Figure 1.2). Now, which is the degree distribution $P(k)$ of this visibility graph? Since the graph is just a motif’s repetition, the degree distribution will be formed by a finite number of non-null values, this number being related to the period of the associated periodic series. This behaviour reminds us the Discrete Fourier Transform (DFT), which for periodic series is formed by a finite number of peaks (vibration modes) related to the series period. Using this analogy, we can understand the visibility algorithm as a geometric (rather than integral) transform. Whereas a DFT decomposes a signal in a sum of (eventually infinite) modes, the visibil-

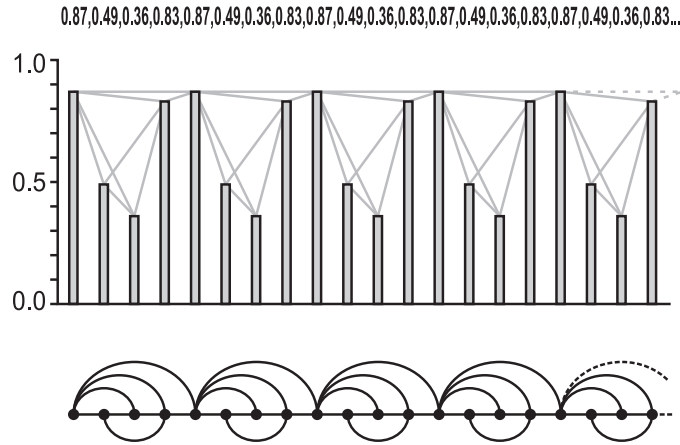


Figure 1.2: Illustrative sketch of the visibility algorithm. In the top a periodic time series is plotted while in the bottom the corresponding graph generated through the visibility algorithm is shown. Each datum in the series corresponds to a node in the graph, such that two nodes are connected if their corresponding data heights fulfil the visibility criterion of equation 1.1. Taken from [39].

ity algorithm decomposes a signal in a concatenation of graph's motifs, and the degree distribution simply makes a histogram of such 'geometric modes'. While the time series is defined in the time domain and the DFT is defined on the frequency domain, the visibility graph is then defined on the 'visibility domain'. We will see in Chapter 3 that the visibility algorithm is able to distinguish between stochastic and chaotic series whereas a generic DFT fails to capture the presence of nonlinear correlations in time series (such as the presence of chaotic behaviour).

A more simple, alternative criterion for the construction of the visibility graph is defined as follows: let $\{x_i\}_{i=1..N}$ be a time series of N data. The so called horizontal visibility algorithm [40] assigns each datum of the series to a node in the horizontal visibility graph (HVG). Two nodes i and j in the graph are connected if one can draw a horizontal line in the time series represented with bars from x_i to x_j that does not intersect any intermediate data height (see Figure 1.3 for an illustration). Hence, i and j are two connected nodes if the following geometrical criterion is fulfilled within the time series:

$$x_i, x_j > x_n \text{ for all } n \text{ such that } i < n < j \quad (1.2)$$

The resulting HVG is always a subgraph of the NVG associated to the same time

series (see Figure 1.3). The HVG graph will also be (i) connected, (ii) undirected, (iii) invariant under affine transformations of the series and (iv) “lossy”. Some concrete properties of these graphs can be found in [40–42].

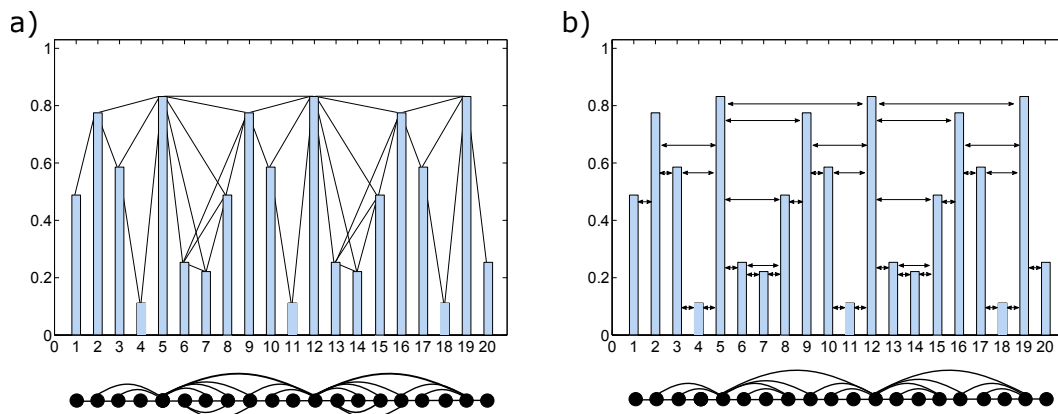


Figure 1.3: Schematic of two families of visibility algorithms. (a): natural visibility algorithm applied to 20 data points of a periodic time series (top) and the corresponding Visibility Graph (VG) (bottom); each datum in the series corresponds to a node in the graph and two nodes are connected if their corresponding data heights show mutual visibility (Eq. 1.1). (b): horizontal visibility algorithm applied to the same series (top) and the corresponding Horizontal Visibility Graph (HVG) (bottom); each datum in the series corresponds to a node in the graph and two nodes are connected if their corresponding data heights show horizontal visibility (Eq. 1.2). Taken from [43].

In Chapter 3 we will focus on the mesoscale structure of visibility graphs, in particular we will introduce the sequential visibility graph motifs, small subgraphs that can be detected along the Hamiltonian path and that can be used to distinguish and classify time series structure and dynamical processes of different nature.

Chapter 2

Triads' formation in networks and emergence of modular structure

2.1 Defining and detecting modules

One of the most relevant feature of networks representing real complex systems is the community (or modular) structure, i.e. the organization of nodes into clusters, with many links joining nodes of the same cluster and comparatively few links joining nodes of different clusters [44]. In the same cluster (or module) nodes likely share common properties and/or play similar roles within the network.

In biological networks such as metabolic, cellular or neural networks the modules can be seen as coherent subsystems performing distinct functions [8, 10]. We can think for example of groups of co-expressed genes in gene regulatory networks, proteins in the cytoskeleton forming supra-molecular structures such as microtubules via physical interactions, or cortical areas of the brain performing different cognitive or motor functions [9, 10]. In social networks the communities can be friendship circles or groups of people sharing common interests or doing activities together, and they reflect, in general, the self-organisation of individuals in order to optimise some task performance [12].

Detecting communities is a very hard problem, not yet satisfactorily solved despite the huge effort of a large interdisciplinary community of scientists who has been working on it over the past few years [44]. The fundamental question is that no rigorous definition of community exists in graph theory. Usually a community is defined as a subset of nodes inside the graph which are more densely linked together when compared to the rest of the network [44, 45]. The limitations of such a definition become clear when we try to

apply it to specific network structures, for example to networks consisting of multiple disconnected components or to bipartite (multipartite) networks in which there are two (multiple) disjoint sets of nodes and links between nodes may occur only if the nodes belong to different sets, making necessary to define clusters of nodes in the same set differently.

For this reason different algorithms have been developed during the last decades based on different principles [44–46]. Until 2002 the traditional way of detecting network communities has been the hierarchical clustering [13, 47]. Hierarchical clustering is based on the idea of assigning certain weights to the nodes of a network that describe how close or similar the nodes are; then nodes are grouped iteratively into clusters, weights between clusters are calculated and clusters are clustered together till all nodes are grouped in just one cluster. In this sense the core of this method is (I) to find an appropriate similarity measure between the nodes to perform the iterative clustering algorithm and (II) to specify a rule to determine at which level of the clusters' hierarchy one finds the optimal partition of nodes into communities.

In 2002 Girvan and Newman proposed an algorithm [13] based on the idea that links connecting highly clustered communities have a higher edge betweenness, where edge betweenness is defined as the number of shortest paths between pairs of nodes that run through that link [3], and the shortest path between two nodes is the shortest sequence of links which is needed to jump from one of the two to the other. The algorithm proceeds in an iterative way by calculating at each iteration step the edge betweenness of all links inside the network, and by identifying and removing the links with highest betweenness. The iterated removal of links with highest betweenness eventually causes the initial connected network to gradually split into more and more separate clusters.

In 2004, again Girvan and Newman [48] introduced the quality function Q , *modularity*, as a stopping criterion for their algorithm. The modularity definition is based on the intuitive idea that random networks do not exhibit community structure. Thus if we think of having an arbitrary partition of a network in n_c communities with m links we can write

$$Q = \frac{1}{2m} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \quad (2.1)$$

where c_i is the community assignment of the generic node i and the Kronecker symbol δ is selecting only the nodes joining the same community; Q is thus summing up for each cluster c the differences between the actual fraction of links found inside that cluster $\frac{1}{2m} \sum_{i,j} a_{ij}^c$ and the expected fraction of links one would have found in the case of a random network $\frac{1}{2m} \sum_{i,j} \frac{k_i^c k_j^c}{2m}$ (indeed the probability for two generic nodes i and j to be connected in a random network is $p_{ij} = \frac{k_i k_j}{\langle k \rangle} = \frac{k_i k_j}{2m}$).

If the network does not exhibit community structure, or if the partitions are allocated

without any regard to the underlying structure, the expected number of intra-community links $\sum_{i,j} a_{ij}^c$ equals $\sum_{i,j} \frac{k_i^c k_j^c}{2m}$ in each cluster and $Q = 0$. In general Q is strictly less than 1 and can be positive or negative if we found respectively more or less intra-community links than expected in the random null model.

Soon after the Girvan and Newman paper, optimization of modularity has rapidly become *per se* the essential element of many clustering methods [44]. The Louvain algorithm developed in 2008 is a heuristic method based on an iterative optimization of modularity starting from an arbitrary partition of the network into clusters, and it outperforms all other known community detection methods based on the same idea in terms of computational time [49].

Although modularity is the first rigorous essential ingredient for the theoretical comprehension of clustering in networks the fundamental assumption for which a random network exhibits very small values of modularity is not completely true: it has been shown that it is possible for a random network to find a partition which gives high values of modularity due to the fluctuations in the link distribution [50].

For this reason the development of others algorithms not based on modularity optimization is of great importance. Currently the one which is thought to be the best is Infomap [44, 51]. This algorithm uses the probability flow of random walks on a network as a proxy for the actual information flows in the real system and decompose the network into modules by compressing a description of the probability flow. The result is a map that both simplifies and highlights the regularities in the structure of the network and their relationships (see Figure 2.1).

Despite the fact that on empirical cases some detection algorithm works better than others, the field of community detection remains open and it is often an hard task to prove in general that one algorithm can outperform another.

In the following we will try to understand how communities in real evolving networks can emerge during their evolution and which are the mechanism responsible for the community formation. To this aim we will introduce the mechanism none as *triadic closure* and we will give a brief historical overview of the scientific literature that has focused on this mechanism.

2.2 The triadic closure mechanism

Triadic closure in network science can be defined as the general mechanism by which a node establishes a connection with a neighbour of a node it is already connected with. From a network topology point of view the triadic closure mechanism characterises the

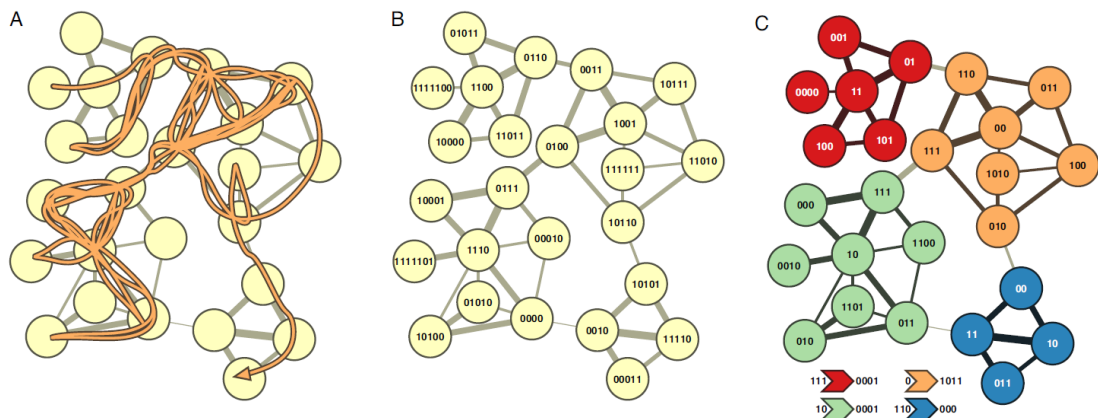


Figure 2.1: The Infomap algorithm detects communities by compressing the description of information flows on networks to reveal important mesoscopic structures. A random walker (orange trajectory) explores the network (A) and a unique code is assigned to every node in the network (B) based on the number of times the walker has visited the node (Huffman code). This code is then compressed to obtain a coarse grained description of the network (C). (Taken from [51])

evolution of a network with the formation of triangles or *triads*, motifs of three nodes all connected with each other.

The importance of the triads as relational structures was first recognized by sociologists long before the development of modern network theory. In 1908 George Simmel [52, 53], tried to develop a theory of triads by studying the social interactions in groups of three individuals, where each member seeks to control the others, and may differ in ‘strength’, an additive mathematical quantity specifying the ability to control a weaker member or to get controlled by a stronger one (see Figure 2.2). Simmel’s social experiments revealed a universal tendency of triads to become a coalition of two against one. The German sociologist also suggested that his theory of the triads can be generalised to situations in which interacting units represent groups of individuals, and thus is applicable at different scales of social interaction, an idea that interestingly anticipates somehow the concepts of hierarchy and hierarchical modularity proper of the the modern network theory.

In the 50’s sociologists continued to study triads and, thanks to the introduction of sociograms in the late 30’s, graphs describing social interactions between individuals, they combined the analysis of triads with the network description of social interactions. In 1953 A. Rapoport [54, 55] studied the process of spreading of information through a population. By the analysis of some data he observed that the measured spread velocity in the network of communication between people couldn’t be obtained by simply considering a random network model, but it could have been explained by assuming a

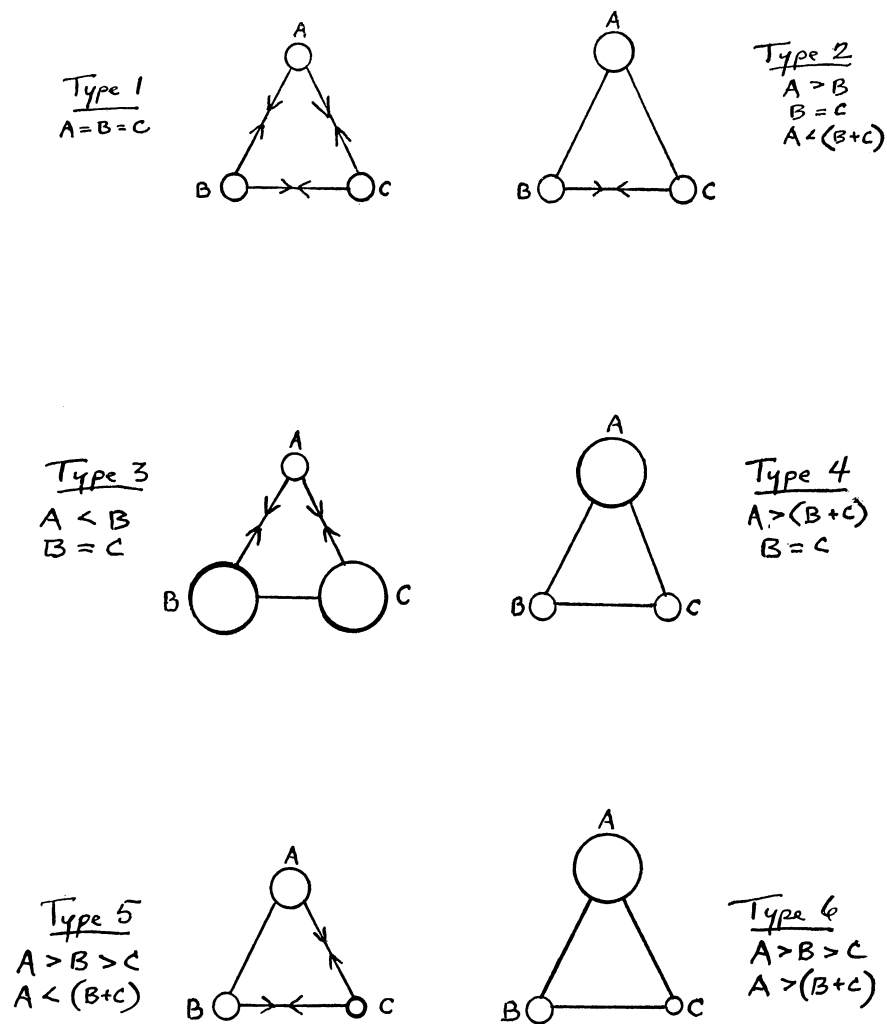


Figure 2.2: Theory of coalitions into triads (taken from [53]). Each of the three individuals A, B and C can use his/her strength (size of the node) to seek control over the others. Despite different possible scenarios of strength distribution often a coalition (directed links) of two individuals against one emerges.

socio-structural bias for which “the likely contacts of two individuals who themselves have been in contact are expected to be strongly overlapping”. He thus suggested experimental procedures to test the statistical properties of the *acquaintance net* of the population, and his paper can be considered the first work on the geometry of social networks.

At the end of the 90’s the modern theory of complex networks began. It was recognized that networks can describe systems of very diverse origin, from nature, society and technology [1, 56, 57] and that those systems can be characterized by a number of general network properties. The feature that received most attention in the literature was the distribution of the degree (the number of neighbours of a node), which is highly skewed,

with a tail that can be often well approximated by a power law [58]. Such property indeed explains a number of striking characteristics of complex networks, like their high resilience to random failures [59] and the very rapid dynamics of diffusion phenomena, like epidemic spreading [60]. To explain board degree distributions the first proposed (and still commonly accepted) mechanism is preferential attachment [61], introduced by Barabási and Albert in 1999. In their (BA) growing network model, new nodes set links with existing nodes with a probability proportional to the degree of the latter. This way the rate of accretion of neighbours is higher for nodes with more connections, and the final degrees will be distributed according to a power law.

The study of large social systems revealed, on the other hand, the presence of a surprisingly high number of triangles in the corresponding networks [62–64] and it was soon realized that models of social networks must have been able to reproduce high values of the clustering coefficient. The small world model (WS) proposed by Watts and Strogatz in 1998 [62] was able to do that. In the WS model one starts by considering a ring lattice of order n with each node connected with all the others within a lattice distance k , then according to a certain probability p each connection is rewired uniformly at random. Since the model is an interpolation between a regular lattice $p = 0$ and a random network $p = 1$ it inherits two important features, namely the high clustering proper of the lattice and the short diameter proper of the random network.

The problem was that none of the two models is able to reproduce together board degree distribution, missing in the WS model, and high clustering coefficient, missing in the BA model. One of the early purposes for which triadic closure was introduced in models of evolving networks was to go beyond the limits of the BA model and of the WS model taken alone. The model by Holme and Kim [65] introduced in 2002 is a variant of the BA model able to generate scale free networks with high clustering. The new node joining the network sets a link with an existing node, chosen with a probability proportional to the degree of the latter, just like in the BA model. To incorporate the high clustering, the other $m - 1$ links of the new node are attached with a probability P_t to a random neighbour of the node which received first link, forming a triad, and with a probability $1 - P_t$ to another node chosen with preferential attachment. The parameter P_t controls the process of triad formation, and by varying it it is possible to tune the level of clustering into the network, while the degree distribution is always a power law.

With the similar spirit of trying to unify the ‘small-worldness’ and ‘scale-freeness’ David-son et al. [66] proposed in the same year a model of acquaintance network that was able to reproduce values of high clustering and short path length (diameter), and to interpolate between networks with scale free and exponential degree distribution, both of which were observed in real social networks. In their network model with fixed number of nodes, at each time step, one randomly chosen node introduces two of his random neighbours

to each other, a link is made between these two neighbours and thus a triad is formed. Instead in the case that the chosen node has only one neighbour, it establishes a link with a randomly chosen node. Then with a certain probability P_D a randomly chosen node is removed along his connections (individual who dies) and a new node (individual who is born) enters the network connecting with a random chosen node. The parameter $P_D < 1$ determines the separation between the temporal scales of the social relations and of the death-birth event: for $P_D \ll 1$ the triad formation dominates the Poissonian death process and the degree distribution is scale free while for large value of P_D the opposite is true and the distribution becomes a stretched exponential. The importance of the model proposed by Davidsen et al. resides in the dynamical explanation it provides of the small world topology, and in the fact that this explanation can be found in a *local* mechanism such as triadic closure.

The *locality* nature of triadic closure has been indeed the second important characteristic that increased the interest of network scientists for this mechanism. The BA model's philosophical implication is that whenever we have a growing network which grows according to an effective preferential attachment process, then the resulting degree distribution is power law. However the model in its details presents a problem that makes it difficult to generalize to all real networks' scenarios: if the new node links preferentially to nodes with high degree this implies that the new node must possess all the information regarding the degree of the nodes already in the network, meaning that explicit preferential attachment is not a local rule. In 2001 Dorogovtsev et al. [67] implemented a model of a growing scale free network in which preferential attachment arises naturally without any rule depending explicitly on the nodes' degree. The model, that in graph theory can be described as an evolving simplicial complex [68], is the simplest model of scale free network. One starts by considering a closed triad as the starting network and, at each time step of the evolution, adds a new node which connects to both ends of a randomly chosen link, as reported in Figure 2.3, closing in this way another triangle. Thus this network evolves throughout a continuous aggregation of triads to its structure. Since the probability for a node in the network to be one end of the chosen link is proportional to its degree, the attachment is preferential. The interesting implication of this model is that triadic closure is the the simplest implicit mechanism of preferential attachment.

Soon after in 2003 Vázquez [69] discussed various *local* mechanisms of network evolution generating power law degree distributions, high clustering and degree correlations. Starting from the assumption that in social graphs if two nodes have a common neighbour have also more chances to get connected, he proposed a *connecting nearest neighbors* model where pairs of nodes having a common neighbour are connected by a *potential* edge. At each time step of the network evolution (1) with probability $1 - P_V$ a new

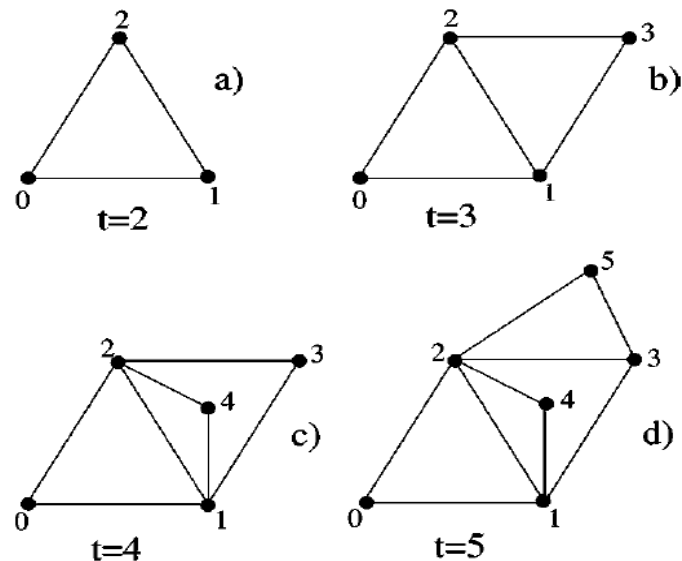


Figure 2.3: In the simplest model of growing scale free network proposed by Dorogovtsev et al. [67], the network evolves through a continuous aggregation of triangles. Taken from [67].

node is linked to a random chosen node in the network and *potential* edges are established between the new node and all the neighbours of the chosen one; (2) a random chosen *potential* edge becomes an edge with probability P_V . Vázquez understood that the triadic closure mechanism in the variant enforced in its model also induces an effective preferential attachment because although new nodes connects to an existing node at random, the potential edges favour nodes with higher degree in acquiring links when triads are subsequently closed. In other words, if a node has high degree then it has more chances of acquiring a potential link because the probability that one of his neighbours is randomly chosen by new nodes increases with its degree. Vázquez [69] claimed that the model yields always power law degree distributions, that is not correct in general as we will see in detail in the next Section, and the reason is that if the growing network exhibits degree correlations the probability of acquiring links is not linearly proportional to the node's degree. The same wrong conclusions on the power law nature of the degree distribution was presented later by Jackson and Rogers [70], who studied a similar model of social network formation where individuals make new connections in two ways: uniformly at random (random links) and via local search to 'friends of their friends' (triadic closure).

Finally, the third reason, perhaps the most important, for which the triadic closure mechanism has received lot of attention in the scientific literature resides in its ability to explain the dynamical emergence of modularity in network systems.

In the contest of social networks the connection between triadic closure and communities has been understood slowly and never really formalized through an extended and comprehensive analysis of the various models that we could consider joining the class of triadic closure models. The ones who first realised this connection were Jin, Girvan and Newman with a work published in 2001 [71] where they proposed a model of social network evolution with fixed number of nodes where (1) links between individuals happen preferentially when there is the presence of a mutual friend, (2) the number of friends of each individual is limited, and (3) the friendship connections can decay over time. They observed a clearly defined community structure in the evolved network, although the resulting network was not in the form of a single connected component and some communities were formed by isolated clusters of individuals. They found that links between different communities were the ones more likely to decay in time, while communities were characterized by high density of 'triangles' with long-lasting links, a clue that communities were produced and reinforced by the acquaintance mechanism.

Despite triadic closure was firstly recognized as a possible mechanism of communities' formation, the many ingredients in the social dynamics made this acquaintance network model quite specific and diverted the attention from the potentiality of the mechanism alone. In 2006 Toivonen et al. [72] found that community structure emerges from a simple model of network growing through a mixture of global search via random attachment and local search through triadic closure, a model quite similar to the ones proposed by Vázquez and Jackson and Rogers, which incorporates the two simplest fundamental processes shaping the topology of social networks and that inspired later an analogous model for weighted networks [73]. They attributed the emergence of the observed significant community to the interplay between the two processes. They pointed out that triadic closure tends to enlarge existing communities because when the triangles start to be formed in regions with high density of links they also tend to be confined there. Instead the random links made by new nodes tends to form bridges between communities. In a certain sense they wrote about the emergence of communities but they never really explained how communities originate. We will explain the mechanism of the emergence in detail in the next Section.

Before concluding we have to say that although social networks represent the contest in which the triad formation has the most intuitive interpretation in terms of acquaintance, triadic closure can be considered a 'universal' mechanism in the sense that it can be observed in a variety of different real complex systems under different forms.

For example in the contest of technological networks the triadic closure is a direct consequence of what is often called the *copying* mechanism [74, 75]. If we consider the evolution of the World Wide Web, new pages are usually created by making a copy of

the hyperlinks of other pages. A single copying process can form either closed triads, if the copied hyperlinks refer to pages hyperlinked in turn, or otherwise open triads that later could become closed thanks to other mechanisms. It's intuitively clear that the reiteration of the *copying* mechanism alone is a powerful tool for the formation of triads. Analogously in a network of scientific citations, where nodes represent papers and directed links represent citation from a paper to another, authors who publish a new article usually cite a paper along with the relevant papers that they find in its bibliography. Popular in this sense is the model of growing network by copying (GNC) proposed by Karpivsky and Redner [75]. The network is directed and it grows by addition of one link at a time. New nodes randomly select a target node and connect to it (outgoing link), as well as to all ancestor nodes of the target node (neighbours to which the target node points). Thus if new nodes choose only the root node as the target, the resulting network is a star graph. In the opposite scenario, if the target node is always the most recent node added in the network the copying mechanism generates a complete graph.

Last, in network biology, the fundamental mechanism which leads to triadic closure is called *duplication – divergence* and has been successfully implemented in models of growing networks to explain the evolution of the proteome, the network describing the physical interactions between the proteins in the cell [69, 76–79]. When the genome is duplicated, a single gene can undergo a mutation and starts to code for a protein which will be similar to the one originally coded one (before the mutation) but not exactly the same. From a network biology perspective we could say that this new protein will inherit *almost* the same set of physical interactions of the protein coded by the original gene. Thus in a network model of the proteome the evolution determined by a gene's mutation can be represented as the *duplication* of a node, representing the original protein, along with all its connections, and the deletion and eventually the creation of some links (*divergence*), representing the lost of some interactions and the acquisition of new ones due to the mutation in the protein itself (see Figure 2.4). It is easy to figure that this biological version of the *copying* easily leads to the closure of triads in the network evolution. Using this model Pastor-Satorras, Smith and Solé determined the two average probabilities respectively of deletion of interactions and of appearance of new ones in the proteome of the *S.cerevisiae* by tuning the exponent of the power-law degree distribution of the model to the value measured experimentally. They also found that for those values of the probabilities more than 40 % of the nodes in the network model become disconnected, consistently with the experimental observation that 50 % of duplicated genes lose their function after the duplication and the other half experience functional divergence. Later Solé and Valverde realized that this model of biological network show spontaneous emergence of modularity in its structure. They suggested that random mutations, which lead to *duplication – divergence*, are able to explain alone

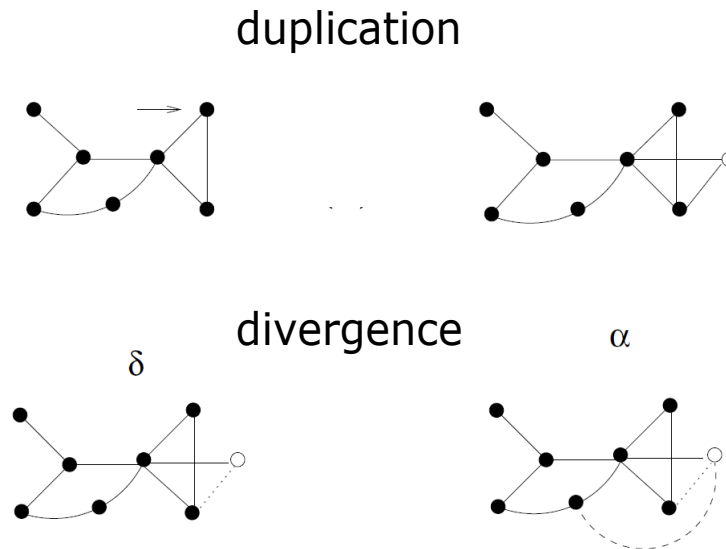


Figure 2.4: In the duplication-divergence model a node (black arrow) is chosen to be duplicated along with its connections. Then its links are removed with a certain probability δ and new links are created with probability α .

the modularity observed everywhere in biological systems. Moreover, being modularity a result of biological evolution, they suggested that evolution itself results from the inner complexity of the system: modularity structure emerge without any use of any explicit parameters, such as *fitness*, or of any external environmental pressure, proper or other classes of models [9], in an opposite perspective to the one of Kashtan et al. [80] who showed that modularity can emerge in a logical network circuit when it evolves under a modular changing environment (modular varying logic tasks).

The truth is that in the *duplication – divergence* model the role of the environment is actually hidden in those values of the probabilities that determine the deletion and addition of interactions. Nevertheless the important and interesting debate of whether modularity arises in biological systems because of internal complexity generated by small random processes or because of the external pressure of the environment is still open.

All the examples seen reveal the universal character of the triadic closure process and its crucial role in shaping the mesoscale structure of a variety of different real-world networks. In the next Section we will discuss under which conditions the triadic closure models are able to generate communities and how the communities originate. We will incorporate in a model of network growth driven by triadic closure the idea of fitness through a node parameter η quantifying the different degree of ability of each node to attract and connect new links during the evolution, and we we will show how the distribution of the fitness of the nodes importantly affect the evolution and the size distribution of the communities.

2.3 Emergence of community structure in an evolving network model

2.3.1 The basic model including triadic closure

We begin with what is possibly the simplest model of network growth based on triadic closure. The starting point is a small connected network of n_0 nodes and $m_0 \geq m$ links. The basic model contains two ingredients:

- *Growth.* At each time a new node is added to the network with m links.
- *Proximity bias.* The probability to attach the new node to node i depends on the order in which the links are added.

The first link of the new node is attached to a random node i_1 of the network. The probability that the new node is attached to node i_1 is then given by

$$\Pi^{[0]}(i_1) = \frac{1}{n_0 + t}. \quad (2.2)$$

The second link is attached to a random node of the network with probability $1 - p$, while with probability p it is attached to a node chosen randomly among the neighbours of node i_1 . Therefore in the first case the probability to attach to a node $i_2 \neq i_1$ is given by

$$\Pi^{[0]}(i_2) = \frac{(1 - \delta_{i_1, i_2})}{n_0 + t - 1}, \quad (2.3)$$

where δ_{i_1, i_2} indicates the Kronecker delta, while in the second case the probability $\Pi^{[1]}(i_2)$ that the new node links to node i_2 is given by

$$\Pi^{[1]}(i_2) = \frac{a_{i_1, i_2}}{k_{i_1}}, \quad (2.4)$$

where a_{ij} is the adjacency matrix of the network and k_{i_1} is the degree of node i_1 .

- *Further edges.* For the model with $m > 2$, further edges are added according to the “second link” rule in the previous point. With probability p , an edge is added to a random neighbour without a link of the *first* node i_1 . With probability $1 - p$, a link is attached to a random node in the network without a link already. A total of m edges are added, 1 initial random edge and $m - 1$ involving triadic closure or random attachment.

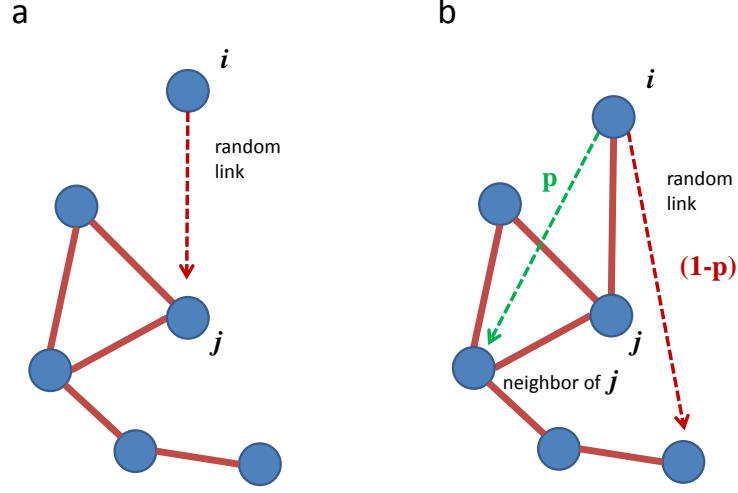


Figure 2.5: Basic model. One link associated to a new node i is attached to a randomly chosen node j , the other links are attached to neighbours of j with probability p , closing triangles, or to other randomly chosen nodes with probability $1 - p$.

In Figure 2.5 the attachment mechanism of the model is schematically illustrated.

For simplicity we discuss here the case $m = 2$. In the basic model the probability that a node i acquires a new link at time t is given by

$$\frac{1}{t} \left[(2 - p) + p \sum_j \frac{a_{ij}}{k_j} \right]. \quad (2.5)$$

In an uncorrelated network, where the probability p_{ij} that a node i is connected to a node j is $p_{ij} = \frac{k_i k_j}{\langle k \rangle n}$ (n being the number of nodes of the network), we might expect that the proximity bias always induces a linear preferential attachment, i.e.

$$\sum_j \frac{a_{ij}}{k_j} \propto k_i, \quad (2.6)$$

but for a correlated network this guess might not be correct. Therefore, assuming, as supported by the simulation results (see Figure 2.6), that the proximity bias induces a linear or sublinear preferential attachment, i.e.

$$\Theta_i = p \sum_j \frac{a_{ij}}{k_j} \simeq c k_i^\theta, \quad (2.7)$$

with $\theta = \theta(p) \leq 1$ and $c = c(p)$, we can write the master equation [81] for the average number $n_k(t)$ of nodes of degree k at time t . From the simulation results it is found that the function $\theta(p)$ is an increasing function of p for $m = 2$. Moreover the exponent θ is also an increasing function of the number of edges of the new node m . Assuming the scaling in Eq. 2.7, the master equation for $m = 2$ reads

$$n_k(t+1) = n_k(t) + \frac{2-p+c(k-1)^\theta}{t} n_{k-1}(t)(1-\delta_{k,2}) - \frac{2-p+ck^\theta}{t} n_k(t) + \delta_{k,2}. \quad (2.8)$$

In the limit of large values of t , we assume that the degree distribution $P(k)$ can be found as $n_k/t \rightarrow P(k)$. So we find the solution for $P(k)$

$$P(k) = C \frac{1}{3-p+ck^\theta} \prod_{j=1}^{k-1} \left(1 - \frac{1}{3-p+cj^\theta} \right), \quad (2.9)$$

where C is a normalization factor. This expression for $\theta < 1$ can be approximated in the continuous limit by

$$P(k) \simeq D \frac{1}{3-p+ck^\theta} e^{-(k-1)G(k-1,\theta,c)}, \quad (2.10)$$

where D is the normalization constant and $G(k, \theta, c)$ is given by

$$\begin{aligned} G(k, \theta, c) &= -\theta {}_2F_1 \left(1, \frac{1}{\theta}, 1 + \frac{1}{\theta}, -\frac{ck^\theta}{3-p} \right) \\ &+ \theta {}_2F_1 \left(1, \frac{1}{\theta}, 1 + \frac{1}{\theta}, -\frac{ck^\theta}{2-p} \right) \\ &+ \log \left(1 - \frac{1}{3-p+ck^\theta} \right). \end{aligned} \quad (2.11)$$

In this case the distribution is broad but not power law. For $\theta = 1$, instead, the distribution can be approximated in the continuous limit by a power law, given by

$$P(k) \simeq D \frac{1}{(3-p+ck)^{1/c+1}}, \quad (2.12)$$

where D is a normalization constant. Therefore we find that the network is scale free only for $\theta = 1$, i.e. only in the absence of degree correlations. In order to confirm the result of our theory, we have extracted from the simulation results the values of the exponents $\theta = \theta(p)$ as a function of p . With these values of the exponents $\theta = \theta(p)$, that turn out to be all smaller than 1, we have evaluated the theoretically expected degree distribution $P(k)$ given by Eq. 2.10 and we have compared it with simulations (see Figure 2.7), finding optimal agreement.

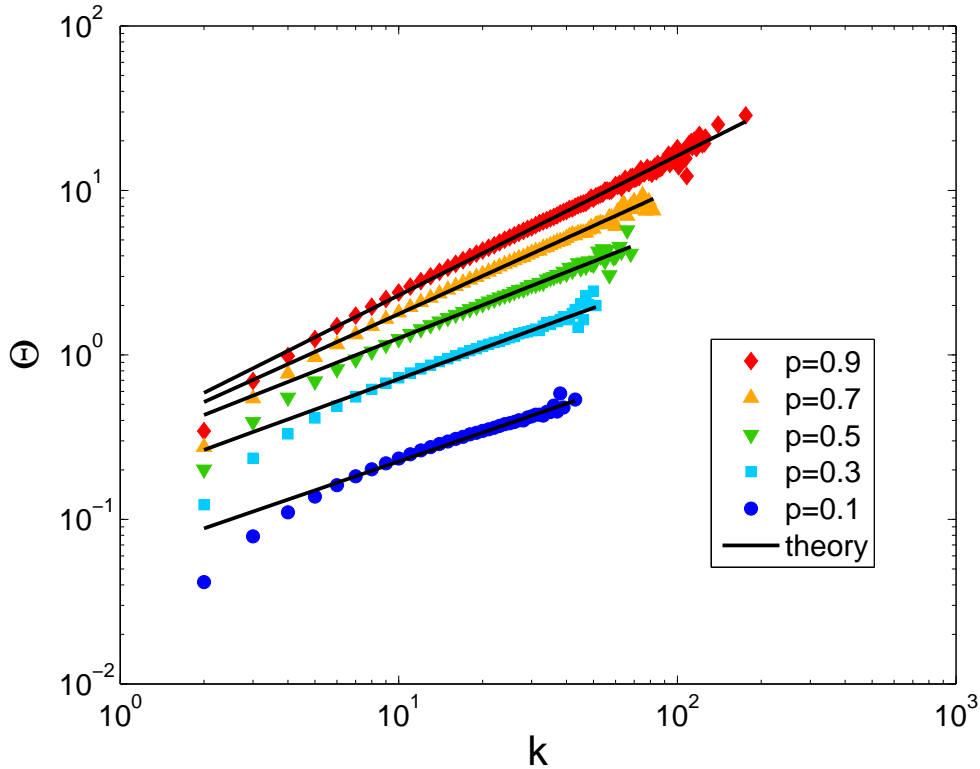


Figure 2.6: Scaling of $\Theta = \langle \Theta_i \rangle_{k_i=k}$, the average of Θ_i , performed over nodes of degree $k_i = k$, versus the degree k . This scaling allows us to define the exponents $\theta = \theta(p)$ defined by Eq. 2.7. The figure is obtained by performing 100 realizations of networks of size $n = 100\,000$.

We remark that this model has been already studied in independent papers by Vázquez [69] and Jackson [70], who claimed that the model yields always power law degree distributions. Our derivation for $m = 2$ shows that this is not correct, in general, and in particular it is not correct when the growing network exhibits degree correlations, in which case we do not expect that the probability to reach a node of degree k_A by following a link is proportional to k_A . When the network is correlated we always find $\theta < 1$, i.e. the effective link probability is *sublinear* in the degree of the target node. We note however, that the duplication model [75, 77, 78], in which every new node is attached to a random node and to each of its neighbor with probability p , displays at the same time degree correlations and power-law degree distribution.

We also find that the model spontaneously generates communities during the evolution of the system. To quantify how pronounced communities are, we use a measure

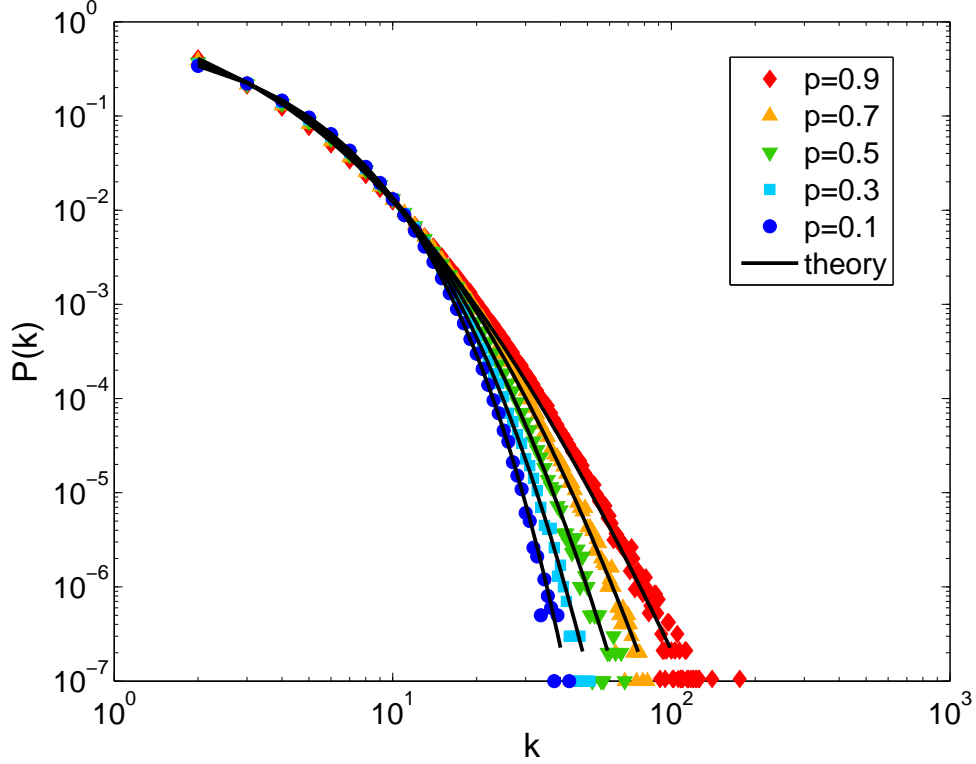


Figure 2.7: Degree distributions of the basic model, for different values of the parameter p . The continuous lines indicate the theoretical predictions of Eq. 2.10, the symbols the distributions obtained from numerical simulations of the model. The figure is obtained by performing 100 realizations of networks of size $n = 100\,000$.

called *embeddedness*, which estimates how strongly nodes are attached to their own cluster. Embeddedness, which we shall indicate with ξ , is defined as follows:

$$\xi = \frac{1}{n_c} \sum_c \frac{k_{\text{in}}^c}{k_{\text{tot}}^c}, \quad (2.13)$$

where k_{in}^c and k_{tot}^c are the internal and the total degree of community c and the sum runs over all n_c communities of the network. If the community structure is strong, most of the neighbours of each node in a cluster will be nodes of that cluster, so k_{in}^c will be close to k_{tot}^c and ξ turns out to be close to 1; if there is no community structure ξ is close to zero. However, one could still get values of embeddedness which are not too small, even in random graphs, which have no modular structure, as k_{in}^c might still be sizeable there. To eliminate such borderline cases, we introduce a new variable, the *node-based embeddedness*, that we shall indicate with ξ_n . It is based on the idea that for a node to be properly assigned to a cluster, it must have more neighbours in that cluster than in

any of the others. This leads to the following definition

$$\xi_n = \frac{1}{n} \sum_i \frac{k_{i,\text{in}} - k_{i,\text{ext}}^{\max}}{k_i}, \quad (2.14)$$

where $k_{i,\text{in}}$ is the number of neighbours of node i in its cluster, $k_{i,\text{ext}}^{\max}$ is the maximum number of neighbours of i in any one other cluster and k_i the total degree of i . The sum runs over all n nodes of the graph. For a proper community assignment, the difference $k_{i,\text{in}} - k_{i,\text{ext}}^{\max}$ is expected to be positive, negative if the node is misclassified. In a random graph, and for subgraphs of approximately the same size, ξ_n would be around zero. In a set of disconnected cliques (a clique being a subgraph where all nodes are connected to each other), which is the paradigm of perfect community structure, ξ_n would be 1.

In Figure 2.8a we show a heat map for ξ_n as a function of the two main variables of the model, the probability p and the number of edges per node m , which is half the average degree. Communities were detected with non-hierarchical Infomap [51] in all cases. Results obtained by applying the Louvain algorithm [49] (taking the most granular level to avoid artefacts caused by the resolution limit [82]) yield a consistent picture. All networks are grown until $n = 50\,000$ nodes. We see that large values of ξ_n are associated to the bottom left portion of the diagram, corresponding to high values of the probability of triadic closure and to low values of degree. So, a high density of triangles ensures the formation of clusters, provided the network is sufficiently sparse. In Figure 2.8b we present an analogous heat map for the average clustering coefficient C , which is defined [62] as

$$C = \frac{1}{n} \sum_i \sum_{j,k} \frac{a_{ij}a_{jk}a_{ki}}{k_i(k_i - 1)} \quad (2.15)$$

where a_{ij} is the element of the adjacency matrix of the graph and k_i is again the degree of node i . Figure 2.8b confirms that C is the largest when p is high and m is low, as expected.

The mechanism of formation and evolution of communities is schematically illustrated in Figure 2.9. When the first denser clumps of the network are formed (a), out of random fluctuations in the density of triangles newly added nodes are more likely to close triads within the protocusters than between them (b). As more nodes and links are added, the protocusters become larger and larger and their internal density of links becomes inhomogeneous, so there will be a selective triadic closure within the denser parts, which yields a separation into smaller clusters (c). This cycle of growing and splitting plays repeatedly along the evolution of the system.

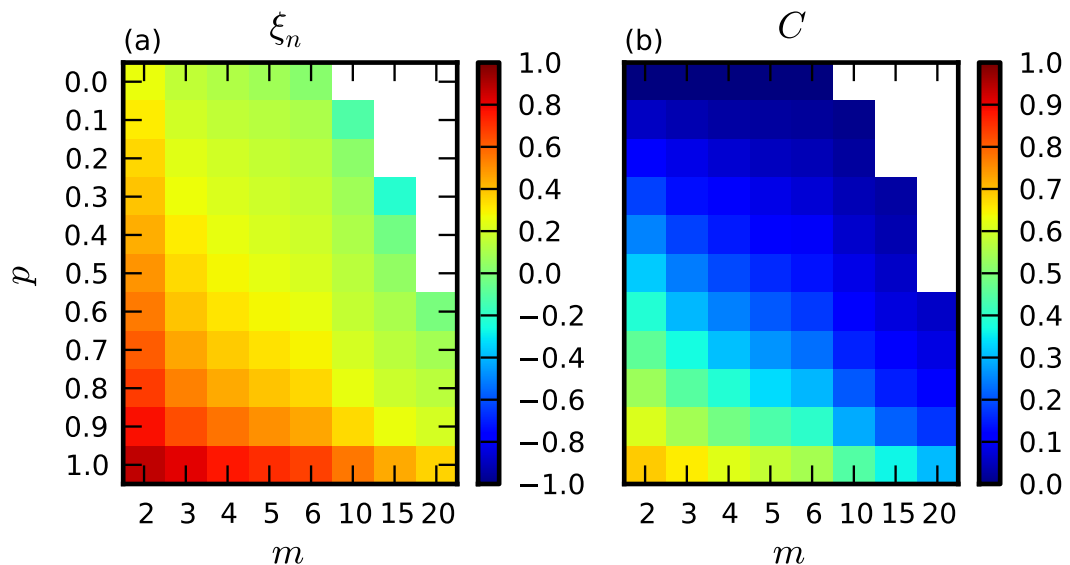


Figure 2.8: Heat map of node-based embeddedness (a) and average clustering coefficient (b) as a function of p and m for the basic model. Community structure (higher embeddedness and clustering coefficient) is pronounced in the lower left region when m is not too large (sparse graphs) and when the probability of triadic closure p is very high. For each pair of parameter values we report the average over 50 network realizations. The white area in the upper right corresponds to systems where a single community, consisting of the whole network, is found. Here one would get a maximum value 1 for ξ_n , but it is not meaningful, hence we discard this portion of the phase diagram, as well as in Figures 2.11 and 2.12.

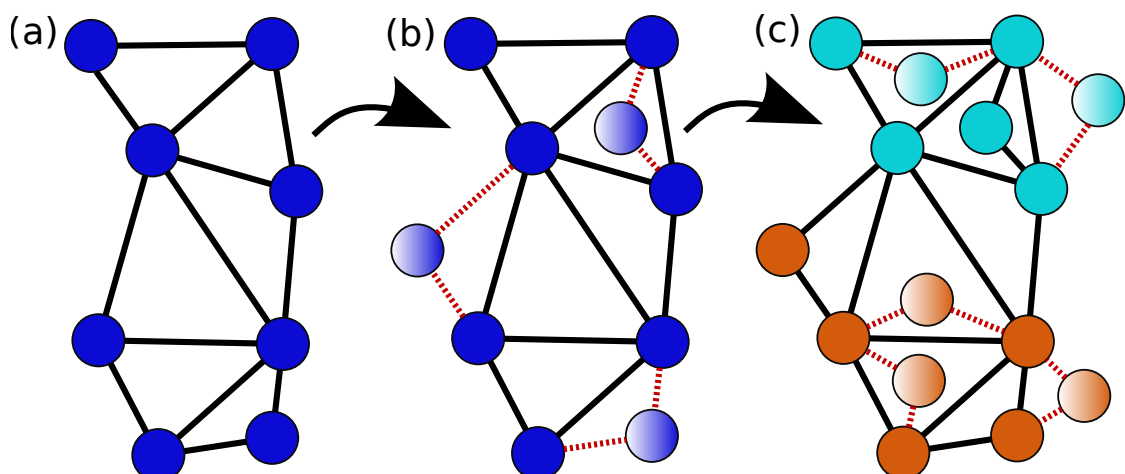


Figure 2.9: Schematic illustration of the formation and evolution of communities. Initial inhomogeneities in the link density make more likely the closure of triads in the denser parts, that keep growing until they become themselves inhomogeneous, leading to a split into smaller communities (different colors).

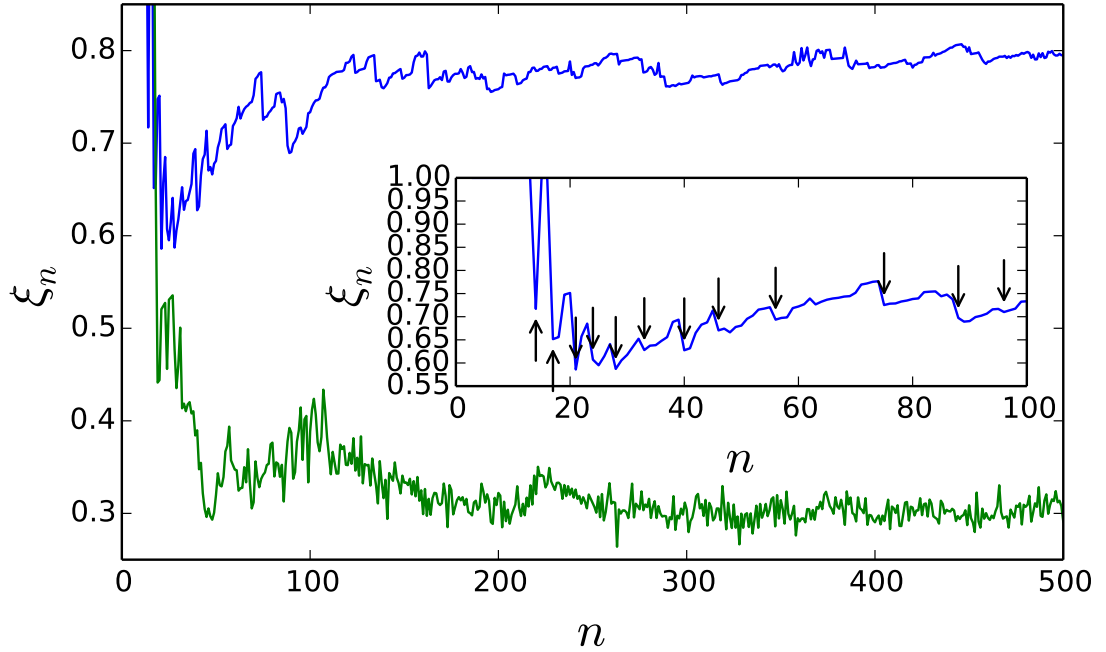


Figure 2.10: Evolution of node-based community embeddedness ξ_n along the growth of the network. The curves refer to the extreme cases of absence of triadic closure (lower curve), yielding a random graph without communities, and of systematic triadic closure (upper curve), yielding a graph with pronounced community structure. For the latter case, we magnify in the inset the initial portion of the curve, to highlight the sudden drops of ξ_n , indicated by the arrows, which correspond to the breakout of clusters into smaller ones.

In Figure 2.10 we show the time evolution of the node-based embeddedness ξ_n during the growth of the system, until 500 nodes are added to the network, $m = 2$. We consider the two extreme situations $p = 0$, corresponding to the absence of triadic closure and $p = 1$, where both links close a triangle every time and there is no additional noise. In the first case (green line), after a transient, ξ_n sets to a low value, with small fluctuations; in the case with pure triadic closure, instead, the equilibrium value is much higher, indicating strong community structure, and fluctuations are modest. In contrast with the random case, we recognize a characteristic pattern, with ξ_n increasing steadily and then suddenly dropping. The smooth increase of ξ_n signal that the communities are growing, the rapid drop that a cluster splits into smaller pieces: in the inset such breakouts are indicated by arrows. Embeddedness drops when clusters break up because the internal degrees $k_{i,\text{in}}$ of the nodes of the fragments in Eq. 2.14 suddenly decrease, since some of the old internal neighbours belong to a different community, while the values of $k_{i,\text{ext}}^{\text{max}}$ are typically unaffected.

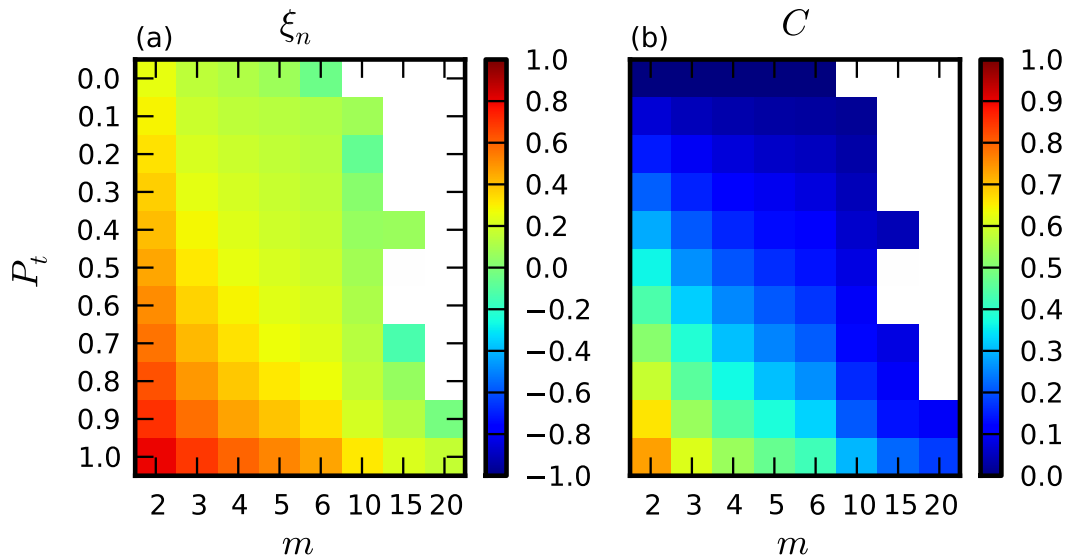


Figure 2.11: Heat map of node-based embeddedness (a) and average clustering coefficient (b) as a function of P_t and m for the model by Holme and Kim [65]. For each pair of parameter values we report the average over 50 network realizations. The white area in the upper right corresponds to systems where a single community, consisting of the whole network, is found, which is not interesting. The diagrams look qualitatively similar to that of the basic model (Figure 2.8), with highest embeddedness and clustering coefficient in the lower left region.

To show that the emergence of the community structure by triadic closure is not model-dependent we will analyse two other different models of evolving network enforcing triadic closure.

The first one is the model by Holme and Kim [65], already introduced in Section 2.2, where we have a growing networks in which new nodes sets a link with an existing node, chosen with a probability proportional to the degree of the latter, and other $m - 1$ by attaching with a probability P_t to a random neighbour of the node which received the most recent preferentially-attached link, otherwise with a probability $1 - P_t$ to another node chosen with preferential attachment. By varying P_t it is possible to tune the level of clustering into the network, while the degree distribution is the same as in the BA model, i.e. a power law with exponent -3 , for any value of P_t . In Figure 2.11 we show the same heat map as in Figure 2.8 for this model, where we now report the probability P_t on the y-axis. Networks are again grown until $n = 50\,000$ nodes. The picture is very similar to what we observe for the basic model.

The second one is the networked society model by Marsili et al. [83], which is not based on a growth process. The model is a model for temporal networks [84], in which the links are created and destroyed on the fast time scale while the number of nodes remains

constant. The starting point is a random graph with n nodes. Then, three processes take place, at different rates:

1. any existing link vanishes (rate λ);
2. a new link is created between a pair of nodes, chosen at random (rate η);
3. a triangle is formed by joining a node with a random neighbour of one of his neighbours, chosen at random (rate ξ_M).

In our simulations we start from a random network of $n = 50\,000$ nodes with average degree 10. The three rates λ , η and ξ_M can be reduced to two independent parameters, since what counts is their relative size. The number of links deleted at each iteration is proportional to λM , where M is the number of links of the network, while the number of links created via the two other processes is proportional to ηn and $\xi_M n$, respectively. The number of links M varies in time but in order to get a non-trivial stationary state, one should reach an equilibrium situation where the numbers of deleted and created links match. A variety of scenarios are possible, depending on the choices of the parameters. For instance, if ξ_M is set equal to zero, there are no triads, and what one gets at stationarity is a random graph with average degree $2\eta/\lambda$. So, if $\eta \ll \lambda$, the graph is fragmented into many small connected components. In one introduces triadic closure, the clustering coefficient grows with ξ_M if the network is fragmented, as triangles concentrate in the connected components. Moreover the model can display a veritable first order phase transition and in a region of the phase diagram displays two stable phases: one corresponding to a connected network with large average clustering coefficient and the other one corresponding to a disconnected network. Interestingly, if there is a dense single component, the clustering coefficient decreases with ξ_M . The degree distribution can follow different patterns too: it is Poissonian in the diluted phase, where the system is fragmented, and broad in the dense phase, where the system consists of a single component with an appreciable density of links. In Figure 2.12 we show the analogous heat map as in Figure 2.8 and 2.11, for the two parameters λ and ξ_M . The third parameter $\eta = 1$. We consider only configurations where the giant component covers more than a half of the nodes of the network. The diagrams are now different because of the different role of the parameters, but the picture is consistent nevertheless. The clustering coefficient C is highest when the ratio of λ and ξ_M lies within a narrow range, yielding a sparse network with a giant component having a high density of triangles and a corresponding presence of strong communities.

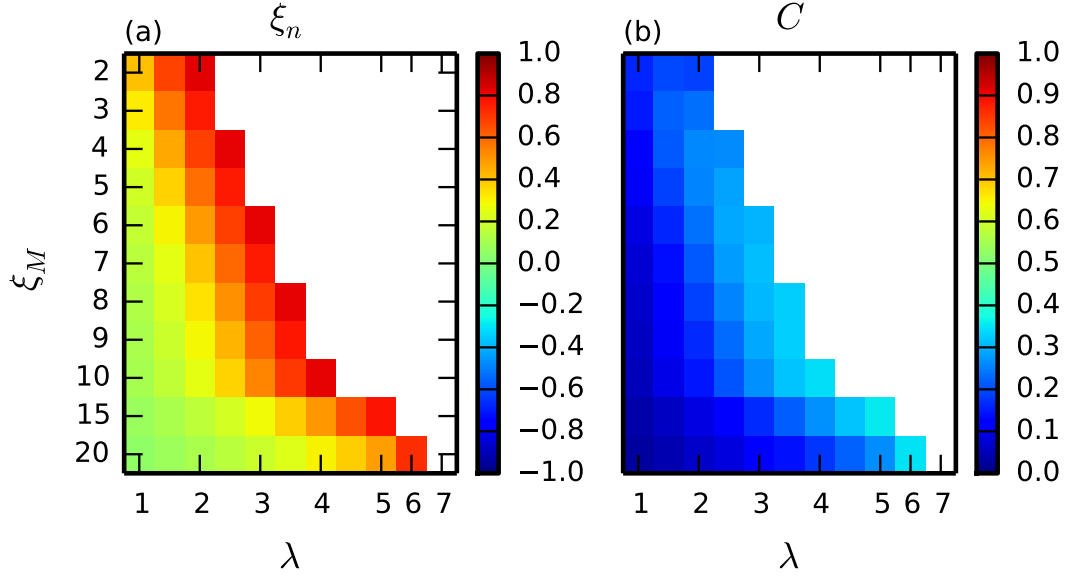


Figure 2.12: Heat map of node-based embeddedness (a) and average clustering coefficient (b) as a function of the rates λ and ξ_M for the model by Marsili et al. [83] ($\eta = 1$). For each pair of parameter values we report the average over 50 network realizations. The white area in the upper right corresponds to systems where a single community, consisting of the whole network, is found, which is not interesting. These diagrams have better communities (higher embeddedness and clustering coefficient) towards the upper right, different from those in Figures 2.8 and 2.11, because of the different meaning and effect of the parameters. However, there is a strong correspondence between high clustering coefficient and strong community structure, as in the other models.

2.3.2 The triadic closure model including fitness of the nodes

We now introduce a variant of the basic model, where the link attractivity depends on some intrinsic fitness of the nodes [85, 86]. We will assume that the nodes are not all equal and assign to each node i a fitness η_i representing the ability of a node to attract new links. We have chosen to parametrize the fitness with a parameter $\beta > 0$ by setting

$$\eta_i = e^{-\beta\epsilon_i}, \quad (2.16)$$

with ϵ chosen from a distribution $g(\epsilon)$ and β representing a tuning parameter of the model. We take

$$g(\epsilon) = (1 + \nu)\epsilon^\nu, \quad (2.17)$$

with $\epsilon \in (0, 1)$. When $\beta = 0$ all the fitness values are the same, when β is large small differences in the ϵ_i cause large differences in fitness. For simplicity we assume that the fitness values are quenched variables assigned once for all to the nodes. As in the basic model without fitness, the starting point is a small connected network of n_0 nodes and $m_0 \geq m$ links. The model contains two ingredients:

- *Growth.* At time t a new node is added to the network with $m \geq 2$ links.
- *Proximity and fitness bias.* The probability to attach the new node to node i_1 depends on the order in which links are added.

The first link of the new node is attached to a random node i_1 of the network with probability proportional to its fitness. The probability that the new node is attached to node i_1 is then given by

$$\Pi^{[0]}(i_1) = \frac{\eta_{i_1}}{\sum_j \eta_j}. \quad (2.18)$$

For $m = 2$ the second link is attached to a node of the network chosen according to its fitness, as above, with probability $1 - p$, while with probability p it is attached to a node chosen randomly between the neighbours of the node i_1 with probability proportional to its fitness. Therefore in the first case the probability to attach to a node $i_2 \neq i_1$ is given by

$$\Pi^{[0]}(i_2) = \frac{\eta_{i_2}(1 - \delta_{i_1, i_2})}{\sum_{j \neq i_1} \eta_j}, \quad (2.19)$$

with δ_{i_1, i_2} indicating the Kronecker delta, while in the second case the probability $\Pi^{[1]}(i_2)$ that the new node links to node i_2 is given by

$$\Pi^{[1]}(i_2) = \frac{\eta_{i_2} a_{i_1, i_2}}{\sum_j \eta_j a_{i_1, j}}, \quad (2.20)$$

where a_{ij} indicates the matrix element (i, j) of the adjacency matrix of the network.

- *Further edges.* For $m > 2$, further edges are added according to the ‘second link’ rule in the previous point. With probability p an edge is added to a neighbour of the *first* node i_1 , not already attached to the new node, according to the fitness rule. With probability $1 - p$, a link is set to any node in the network, not already attached to the new node, according to the fitness rule.

For simplicity we shall consider here the case $m = 2$. The probability that a node i acquires a new link at time t is given by

$$\frac{e^{-\beta\epsilon_i}}{t} \left[(2-p) + p \sum_j \frac{a_{ij}}{\sum_r \eta_r a_{jr}} \right]. \quad (2.21)$$

Similarly to the case without fitness, here we will assume, supported by simulations, that

$$\Theta_i = p \sum_j \frac{\eta_j a_{ij}}{\sum_r \eta_r a_{jr}} \simeq ck_i^{\theta(\epsilon)}, \quad (2.22)$$

where, for every value of p , $\theta = \theta(\epsilon) \leq 1$ and $c = c(\epsilon)$.

We can write the master equation for the average number $n_{k,\epsilon}(t)$ of nodes of degree k and energy ϵ at time t , as

$$\begin{aligned} n_{k,\epsilon}(t+1) &= n_{k,\epsilon}(t) \\ &+ \frac{e^{-\beta\epsilon}[2-p+c(\epsilon)(k-1)^\theta]}{t} n_{k-1,\epsilon}(t)(1-\delta_{k,2}) \\ &- \frac{e^{-\beta\epsilon}[2-p+c(\epsilon)k^\theta]}{t} n_{k,\epsilon}(t) + \delta_{k,2}g(\epsilon). \end{aligned} \quad (2.23)$$

In the limit of large values of t we assume that $n_{k,\epsilon}/t \rightarrow P^\epsilon(k)$, and therefore we find that the solution for $P^\epsilon(k)$ is given by

$$\begin{aligned} P^\epsilon(k) &= C(\epsilon) \frac{1}{1 + e^{-\beta\epsilon}[2-p+c(\epsilon)k^\theta]} \\ &\times \prod_{j=1}^{k-1} \left\{ 1 - \frac{1}{1 + e^{-\beta\epsilon}[2-p+c(\epsilon)j^\theta]} \right\}, \end{aligned} \quad (2.24)$$

where $C(\epsilon)$ is the normalization factor. This expression for $\theta(\epsilon) < 1$ can be approximated in the continuous limit by

$$P^\epsilon(k) \simeq D(\epsilon) \frac{e^{-(k-1)G[k-1,\epsilon,\theta(\epsilon),c(\epsilon)]}}{1 + e^{-\beta\epsilon}[2-p+c(\epsilon)k^\theta]}, \quad (2.25)$$

where $D(\epsilon)$ is the normalization constant and $G(k, \epsilon, \theta, c)$ is given by

$$\begin{aligned} G(k, \epsilon, \theta, c) &= -\theta {}_2F_1\left(1, \frac{1}{\theta}, 1 + \frac{1}{\theta}, -\frac{ck^\theta}{2-p+e^{\beta\epsilon}}\right) \\ &\quad + \theta {}_2F_1\left(1, \frac{1}{\theta}, 1 + \frac{1}{\theta}, -\frac{ck^\theta}{2-p}\right) \\ &\quad + \log\left(1 - \frac{1}{1 + \frac{e^{\beta\epsilon}}{2-p+ck^\theta}}\right). \end{aligned} \quad (2.26)$$

When $\theta(\epsilon) = 1$, instead, we can approximate $P^\epsilon(k)$ with a power law, i.e.

$$P^\epsilon(k) \simeq D(\epsilon) \left[1 + e^{-\beta\epsilon} (2-p + c(\epsilon)k)\right]^{-\frac{\epsilon-\beta\epsilon}{c(\epsilon)} - 1}. \quad (2.27)$$

Therefore, the degree distribution $P(k)$ of the entire network is a convolution of the degree distributions $P^\epsilon(k)$ conditioned on the value of ϵ , i.e.

$$P(k) = \int d\epsilon P^\epsilon(k). \quad (2.28)$$

As a result of this expression, we found that the degree distribution can be a power law also if the network exhibits degree correlations and $\theta(\epsilon) < 1$ for every value of ϵ . Moreover we observe that for large values of the parameter β the distribution becomes broader and broader until a condensation transition occurs at $\beta = \beta_c$ with the value of β_c depending on both the parameters ν and p of the model. For $\beta > \beta_c$ successive nodes with maximum fitness (minimum value of ϵ) become ‘superhubs’, attracting a finite fraction of all the links, similarly to what happens in Ref. [85]. In Figure 2.13 we see the degree distribution of model, obtained via numerical simulations, for different values of β . The continuous lines, illustrating the theoretical behaviour, are well aligned with the numerical results, as long as $\beta < \beta_c$.

In Figure 2.14 we show the heat map of ξ_n and C for the model, as a function of the parameters p and β . The number of edges per node is $m = 2$, and the networks consist of 50 000 nodes. Everywhere in this work, we set the parameter $\nu = 6$. For $\beta = 0$ all nodes have identical fitness and the model reduces itself to the basic model. So we recover the previous results, with the emergence of communities for sufficiently large values of the probability of triadic closure p , following a large density of triangles in the system. The situation changes dramatically when β starts to increase, as we witness a progressive weakening of community structure, while the clustering coefficient keeps growing, which appears counterintuitive. In the analogous diagrams for $m = 5$, we see that this pattern holds, though with a weaker overall community structure and lower values of the clustering coefficient.

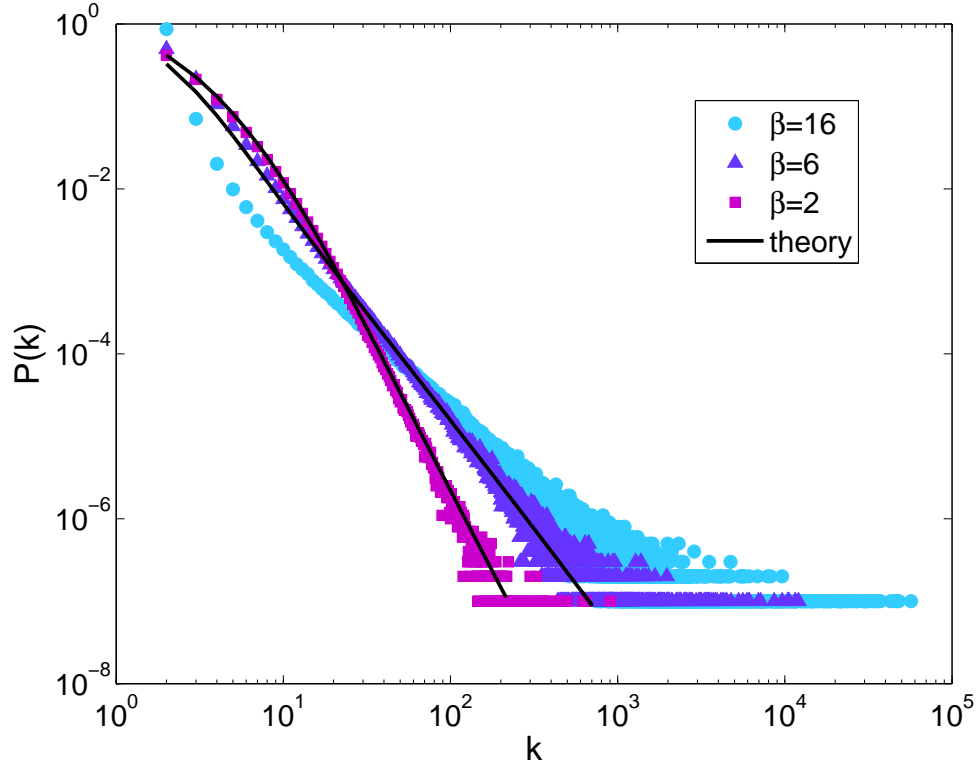


Figure 2.13: Degree distribution of the model with fitness, for three values of the parameter β , which indicates the heterogeneity of the distribution of the fitness of the nodes. Symbols stand for the results obtained by building the network via simulations, continuous lines for our analytical derivations. The figure is obtained by performing 100 realizations of networks of size $n = 100\,000$ with $\nu = 6$.

When β is sufficiently large, communities disappear, despite the high density of triangles. To check what happens, we compute the probability distribution of the scaled link density $\tilde{\rho}$ and the node-based embeddedness ξ_n of the communities of the networks obtained from 100 runs of the model, for three different values of β : 0, 6 and 20. All networks are grown until 100 000 nodes. The scaled link density $\tilde{\rho}$ of a cluster is defined [87] as

$$\tilde{\rho} = \frac{2l_c}{n_c - 1}, \quad (2.29)$$

where l_c and n_c are the number of internal links and of nodes of cluster c . If the cluster is tree-like, $\tilde{\rho} \approx 2$, if it is clique-like it $\tilde{\rho} \approx n_c$, so it grows linearly with the size of the cluster. The distributions of ξ_n and $\tilde{\rho}$ are shown in Figure 2.16. They are peaked, but the peaks undergo a rapid shift when β goes from 0 to 20. The situation resembles what one usually observes in first-order phase transitions. The embeddedness ends up peaking

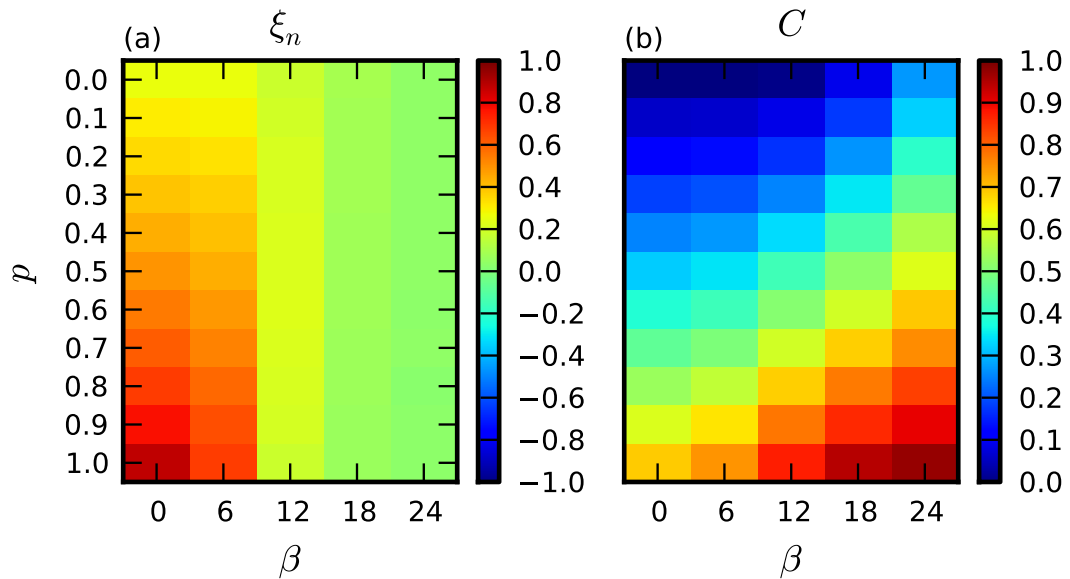


Figure 2.14: Heat map of node-based embeddedness (a) and average clustering coefficient (b) as a function of the probability of triadic closure p and the heterogeneity parameter β of the fitness distribution of the nodes, for the model with fitness. The number of new edges per node is $m = 2$. For each pair of parameter values we report the average over 50 network realizations. When $\beta = 0$ we recover the basic model, without fitness. We see the highest values of embeddedness in the lower left, while highest values of the clustering coefficient are in the lower right. When β increases, we see a drastic change of structure in contrast to the previous pattern: communities disappear, whereas the clustering coefficient gets higher.

at low values, quite distant from the maximum 1, while the scaled link density eventually peaks sharply at 2, indicating that the subgraphs are effectively tree-like.

What kind of objects are we looking at? To answer this question, in Figure 2.17 and 2.18 we display two pictures of networks obtained by the fitness model, for $\beta = 0$ and $\beta = 20$, respectively. The number of nodes is 2000, and the number of edges per node $m = 2$. The probability of triadic closure is $p = 0.97$, as we want a very favourable scenario for the emergence of structure. The subgraphs found by our community detection method (non-hierarchical Infomap, but the Louvain method yields a similar picture) are identified by the different colors. The insets show an enlarged picture of the subgraphs, which clarify the apparent puzzle delivered by the previous diagrams. For the basic model $\beta = 0$ (Figure 2.17), the subgraphs are indeed communities, as they are cohesive objects which are only loosely connected to the rest of the graph. The situation remains similar for low values of β . However, for sufficiently high β (Figure 2.18), a phenomenon of link condensation takes place, with a few superhubs attracting most of

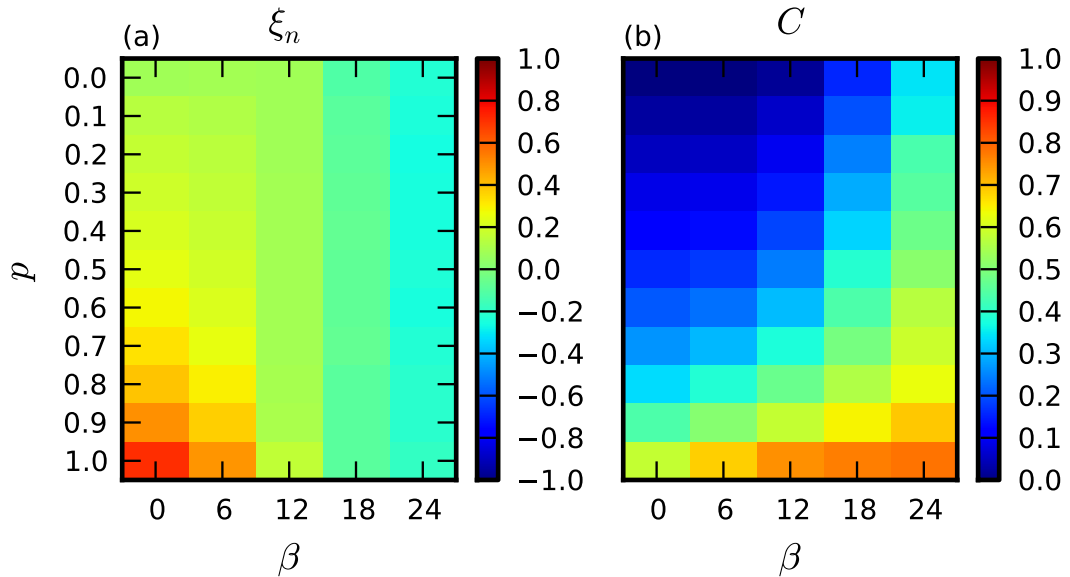


Figure 2.15: Same as Figure 2.14, but for $m = 5$. The picture is consistent with the case $m = 2$, but communities are less pronounced.

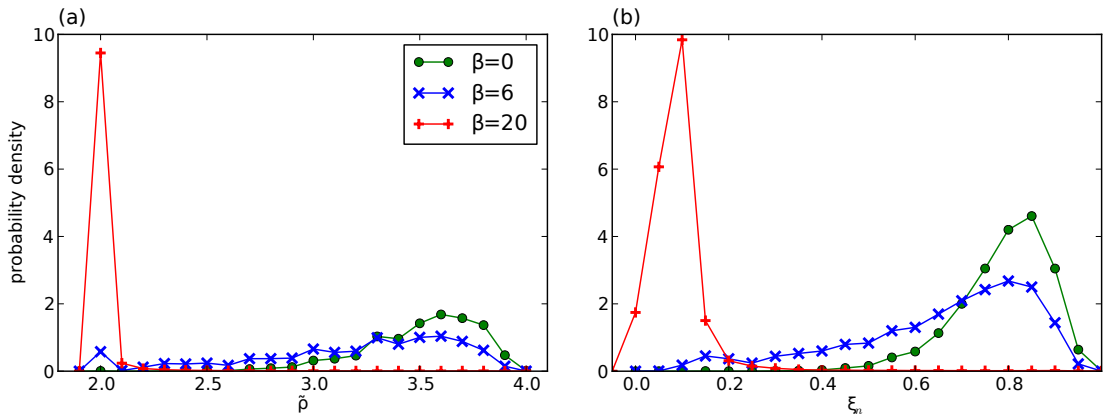


Figure 2.16: Probability distributions of the scaled link density $\tilde{\rho}$ (left) and node-based embeddedness ξ_n (right) of the communities of the fitness model, for $m = 2$ and $\beta = 0, 6, 20$. For each β -value we derived 100 network realizations, each with 100 000 nodes. We see that at $\beta = 0$, the detected communities satisfy the expectations of good communities, while at $\beta = 20$ they do not.

the links of the network [85]. Most of the other nodes are organized in groups which are “shared” between pairs (for $m = 2$, more generally m -ples) of superhubs (see figure). The community embeddedness is low because there are always many links flowing out of the subgraphs, towards superhubs. Besides, since the superhubs are all linked to each other, this generates high clustering coefficient for the subgraphs, as observed in

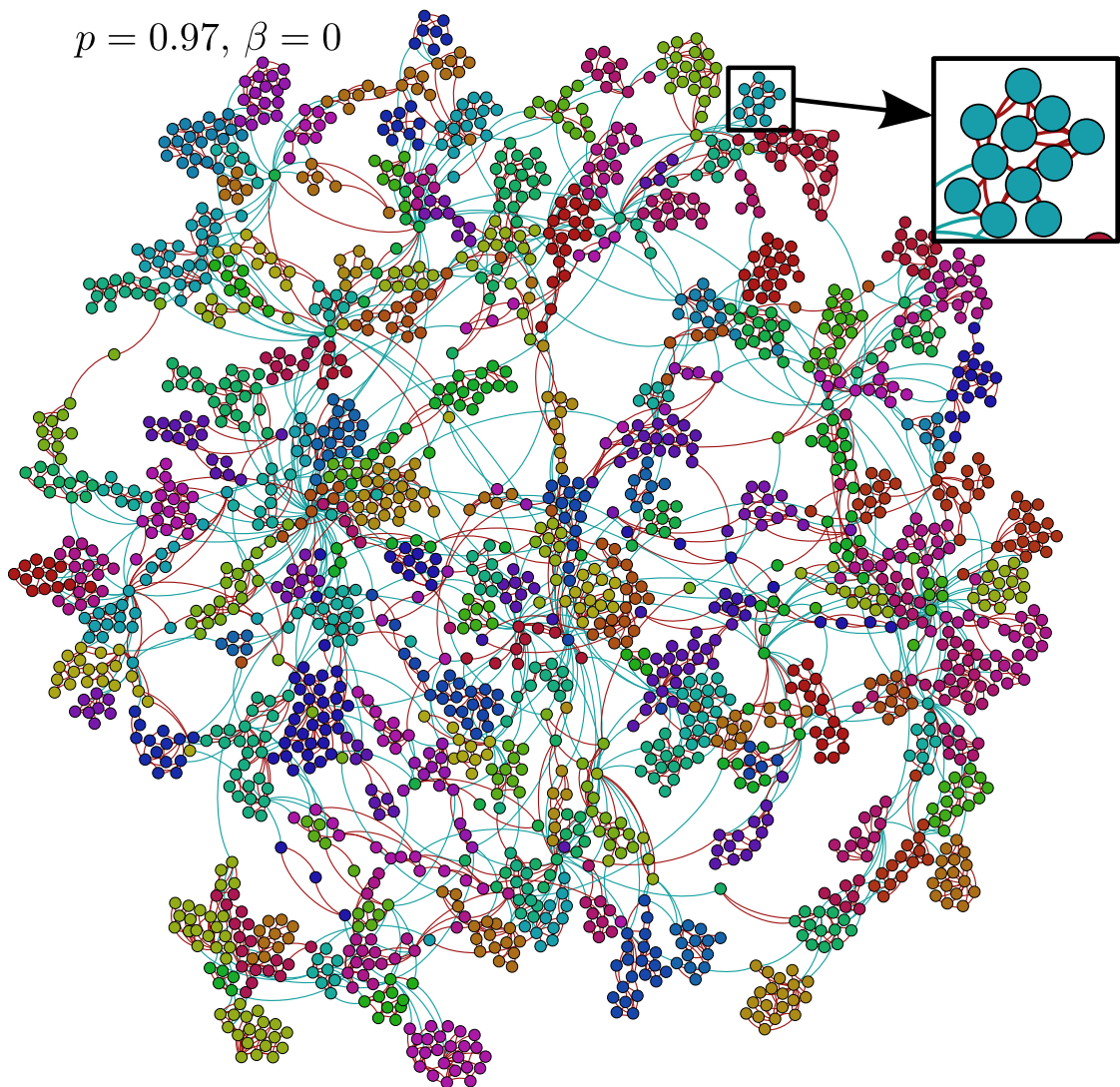


Figure 2.17: Picture of a network with 2000 nodes generated by the fitness model, for $p = 0.97$, $m = 2$ and $\beta = 0$. Since $\beta = 0$ fitness does not play a role and we recover the results of the basic model. Colors indicate communities as detected by the non-hierarchical Infomap algorithm [51].

Figure 2.14 and 2.15. In fact, the clustering coefficient for the non-hubs attains the maximum possible value of 1, as their neighbours are nodes which are all linked to each other.

These results show that triadic closure is alone capable to generate systems with all the characteristic properties of complex networks, from fat-tailed degree distributions to high clustering coefficients and strong community structure. Communities emerge naturally via triadic closure, which tend to generate cohesive subgraphs around portions of the system that happen to have higher density of links, due to stochastic fluctuations.

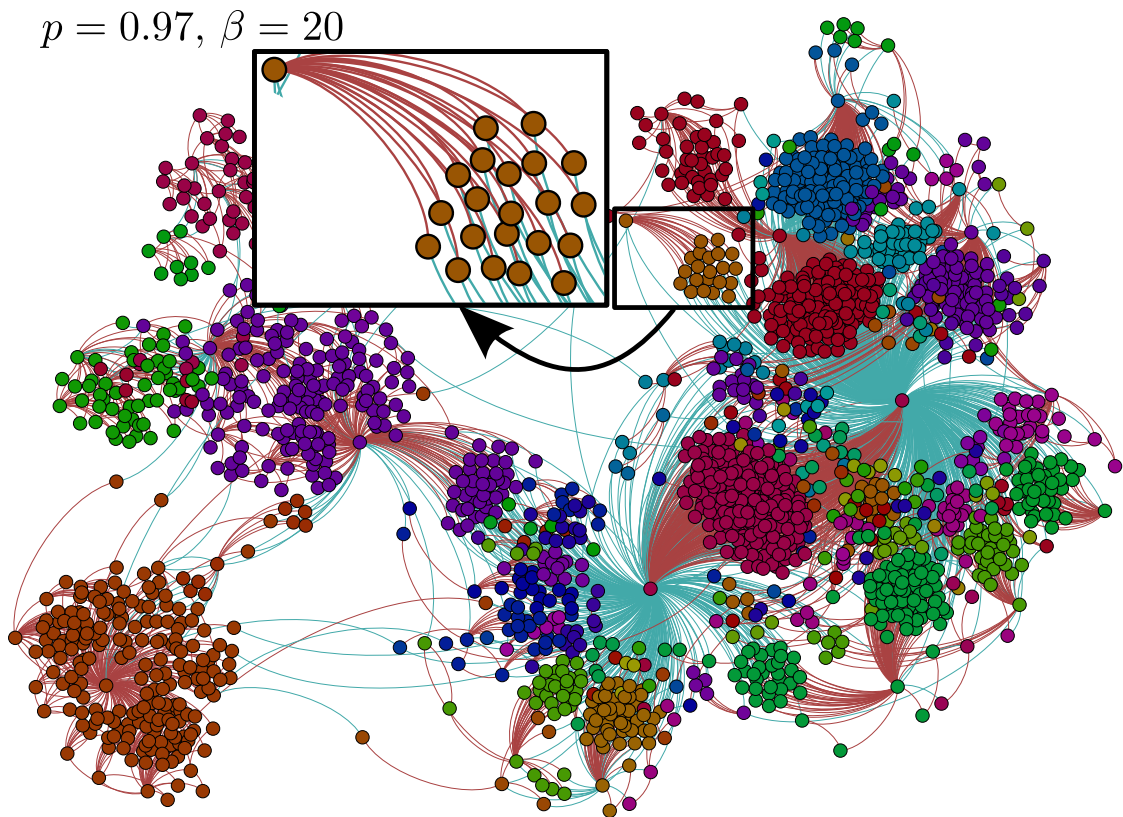


Figure 2.18: Picture of a network with 2000 nodes generated by the fitness model, for $p = 0.97$, $m = 2$ and $\beta = 20$. The growing process is the same as in Figure 2.17, but the addition of fitness changes the structural organization of the network. As seen in the inset, node aggregations form around hub nodes with high fitness. Looking at the inset we see that such aggregations do not satisfy the typical requirements for communities: they are internally tree-like, and there are more external edges (blue or light gray) than internal (red or dark gray) touching its nodes. In particular, internal edges only go from regular nodes to superhubs.

When clusters become sufficiently large, their internal structure exhibits in turn link density inhomogeneities, leading to a progressive differentiation and eventual separation into smaller clusters (separation in the sense that the density of links between the parts is appreciably lower than within them). This occurs both in the basic version of network growth model based on triadic closure, and in more complex variants. The strength of community structure is the higher, the sparser the network and the higher the probability of triadic closure. We have also introduced a new variant, in that link attractivity depends on some intrinsic appeal of the nodes, or fitness. Here we have seen that, when the distribution of fitness is not too heterogeneous, community structure still emerges, though it is weaker than in the absence of fitness. By increasing the heterogeneity of the fitness distribution, instead, we observe a major change in the structural organization of the network: communities disappear and are replaced by special subgraphs, whose

nodes are connected only to superhubs of the network, i.e. nodes attracting most of the links. Such structural phase transition is associated to very high values of the clustering coefficient.

Chapter 3

Characterisation of mesoscopic structures in multiplex networks

3.1 Communities in multiplex networks

In this Chapter we will focus on the mesoscopic structure of multiplex networks, networks that are formed by a set of N nodes interacting through M different layers, as discussed in 1.2. Indeed it has been observed that the communities on different layers of a multiplex network typically overlap among each others, forming mesoscale structures that span across the different layers. These *multiplex communities* have an intrinsic multiplex nature and to characterise their organisation and their emergence is central for generalising the concept of community to multilayer networks [88, 89].

In the following Section we will discuss a methodology to characterise the correlations of multiplex networks at the mesoscopic scale, and to use this information in order to build a network between the layers of multiplex datasets. In particular we propose an information theory measure $\tilde{\Theta}^S$, able to define similarities between the layers of a multiplex respect to their mesoscopic structures. This similarity is more significant when groups of nodes densely connected with each others are simultaneously present on different layers, forming overlapping communities. We will apply the proposed methodology to characterise the American Physical Society (APS) Collaboration Multiplex Networks extracted from the APS dataset [90], that is formed by the authors of the APS papers, and by layers corresponding to the Physics and Astronomy Classification Scheme (PACS) codes [91]: two authors are linked on layer α if they have co-authored a paper with PACS code corresponding to layer α and since the PACS codes are organized in hierarchical

levels we will see that it is possible to construct two APS Collaboration Multiplex Networks corresponding to layers describing either the first or the second level of the PACS hierarchy.

In Section 3.3 we address the problem of understanding how the multiplex community structure can emerge in real world collaboration multiplex networks. We will propose a simple model to explain the appearance, coexistence and co-evolution of communities at the different layers of a multiplex. Our hypothesis is that the formation of communities in collaboration networks is an intrinsically multiplex process, which is the result of the interplay between an intra-layer triadic closure process and an inter-layer triadic closure process. For instance, in the case of scientific collaborations, multiplex communities naturally arise from the fact that scientists may collaborate with other researchers in their principal field of investigation and with colleagues coming from other scientific disciplines. Analogously, actors can prefer either to specialise in a specific genre or instead to explore different (sometimes dissonant) genres, and these two opposite behaviours undoubtedly have an impact on the kind of meso-scale structures observed on each of the layers of the system. The generative model we propose here mimics two of the most basic processes that drive the evolution of collaborations in the real world, namely intra- and inter-layer triadic closure, and is able to explain the appearance of overlapping modular organisations in multi-layer systems. We will show that the model is able to reproduce the salient micro-, meso- and macro-scale structure of different real-world collaboration networks including the APS Collaboration Multiplex Network and the Multiplex Co-starring Network of actors obtained from the Internet Movie Database (IMDb).

3.2 Information theory to characterise multiplex mesoscopic structures: the $\tilde{\Theta}^S$ indicator

Our goal here is to construct an information theory indicator function $\tilde{\Theta}^S$ to characterise the similarity in the mesoscopic structure of the layers of a multiplex network. This indicator function is based on the entropy of network ensembles [92–95], a quantity which plays a key role when inference problems are addressed using an unbiased information theory approach [94, 95]. In this Section we define how the indicator function $\tilde{\Theta}^S$ is defined. We consider a multiplex network formed by N nodes $i = 1, 2, \dots, N$ and M layers $\alpha = 1, 2, \dots, M$. The structure of the multiplex network is characterised by M adjacency matrices \mathbf{a}^α of elements $a_{ij}^\alpha = 1$ if node i is connected to node j in layer α , or $a_{ij}^\alpha = 0$ otherwise. We indicate with k_i^α the degree of a node i on layer α , i.e. the number of neighbours that node i has on α . The nodes having degree $k_i^\alpha = 0$ in layer α , are the isolated nodes, i.e. nodes that are not connected to any other node in the layer α , also called [20] in the context of multilayer networks “inactive” nodes in layer α . Conversely all the nodes with $k_i^\alpha > 0$ are called “active” nodes in layer α .

We assume that each node i of layer α has a characteristic $q_i^\alpha \in \{1, \dots, Q^\alpha\}$. The quantity q_i^α can for example indicate the community to which the node i belongs. More in general q_i^α can represent any feature of the nodes in layer α . Starting from this information we can classify the nodes in P^α classes $p_i^\alpha \in \{1, \dots, P^\alpha\}$ which take into account at the same time the information about the degree of the nodes and their characteristic q_i^α . This is the minimal assumption to capture the structure of networks with communities induced by the characteristics $q^\alpha = \{q_i^\alpha\}_{i=1,2,\dots,N}$, and strong heterogeneities in the degree. Considering only the partition induced by the characteristics would imply that in the network we do not consider the structure induced by the degrees, which is clearly not a viable option for broadly distributed networks.

Including other features of the nodes to define node classes could be a viable option. In this case the characteristics q^α will take into account different features which might depend on the specific network under consideration. Therefore here we take the class p_i^α to be a function of degree k_i^α and of the characteristic q_i^α , i.e. $p_i^\alpha = f(k_i^\alpha, q_i^\alpha)$. The block structure of the network induced by the classes $p_i^\alpha = f(k_i^\alpha, q_i^\alpha)$ is described by the matrices \mathbf{e}^α of elements $e^\alpha(p, p')$ indicating the total number of links on the layer α between nodes of class p and nodes of class p' . We define the entropy $\Sigma_{k^\alpha, q^\alpha}$ [92–95] of a layer α as the logarithm of the number of graphs preserving the block structure \mathbf{e}^α in a given layer. By considering the number of graphs preserving a given block structure, we have that this entropy takes the simple expression,

$$\Sigma_{k^\alpha, q^\alpha} = \log \left[\prod_{p < p'} \binom{n_p^\alpha n_{p'}^\alpha}{e^\alpha(p, p')} \prod_p \binom{n_p^\alpha (n_p^\alpha - 1)/2}{e^\alpha(p, p)} \right], \quad (3.1)$$

where

$$e^\alpha(p, p') = \sum_{i, j} a_{ij}^\alpha \delta [p_i^\alpha(k_i^\alpha, q_i^\alpha), p] \delta [p_j^\alpha(k_j^\alpha, q_j^\alpha), p'], \quad (3.2)$$

for $p \neq p'$, and $e(p, p), n(p)$ given respectively by

$$e^\alpha(p, p) = \sum_{i < j} a_{ij}^\alpha \delta [p_i^\alpha(k_i^\alpha, q_i^\alpha), p] \delta [p_j^\alpha(k_j^\alpha, q_j^\alpha), p], \quad (3.3)$$

and

$$n_p^\alpha = \sum_i \delta [p_i^\alpha(k_i^\alpha, q_i^\alpha), p], \quad (3.4)$$

with $\delta[x, y]$ indicating the Kronecker delta. The entropy $\Sigma_{k^\alpha, q^\alpha}$ is a measure to assess how much information is encoded in the constraint imposed to the network i.e. the block structure \mathbf{e}^α . The smaller is the entropy the smaller is the number of networks that share the block structure \mathbf{e}^α . Therefore the smaller is the entropy of an ensemble the larger is the level of information encoded by the constraint. If for a given assignment of the characteristics $\{q_i^\alpha\}$ the entropy is much smaller than in a random hypothesis (when the characteristics are reshuffled randomly between the nodes), then the network structure reflects the characteristic assignment $\{q_i^\alpha\}$ and thus the characteristics $\{q_i^\alpha\}$ capture relevant information respect to the network structure. Following this argument the quantity Θ proposed in [95], which is based on the entropy of network ensembles, has been shown to be an unbiased indicator able to quantify the specificity of a generic layer α to the assignment q_i^α . This information theory quantity is defined as:

$$\Theta_{k^\alpha, q^\alpha} = \frac{E_\pi[\Sigma_{k^\alpha, \pi(q^\alpha)}] - \Sigma_{k^\alpha, q^\alpha}}{\sqrt{E_\pi[(\Sigma_{k^\alpha, \pi(q^\alpha)} - E_\pi[\Sigma_{k^\alpha, \pi(q^\alpha)}])^2]}}, \quad (3.5)$$

where $E_\pi[\dots]$ is the expected value over random uniform permutations $\pi(q^\alpha)$ of the node characteristics q^α in layer α .

Here we propose to use this quantity to compare the similarity between the different layers in a multiplex network. Indeed we can consider the characteristics q^β of the nodes in layer β as an induced feature of nodes in layer α and measure by the corresponding indicator $\Theta_{k^\alpha, q^\beta}$ how much information the characteristics q^β contain respect to the node structure of layer α . In particular the indicator $\Theta_{k^\alpha, q^\beta}$ is given by

$$\Theta_{k^\alpha, q^\beta} = \frac{E_\pi[\Sigma_{k^\alpha, \pi(q^\beta)}] - \Sigma_{k^\alpha, q^\beta}}{\sqrt{E_\pi[(\Sigma_{k^\alpha, \pi(q^\beta)} - E_\pi[\Sigma_{k^\alpha, \pi(q^\beta)}])^2]}}. \quad (3.6)$$

Therefore $\Theta_{k^\alpha, q^\beta}$ measures the specificity of the layer α respect to the particular set q^β , which is the assignment of the characteristics of the nodes on layer β .

When one considers a single layer, the entropy is independent on the choice adopted for classifying isolated (inactive) nodes in layers belonging to multiplex networks. In fact, we can either group all the isolated nodes in a single class or each isolated node in a different class, and the entropy value given by Eq. 3.1 does not change because the isolated nodes have no links attached to them. Instead the indicator function $\Theta_{k^\alpha, q^\alpha}$ might depend on this choice because its construction involves several reshuffling of the characteristics of the nodes.

When comparing different layers of a multiplex network, the nodes that are active in one layer might not be active in another layer. Nevertheless, the information carried by the activity of the node might be significant. For example if two layers have very different activity patterns, it might occur that the nodes inactive in one layer form a well defined cluster in the other layer resulting in a very significant information that is important to capture. Therefore to distinguish between nodes active and inactive in a layer it is a very convenient choice to classify all the inactive nodes in one layer under a given common characteristic. A similar type of argument can be made about connected clusters of small sizes, which are “quasi-isolated” as the nodes belonging to connected clusters of size 2 or 3 etc. Depending on the number of such clusters it might be convenient to classify also nodes in connected components of size 2 or 3 etc. into given common characteristics as we will show in the next sections using the concrete examples of the APS Collaboration Multiplex Networks. Here, if not stated otherwise, we will consider the case in which the features q^α indicates the community of the nodes in layer α and the characteristic p_i^α takes a different value for each distinct pair (k_i^α, q_i^α) where $k_i^\alpha \neq 0$, while all the nodes with $k_i^\alpha = 0$ form another class of nodes.

In order to compare the level of information carried in layer α by the community structure in layer β , q^β , with the level of information carried by the proper community structure, q^α , we define the quantity

$$\tilde{\Theta}_{\alpha, \beta} = \frac{\Theta_{k^\alpha, q^\beta}}{\Theta_{k^\alpha, q^\alpha}}. \quad (3.7)$$

The quantity $\tilde{\Theta}_{\alpha, \beta}$ is a measure of how layer β is similar to α respect to the community

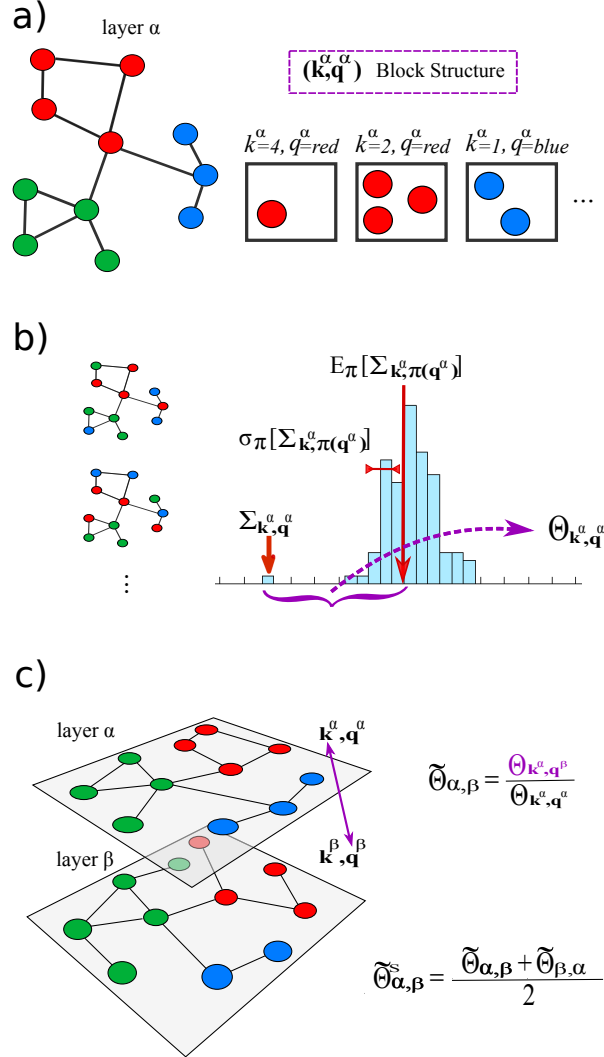


Figure 3.1: Diagram showing the method. Panel a): We consider a layer α in a multiplex network and we define the node classes $p^\alpha = (k^\alpha, q^\alpha)$, where k^α indicates the node degrees and q^α the node characteristics on the layer α . These classes induce a block structure in the network specified by the number of links between the nodes of each class and the number of links connecting the nodes in different classes. Panel b): The entropy $\Sigma_{k^\alpha, q^\alpha}$ given by Eq. 3.1 is calculated and compared with the entropy distribution obtained in a random hypothesis, by performing random uniform permutations $\pi(q^\alpha)$ of the characteristics q^α of the nodes and subsequently measuring the $\Sigma_{k^\alpha, \pi(q^\alpha)}$ values. The mean $E_\pi[\Sigma_{k^\alpha, \pi(q^\alpha)}]$ and standard deviation $\sigma_\pi[\Sigma_{k^\alpha, \pi(q^\alpha)}]$ of the entropy distribution is thus calculated. The indicator function $\Theta_{k^\alpha, q^\alpha}$ measures the difference between $\Sigma_{k^\alpha, q^\alpha}$ and $E_\pi[\Sigma_{k^\alpha, \pi(q^\alpha)}]$ in units of $\sigma_\pi[\Sigma_{k^\alpha, \pi(q^\alpha)}]$. Panel c): Given a second layer β , $\tilde{\Theta}_{\alpha, \beta}$ characterizes the information about the structure in layer α , carried by the characteristics of nodes in layer β . In order to define a symmetric indicator function of the similarity between the layers α and β we define the indicator $\tilde{\Theta}_{\alpha, \beta}^S$ that symmetrizes the indicator function $\tilde{\Theta}_{\alpha, \beta}$.

assignment \mathbf{q} . If $\tilde{\Theta}_{\alpha,\beta} = 1$ the community structure q^β , proper of layer β , carries the same level of information for the structure of layer α as the community structure q^α , proper of the layer α . It is important to notice that the matrix $\tilde{\Theta}$ in principle is not symmetric. We can construct the symmetric measure $\tilde{\Theta}_{\alpha,\beta}^S$ by symmetrising the quantity $\tilde{\Theta}_{\alpha,\beta}$ i.e. by defining

$$\tilde{\Theta}_{\alpha,\beta}^S = \frac{\tilde{\Theta}_{\alpha,\beta} + \tilde{\Theta}_{\beta,\alpha}}{2}. \quad (3.8)$$

This is a symmetric measure indicating how similar layer α and layer β are with respect to their community structure. In Figure 3.1 we give a schematic summary of the method used to construct the similarity measure $\tilde{\Theta}_{\alpha,\beta}^S$.

In a given multiplex network, we can then analyse the entire symmetric matrix $\tilde{\Theta}^S$ measuring the similarity between the community structure of the layers. This matrix characterises the entire multiplex network at the layer level, reducing the information about the network structures to one matrix of similarity between the layers.

In the following Section we will first test this measure on multiplex network benchmark models with non trivial community structure, then in the subsequent Section we will focus on characterising the APS Collaboration Multiplex Networks where the layers are the collaborations networks of scientists using different PACS numbers.

In this paper we are mostly concerned about similarities in the community structure of the layers of a multiplex network, nevertheless it has to be stressed that the proposed approach and similarity measure $\tilde{\Theta}_{\alpha,\beta}^S$ is general and it can be used by considering any available feature of the nodes related to the structure of the layers.

3.2.1 Testing $\tilde{\Theta}^S$ on Benchmark Models

In order to validate on a well defined multiplex architecture our similarity measure $\tilde{\Theta}^S$ respect to the community structures of different layers of a multiplex network, we have developed two benchmark models with communities. In particular we want to construct benchmark multiplex network models with a controlled level of overlap between the communities in different layers. Given in a generic multilayer the community assignment q^α of the nodes on each layer α , we define the community overlap as

$$O_c = \frac{2}{M(M-1)} \frac{1}{N} \max_{\{\pi\}} \left\{ \sum_{\alpha < \beta} \sum_{i=1}^N \delta \left[q_i^\alpha, \pi(q_i^\beta) \right] \right\}, \quad (3.9)$$

where M indicates the total number of layers and N indicates the total number of nodes, $\delta[x, y]$ indicates the Kronecker delta and the maximum is taken over all the permutations $\pi(q^\beta)$ of the label of the communities in layer β .

We define two benchmark models (see Figure 3.2) based respectively on the Girvan and Newman (GN) [13] model and on the Lancichinetti - Fortunato - Radicchi (LFR) model [96], which are very well established benchmarks for single networks with communities. The proposed benchmarks are designed to tune the overlap of communities between different layers of simple multiplex networks having respectively homogeneous or heterogeneous degree distribution and community size distribution.

For the first benchmark model, the Duplex Network GN model (DNGN) we construct a duplex network (a multiplex network made of two layers) in which each layer is formed by a GN network realisation. Therefore each of the layers is formed by N nodes divided into 4 equal size clusters of size N_c .

The network in each layer is a random network in which each node has a probability p_{in} to link to nodes of its same community and a probability p_{out} to link to nodes outside its community. In particular we have chosen p_{in} and p_{out} in order to have for each node, a mean degree $\langle k \rangle = 16$ and a mean number of links outside the community given by $\langle k_{out} \rangle = 4$. The layers generated in this way have a well defined community structure and they are essentially random respect to other network characteristics. The characteristic q_i^α indicates the community to which a node i belongs on layer $\alpha = 1, 2$. Here we consider the possible correlations existing between the community assignment $q_i^{[1]}$ and $q_i^{[2]}$ in the two layers. This community assignment allows us to tune in a control way the level of overlap between the communities. In particular we label the nodes $i = 1 \dots, N$ in layer 1 according to the following community assignment q_i^α ,

$$q_i^{[1]} = \left\lceil \frac{i}{N_c} \right\rceil. \quad (3.10)$$

where the brackets $\lceil x \rceil$ in the right end side of this expression indicate the ceiling function of x . Therefore we have, for $N = 128$ and $N_c = 32$,

$$q_i^1 = \begin{cases} 1 & \text{for } i \in [1, 32] \\ 2 & \text{for } i \in [33, 64] \\ 3 & \text{for } i \in [65, 96] \\ 4 & \text{for } i \in [97, 128] \end{cases}.$$

The community assignment in layer 2 will not be in general the same of layer 1. In order to model overlap of communities we perform a simple ‘‘shift’’ of the labels, parametrised

with the parameter $\rho > 0$. In particular we take

$$q_i^{[2]} = \begin{cases} \left\lceil \frac{i - \rho N_c}{N_c} \right\rceil & \text{if } \left\lceil \frac{i - \rho N_c}{N_c} \right\rceil > 0 \\ \frac{N}{N_c} & \text{if } \left\lceil \frac{i - \rho N_c}{N_c} \right\rceil = 0 \end{cases}.$$

In general the control parameter ρ takes values $0 \leq \rho \leq 0.5$. If $\rho = 0$ there is no “shift” between the layer partitions (they perfectly match); if $\rho > 0$ each community in the first layer overlaps with the corresponding one in the second layer for a fraction of nodes equal to $(1 - \rho) \cdot N_c$; thus $\rho \cdot N_c$ is the number of “shifted” nodes per community. When $\rho = 0.5$, $N = 128$ and $N_c = 32$, we have

$$q_i^2 = \begin{cases} 1 & \text{for } i \in [17, 48] \\ 2 & \text{for } i \in [49, 80] \\ 3 & \text{for } i \in [81, 112] \\ 4 & \text{for } i \in [1, 16] \cup [113, 128] \end{cases}.$$

Therefore $\rho = 0.5$ describes the maximum “shift” between the community of the two layers: each community in the first layer shares 16 nodes with its corresponding community in the second layer. Given a value of ρ the overall community overlap in the network can be easily calculated, being $O_c = (1 - \rho)$, and in the case of maximum “shift” we obtain $O_c = 0.5$.

For the second benchmark model the Duplex Network LFR model (DNLFR), we have taken a duplex network in which the single layers are constructed according to the LFR model [96].

1. The network in the first layer is a LFR network, formed by Q communities. The communities are labelled according to their size in descending order.
2. The network in the second layer is a LFR network with Q communities generated using the same parameters used for the network in the first layer. Additionally we require that the network in the second layer satisfies a further condition, which allows us to modulate the overlap between the communities in the two layers. Specifically, for each second layer candidate, we first label the communities according to their size in descending order. Then we compare each of them to the corresponding one in the first layer (panel B Figure 3.2). We calculate the number of “shifted” nodes N_s given by the sum of the absolute values of the difference

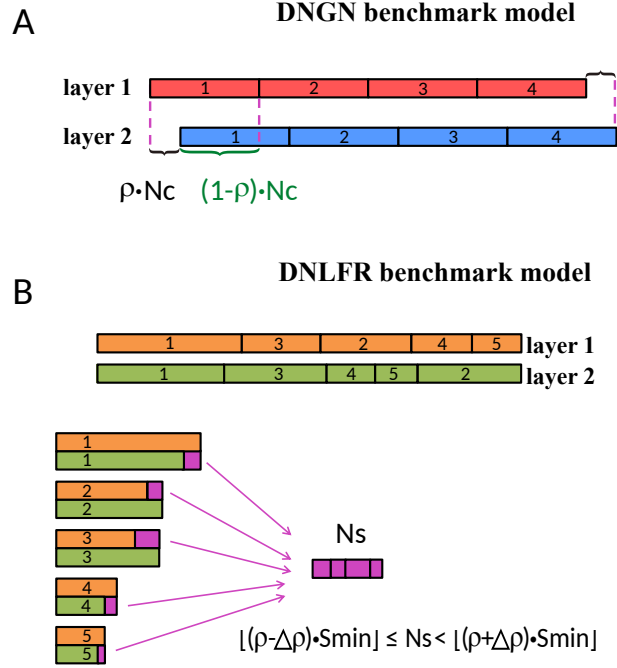


Figure 3.2: Schematic of the benchmark models DNGN and DNFLR. Panel (A). The DNGN benchmark model: nodes on both layers (blue and red) are divided into four communities of equal size N_c , labelled from 1 to 4. Each community of layer 1 overlaps for a fraction of $(1-\rho) \cdot N_c$ nodes with its corresponding community in layer 2. Panel (B). The DNFLR benchmark model: on each layer $Q = 5$ non homogeneous communities are generated and labelled from 1 to 5 according to their size (left). For a given ρ the total number of nodes which do not overlap between communities of the same label, N_s , has values $\lfloor (\rho - \Delta\rho) \cdot S_{min} \rfloor \leq N_s < \lfloor (\rho + \Delta\rho) \cdot S_{min} \rfloor$, where $\lfloor \dots \rfloor$ is the floor function and S_{min} is the minimum bound of the power-law distribution from which the community sizes in the two layers are extracted.

between the corresponding communities sizes, i.e.

$$N_s = \sum_{l=1}^Q \left| n_l^{[1]} - n_l^{[2]} \right|, \quad (3.11)$$

where n_l^α is the size of the community l in layer α . Finally we retain the candidate network as the second layer of the duplex network only if

$$\lfloor (\rho - \Delta\rho) \cdot S_{min} \rfloor \leq N_s < \lfloor (\rho + \Delta\rho) \cdot S_{min} \rfloor, \quad (3.12)$$

where $\lfloor \dots \rfloor$ is the floor function. Here ρ and $\Delta\rho$ are control parameters of the benchmark model that modulate the overlap of the communities, and S_{min} in Eq. 3.12 is the parameter that in the LFR model fixes the lower bound of the community sizes. In this way if one considers a sufficient number of multiple

realisations of the multilayer, and a sufficiently low value of $\Delta\rho$, one gets

$$\langle N_s \rangle \simeq \lfloor \rho \cdot S_{min} \rfloor. \quad (3.13)$$

3. Finally, the nodes are relabelled in both layers in order to allow the maximum community overlap. In particular the labels are reassigned in such a way that the common number of nodes in the communities that have the same label in the two layers, is equal to the minimum of the two community sizes. (see Figure 3.2.)

Therefore the average community overlap of the benchmark network is dependent on ρ and, for a significant number of realisations and low enough values of $\Delta\rho$, is given by

$$\langle O_c \rangle = 1 - \frac{\langle N_s \rangle}{N} \simeq 1 - \frac{\lfloor \rho \cdot S_{min} \rfloor}{N}. \quad (3.14)$$

In order to test the performance of the similarity measure $\tilde{\Theta}^S$, we apply this measure to the two duplex network benchmarks, for different values of ρ . Since ρ modulates the level of community overlap between the layers we expect that the similarity measure $\tilde{\Theta}^S$ is larger for lower value of ρ (corresponding to larger community overlap O_c between the layers) and smaller for larger values of ρ (corresponding to smaller community overlap O_c between the layers). In Figure 3.3 we show the dependence $\tilde{\Theta}^S$ as a function of ρ for the two proposed benchmark models. In both cases the displayed values $\tilde{\Theta}^S$ are averaged over 50 benchmark realisations.

For the DNGN benchmark, we considered $N = 128$, $N_c = 32$ and $\rho \leq 0.5$. The similarity measure $\tilde{\Theta}^S$ is monotonically decreasing with ρ . For the DNLFR benchmark the two single layers are generated according to the LFR algorithm with parameters $N = 600$ (number of nodes) and $Q = 5$ (number of communities). The size of each community is taken from a power-law distribution with lower bound $S_{min} = 60$, upper bound $S_{max} = 180$, and power-law exponent $\tau_1 = 1.5$. Inside the communities the node degree distribution is also extracted from a power-law distribution with parameters $k_{max} = 50$ (maximum degree), $\tau_2 = 2.6$ (power-law exponent), $\langle k \rangle = 16$ (average degree). For building the DNLFR network we used $\Delta\rho = 0.05$ and $\rho \leq 0.95$. Also in the case of the DNLFR benchmark, where the size of the communities is heterogeneous, $\tilde{\Theta}^S$ decreases monotonically with ρ .

This result shows that in benchmark models in which the community overlap is modulated by an external control parameter, $\tilde{\Theta}^S$ decreases together with the community overlap. Since in general measuring the community overlap involves an optimisation over a permutation of the community assignment, measuring the community overlap can be very costly numerically. In this situation calculating $\tilde{\Theta}^S$ could instead give an alternative

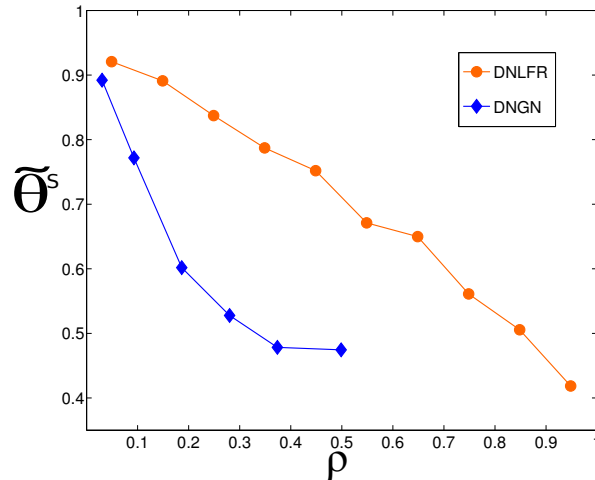


Figure 3.3: The similarity measure $\tilde{\Theta}^S$ between the two layers of the DNGN (blue diamonds) and DNFLR (orange circles) benchmark models is measured as a function of the control parameter ρ . When ρ increases the total community overlap between the layers decreases and $\tilde{\Theta}^S$ decreases monotonically both in the case of homogeneous-size communities (DNGN) and in the case of heterogeneous-size communities (DNFLR). Each data point is averaged over 50 benchmark realisations. For the DNFLR model the parameter $\Delta\rho$ was set to 0.05.

way to assess the similarity between the layers of a multiplex network.

In the following, using the concrete examples of the APS Collaboration Multiplex Networks, we will compare the similarity measure $\tilde{\Theta}^S$ to other existing measures introduced to compare different community assignments in single layers.

3.2.2 The network between the layers of the APS Collaboration Multiplex Networks

In this Section, we use the similarity matrix $\tilde{\Theta}^S$ to analyse the APS Collaboration Multiplex Networks. These multiplex networks are extracted from the APS collaboration dataset [90] recording all the bibliometric information about the papers published in the APS journals.

The network is formed by a set of N nodes representing the APS authors. Since there is no agreement on disambiguation techniques for the author names, we have identified each author with the initials of his/her first name and last name. The layers correspond to different Physics and Astronomy Classification Scheme (PACS) codes [91] describing the subject of the papers. Two authors are linked in a given layer α if they are co-authors of at least one paper having the PACS number corresponding to layer α . Since PACS

numbers are organised in a hierarchical way (the first digit of the number indicates the general field of physics while the second digit specifies the ambit inside that field), we have constructed two multiplex networks whose layers correspond respectively to the first and second hierarchical level of the PACS codes. The APS Collaboration Multiplex Network related to the first level of the hierarchy of PACS codes is made of $M_1 = 10$ layers each one describing the collaboration network in a general field of physics. The APS Collaboration Multiplex Network at the second level of the hierarchy is made of $M_2 = 66$ layers each one describing the collaboration network in a specific ambit of physics (second level of the PACS code hierarchy).

In extracting the APS Collaboration Multiplex Networks we considered all the papers until 2014 with less than ten co-authors. This threshold was introduced to exclude papers coming from big collaborations that follow different statistical properties with respect to the rest of the dataset. With this threshold, our dataset includes a consistent fraction of the whole dataset ($\simeq 97\%$ of the total number of papers) and a number of authors $N = 180,539$.

The layers of the APS Collaboration Multiplex Networks are characterised by a significantly different activity pattern of the nodes. Moreover roughly 0.7% of nodes belong to connected components of size 2 while only about 0.006% of the nodes belongs to connected components of size 3. Therefore we consider here the case in which the characteristics $\{q_i^\alpha\}$ indicate the community of the nodes in layer α and the class p_i^α of node i in layer α takes a different value for each distinct pair (k_i^α, q_i^γ) as long as the node i is not isolated $k_i^\alpha > 0$, and it belongs to a community of more than two nodes. All the isolated nodes belong a the same class \tilde{p} . All the nodes belonging to a two-node community belong to another class \hat{p} .

Let us first characterise the mesoscale similarities between the $M_1 = 10$ layers of the APS Collaboration Multiplex Network in the main subjects of physics, described by the first level of the PACS code hierarchy. The similarity matrix $\tilde{\Theta}^S$ is constructed in two different ways, using either the Infomap community detection algorithm [97] and the Louvain algorithm [49], and averaging in both cases over 350 random permutations of the community assignments. For simplicity we will refer to these two matrices as Infomap- $\tilde{\Theta}^S$ and Louvain- $\tilde{\Theta}^S$. The two matrices are reported in Figure 3.4 in the form of heat-maps. The patterns shown by the two heat-maps are very similar, denoting that from a qualitatively point of view the measure $\tilde{\Theta}^S$ is not affected by the choice of the algorithm used to perform the community detection for the network under study. We can observe that, in general, clusters in the APS Collaboration Multiplex Network extend across multiple layers. As expected, layers describing collaborations in general or interdisciplinary fields such as General Physics or Interdisciplinary Physics, which often

involve people from different specific ambits of physics, show high values of $\tilde{\Theta}^S$ respect to several other layers while more specific fields, such as Gases&Plasma, show lower values of $\tilde{\Theta}^S$ respect to the other layers.

Given this similarity measure between the layers of the multiplex, one can build a network of networks whose nodes represent the $M_1 = 10$ networks of collaboration in general fields of physics and whose weighted edges are the values $\tilde{\Theta}_{\alpha,\beta}^S$ and represent the similarity between the M_1 networks respect to their community structure. This network of layers is thus a weighted fully-connected network showing itself a significant community structure and revealing how the pattern of collaboration between scientists is organized across different fields of physics. In order to characterise this community structure between the layers of the multiplex network, we perform a hierarchical clustering analysis starting from the dissimilarity matrix d of elements $d_{\alpha,\beta}$ given by

$$d_{\alpha,\beta} = 1 - \left| \tilde{\Theta}_{\alpha,\beta}^S \right|. \quad (3.15)$$

Specifically, we use the average linkage clustering method which gave the best cophenetic correlation coefficient compared to other clustering method [98–100]. According to the average method the distance $d_c(C_1, C_2)$ between two clusters C_1 and C_2 is defined as the average distance between all pairs of layers in the two clusters:

$$d_c(C_1, C_2) = \frac{1}{\mathcal{N}(C_1)\mathcal{N}(C_2)} \sum_{\alpha \in C_1} \sum_{\beta \in C_2} d_{\alpha,\beta} \quad (3.16)$$

where $\mathcal{N}(C_i)$ indicates the number of layers in cluster C_i .

In Figure 3.4, together with the matrices Infomap- $\tilde{\Theta}^S$ and Louvain- $\tilde{\Theta}^S$ we show the dendrograms resulting from the hierarchical clustering analysis of the respective dissimilarity matrices Infomap- d and Louvain- d . In order to define an optimal partition of the layers into communities, we looked for the agglomerative stage of the cluster hierarchy at which the weighted modularity Q [101] is maximised, Q defined as:

$$Q = \frac{1}{\langle \eta \rangle M} \sum_{\alpha \neq \beta}^M \left(\left| \tilde{\Theta}_{\alpha,\beta}^S \right| - \frac{\eta_\alpha \eta_\beta}{\langle \eta \rangle M} \right) \delta[\sigma_\alpha \sigma_\beta], \quad (3.17)$$

where σ_α labels the community in which layer α is, $\delta[x, y]$ indicates the Kronecker delta

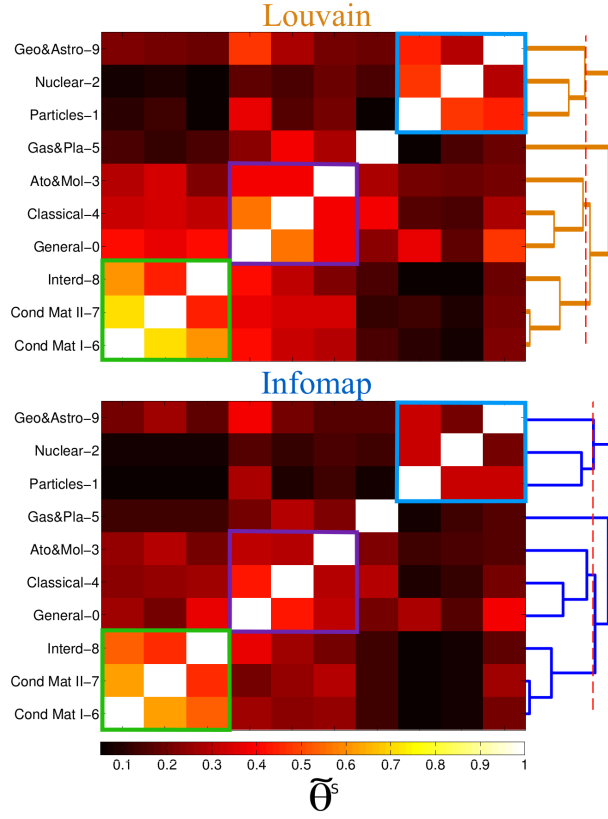


Figure 3.4: The similarity matrices of elements $\tilde{\Theta}_{\alpha,\beta}^S$ calculated respectively using the Louvain and the Infomap community detection algorithms are plotted for the APS Collaboration Multiplex Network with the $M_1 = 10$ layers indicating the collaboration network at the first level of the PACS hierarchy. Each layer refers to a general field of Physics (see Table 3-A for the legend of the layer acronyms). The dendrogram between the layers is shown on the left of each matrix $\tilde{\Theta}_{\alpha,\beta}^S$. The dashed line on top of the dendrogram indicates the partition that correspond to the optimal value of the weighted modularity given by Eq. 3.17).

and $\eta_\alpha, \langle \eta \rangle$ are given respectively by

$$\begin{aligned} \eta_\alpha &= \sum_{\beta \neq \alpha} |\tilde{\Theta}_{\alpha,\beta}^S|, \\ \langle \eta \rangle &= \frac{1}{M} \sum_{\alpha} \eta_\alpha. \end{aligned} \quad (3.18)$$

As shown in Figure 3.4 the optimal partition found is the same either when using the Infomap algorithm or the Louvain algorithm to perform the community detection in the layers of the multiplex. The analysis reveals that the first layers clustering together are Condensed Matter I&II and Interdisciplinary Physics and they form the first block (green

Acronym	PACS	Field
General-0	00	General
Particles-1	10	Physics of Elementary Particles and Fields
Nuclear-2	20	Nuclear Physics
Ato&Mol-3	30	Atomic and Molecular Physics
Classical-4	40	Electromagnetism, Optics, Acoustic, Heat Transfer, Classical Mechanics and Fluid Dynamics
Gas&Pla-5	50	Physics of Gases, Plasmas and Electric Discharges
Cond Mat I-6	60	Condensed Matter: Structural, Mechanical and Thermal Properties
Cond Mat II-7	70	Condensed Matter: Electronic Structure, Electrical, Magnetic and Optical properties
Interd-8	80	Interdisciplinary Physics and Related Areas of Science and Technology
Geo&Astro-9	90	Geophysics, Astronomy and Astrophysics

Table 3-A: The acronyms used in this study for the PACS number at the first level of the PACS hierarchy, the corresponding PACS numbers and corresponding general fields of Physics.

coloured box); the second block includes General Physics, Classical Physics, Atomic and Molecular Physics (purple coloured box); in the third block Particles Physics, Nuclear Physics and Geophysics&Astrophysics group together (cyan coloured box). The layer related to Gases&Plasma Physics is isolated and can be considered as a block by itself.

Once revealed the block (community) structure an interesting issue is to characterise the Minimal Spanning Tree (MST) that allows us to identify the layers which connect the blocks together. Therefore we construct the MST using the dissimilarity measure d defined in Eq. 3.15 calculated either using the Infomap or the Louvain clustering algorithm. The two MSTs are identical (Figure 3.5) and this confirm the robustness of the results with respect to the community detection algorithm used. We can see that the collaboration layer of General Physics connects the three main blocks together.

Block 1	Block 2	Block 3	Block 4
Cond Mat I-6 Cond Mat II-7 Interd-8	General-0 Ato&Mol-3 Classical-4	Particles-1 Nuclear-2 Geo&Astro-9	Gas&Pla-5

Table 3-B: Clusters between the $M_1 = 10$ layers of the APS multiplex network corresponding to the first level of the PACS hierarchy (see for the legend of the layer acronym Table 3-A). The clusters have been obtained from the dendrograms shown in Figure 3.4, cut in order to obtain the partition that optimises the weighted modularity Q defined in Eq. 3.17.

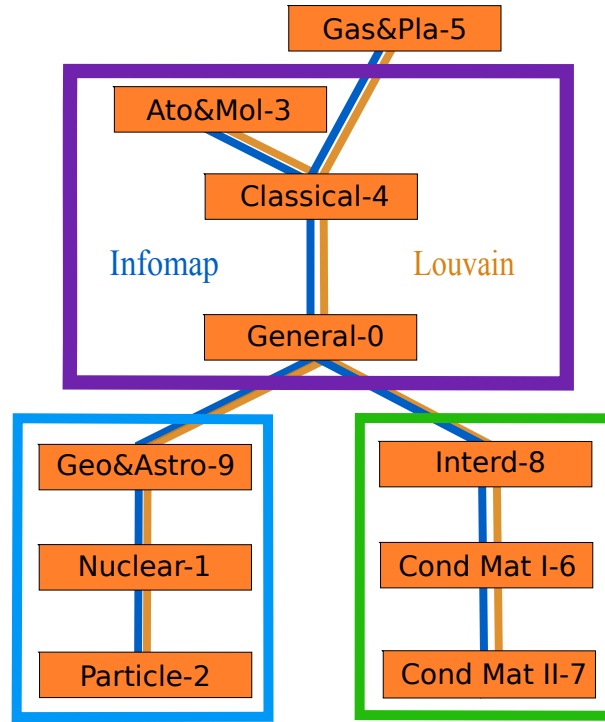


Figure 3.5: Minimal Spanning Tree (MST) using the dissimilarity measure d in the case of Infomap- d dissimilarity (blue) and in the case of Louvain- d dissimilarity (oher). The block structure obtained with the hierarchical clustering analysis is also showed.

In order to have a deeper understanding of the results previously found we now consider the multiplex network of scientific collaborations where the layers are related to the PACS code at the second level of the PACS hierarchy. For this multiplex network we have calculated the similarity matrix $\tilde{\Theta}^S$ between the $M_2 = 66$ layers and found the optimal partition into communities according to the score function Q , following an analogous procedure to the one used previously for first level of the PACS hierarchy. To calculate $\tilde{\Theta}_{\alpha,\beta}^S$ we have performed averages over 350 random permutations of the

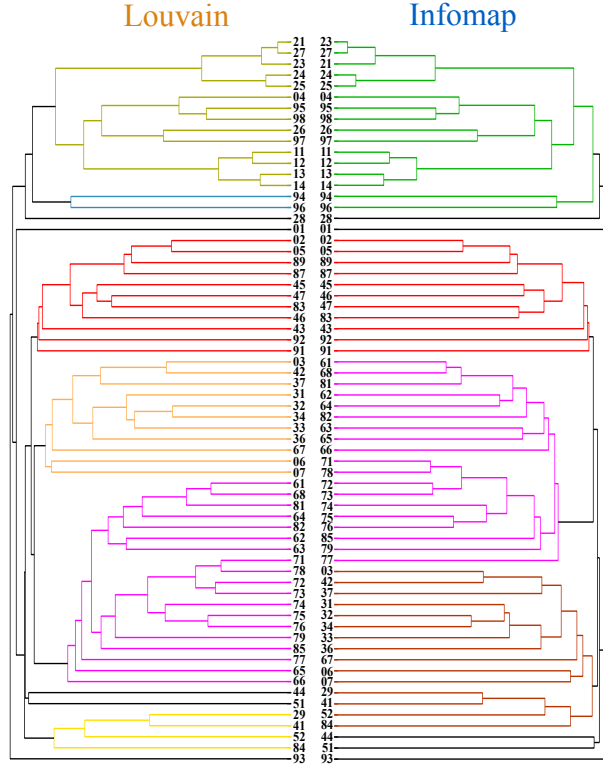


Figure 3.6: Hierarchical clustering of the APS Collaboration Multiplex Network in which each layer represents a collaboration network in a specific area of physics, as described by the second hierarchical level of the PACS code. We show the two dendrograms obtained respectively from the Louvain- $\tilde{\Theta}_{\alpha,\beta}^S$ (left) and from the Infomap- $\tilde{\Theta}_{\alpha,\beta}^S$ (right). In each dendrogram the communities found at the optimal partition (maximum of Q) are represented as branches of the same colours.

community assignments.

In Figure 3.6 we plot the dendrograms resulting from the hierarchical clustering analysis in the case of Louvain- d dissimilarity and Infomap- d dissimilarity. For each dendrogram, the clusters found in the optimal partitions are represented as branches of the same colours. When using the Louvain- d dissimilarity we obtain six clusters plus some isolated layers. When using the Infomap- d dissimilarity we obtain four clusters plus isolated layers. Nevertheless we observe that two of the clusters (the red and the violet clusters) are identically the same in the two partitions. The other two clusters obtained with the Infomap- d dissimilarity are each divided into two clusters when considering the optimal partition using the Louvain- d dissimilarity. In particular the combination of the green-yellow and green-blue clusters in the Louvain partition is identical to the green cluster of the Infomap partition, while the combination of the orange and the yellow clusters in the Louvain algorithm is identical to the brown cluster of the Infomap partition.

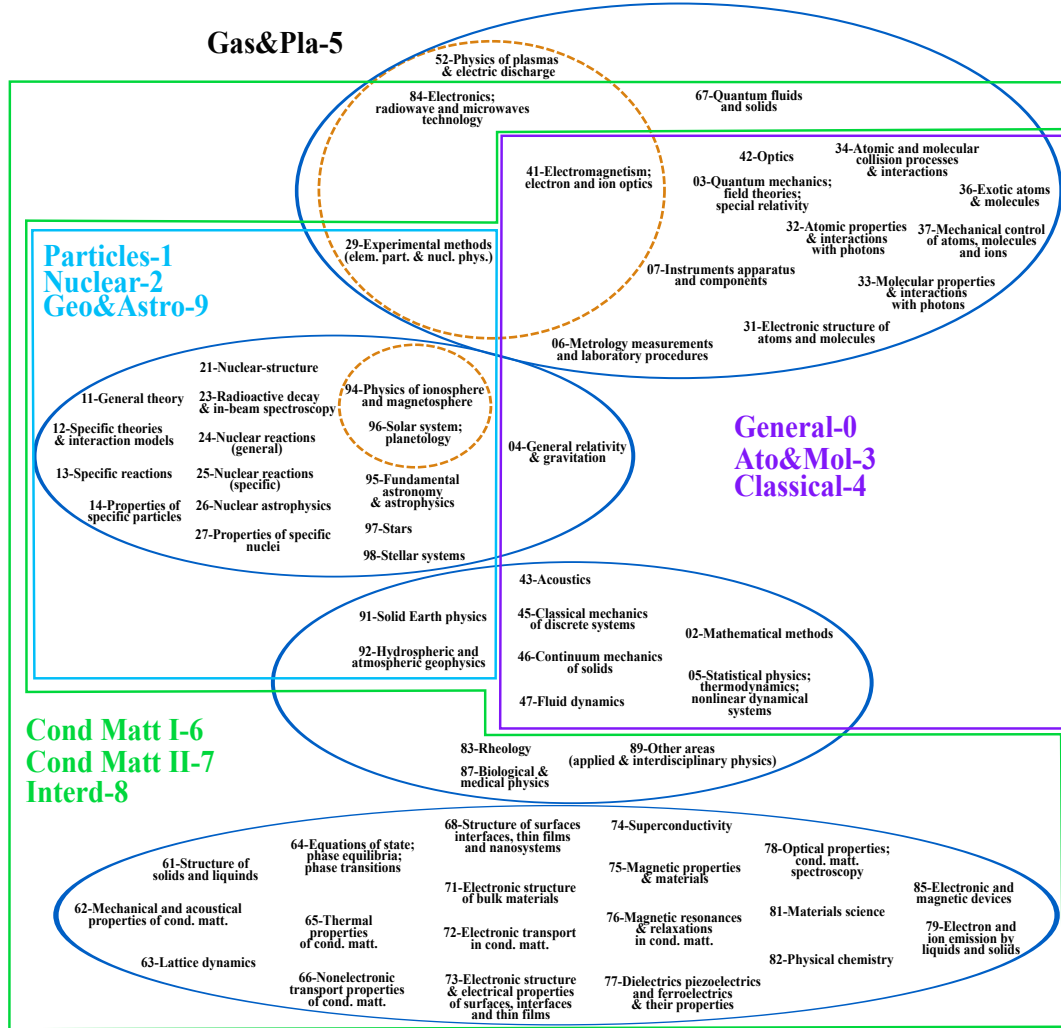


Figure 3.7: Optimal community structure of the layers of the APS Collaboration Network in which each layer represents a collaboration network in a specific area of physics, as described by the second hierarchical level of the PACS code. The four communities found starting from the Infomap- $\tilde{\Theta}_{\alpha,\beta}^S$ matrix are represented by blue solid-line ovals. In the partition obtained from the Louvain- $\tilde{\Theta}_{\alpha,\beta}^S$ two sub-communities (other dashed ovals) are considered separate communities. These communities form the course-grained partition into the three blocks found at the first hierarchical level of the PACS code (coloured solid-line polygons). The nodes displayed in this figure correspond to a subset of 61 layers that are not isolated in the optimal partition in communities which optimises the weighted modularity Q .

In Figure 3.7 we give an overview of the blocks hierarchy found. The four clusters found in the Infomap- d optimal partition matrix are represented by solid-line ovals. Dashed ovals split two clusters in two, according to the results obtained from the Louvain- d optimal partition. The block structure at the first level of the PACS hierarchy is shown using solid-line polygons. This method allows us to characterise with a bottom-

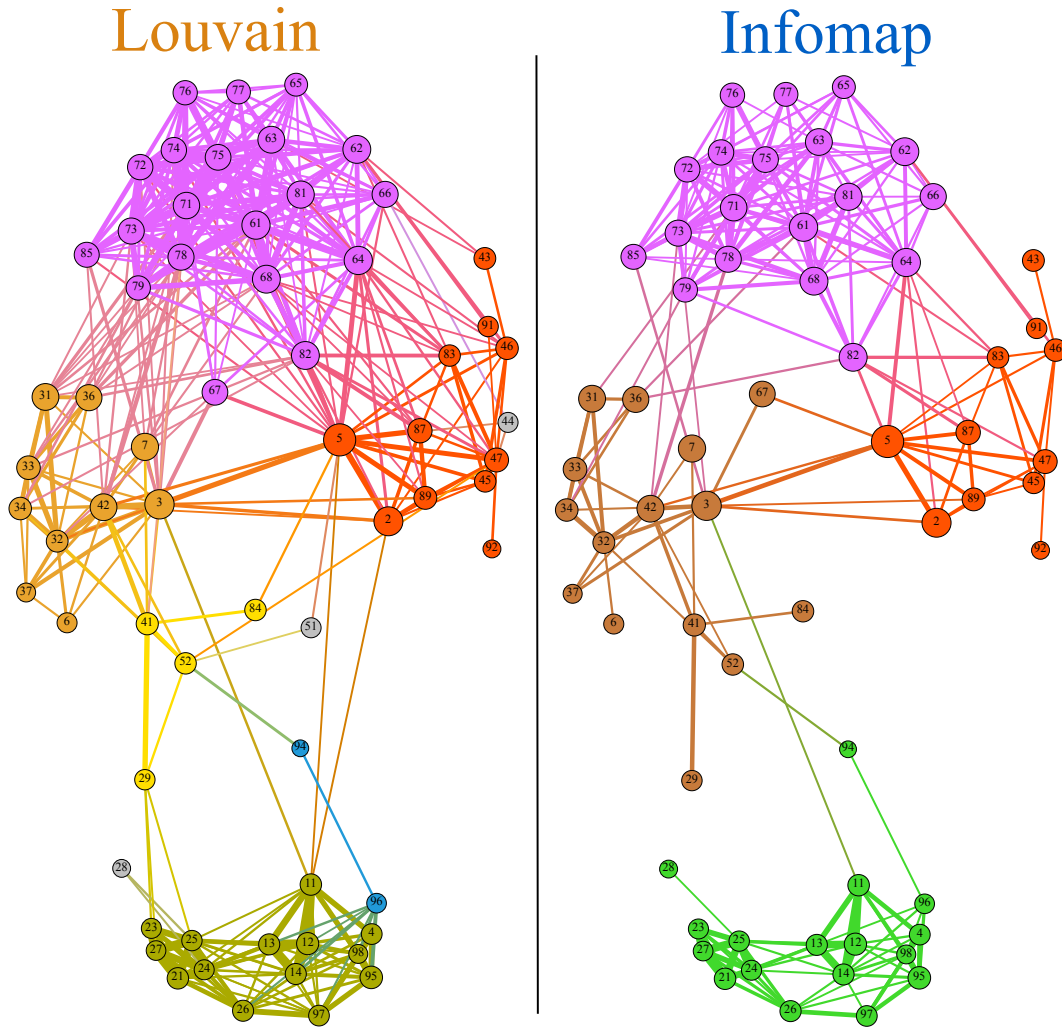


Figure 3.8: The network between the layers of the APS Collaboration Multiplex Network (with layers corresponding to the PACS code at the second level of the PACS hierarchy) is displayed here for the two cases in which the Louvain- $\tilde{\Theta}^S$ or the Infomap- $\tilde{\Theta}^S$ similarity matrix are used. The link weights represent the similarity between the community structure of the two linked layers. The networks are obtained from the $\tilde{\Theta}^S$ similarity matrix by filtering out the links below a given threshold value. The threshold is chosen to be the maximal value that ensures that in the filtered network each layer is connected with at least one layer inside its own cluster. The architecture of the networks describes the interplay between the collaboration networks and the organisation of knowledge in physics. The community structure revealed by the hierarchical clustering analysis is shown making use of the same colour scheme of Figure 3.6.

up method how the organisation of knowledge in physics is effectively perceived by scientists while shaping their collaboration network. We observe that while the PACS hierarchy clearly captures main features of the collaboration network, the analysis of the Collaboration Multiplex Network at the second level of the PACS hierarchy clearly

suggests a hierarchical organisation of these PACS numbers that is not equivalent to the first level of the PACS hierarchy. Finally we used the information gained by this analysis to construct the network of networks between the layers of the Collaboration Multiplex Network at the second level of the PACS hierarchy. To this aim we have constructed the weighted network determined by an opportune thresholding of the Louvain- $\tilde{\Theta}^S$ or Infomap- $\tilde{\Theta}^S$ similarity matrix (see Figure 3.8). The threshold, is here given by the minimum value of the similarity matrix $\tilde{\Theta}^S$ that ensures that each layer is connected to at least one other layer of its own cluster. From these networks, it is possible to appreciate that, although the network between the layer of the Collaboration Mutliplex Network is highly interconnected, the clusters found corresponds to layers much more similar between themselves than with other layers outside their own cluster. Interestingly this visualisation shows that the two clusters detected only by the Louvain algorithm, [94, 96] and [29, 41, 52, 84], contain the nodes that act as bridges between the yellow-green cluster and the red and the orange clusters. This might explain why the Louvain algorithm identifies them as separate clusters.

3.2.3 Comparison of the results obtained with $\tilde{\Theta}^S$ respect to other similarity measures

In this Section we compare the results obtained from the analysis of the APS Collaboration Multiplex Network using the $\tilde{\Theta}^S$ indicator with results from other similarity measures commonly used to compare different network partitions [44] and with the *ACTIVIS* Index, an index able to capture the similarity of the layers of a multiplex due to the activity of the nodes. In particular, focusing on the highest level of the PACS hierarchy, we compute the Normalized Mutual Information *NMI* [45], the Jaccard index *J* [102], the Rand index *R* [103, 104] and the *ACTIVIS* Index for each pair of the $M_1 = 10$ layers. Given two network partitions X and Y , the Normalized Mutual Information *NMI*, is defined as

$$NMI(X, Y) = \frac{2[H(X) - H(X|Y)]}{H(X) + H(Y)}, \quad (3.19)$$

where $H(X) = -\sum_x P(x) \log P(x)$ is the entropy associated to the distribution $P(x)$ of sizes x of the clusters classified by the partition X ; $H(Y)$ corresponds to the entropy associated to the distribution $P(y)$ of the sizes y of the clusters in the partition Y ; $H(X|Y)$ is the conditional entropy associated to the distribution of the community assignment X conditioned on the distribution of the community assignment Y and is given by $H(X|Y) = -\sum_{x,y} P(x, y) \log P(x, y)/P(y)$, $P(x, y)$ the distribution of the number of nodes having community assignment x in partition X and y in partition Y .

The Jaccard index J and the Rand Index R , are instead defined as

$$J(X, Y) = \frac{a_{11}}{a_{11} + a_{10} + a_{01}} \quad (3.20)$$

$$R(X, Y) = \frac{a_{11} + a_{00}}{a_{11} + a_{10} + a_{01} + a_{00}}, \quad (3.21)$$

where a_{11} is the number of pairs of nodes belonging to the same cluster in both partitions X and Y , a_{00} is the number of pairs of nodes classified in different clusters in both the X and Y partitions, and $a_{10}(a_{01})$ is the number of pair of nodes belonging to the same cluster in $X(Y)$ but belonging to different clusters in $Y(X)$.

Finally we define the Activity Similarity *ACTIVIS* Index between the layers α and β of a multiplex network, which compares the activity patterns in different layers. This index is given by

$$ACTIVIS = b_{11} + b_{00}, \quad (3.22)$$

where b_{11} are the fraction of nodes active in both layers and b_{00} are the fraction of nodes inactive in both layers.

In Figure 3.9 we show the similarity matrices for the different measures and their respective dendrograms, obtained with the same hierarchical clustering analysis discussed above for the $\tilde{\Theta}^S$ case. Here the layer partitions are obtained using the Infomap algorithm. When the modularity Q is optimized, the partition obtained with all these alternative measures are different from the one obtained using the $\tilde{\Theta}^S$ indicator function. Moreover the partitions obtained are characterised by having at least 3 out of 10 layers in separate clusters, resulting in significantly less relevant partitions. Moreover, by looking at the dendrograms, we can see that none of the other measure is able to give the optimal partition obtained with $\tilde{\Theta}^S$ even by applying an arbitrary cut to the respective dendrogram.

These results show clearly that the proposed indicator function $\tilde{\Theta}^S$ based on information theory, is not equivalent to previously defined similarity measures between partitions. Moreover the method is not affected significantly by the choice we made for treating inactive nodes or nodes belonging to connected components of two nodes. Although it might be a challenging technical problem to assess which of the similarity measures proposed so far is the best, the similarity measure $\tilde{\Theta}^S$ seems to be more relevant of other similarity measures used in the literature when applied to the APS Collaboration Multiplex Networks. In fact the partition obtained by using the similarity measure $\tilde{\Theta}^S$ reflect much more closely the general perception of the organisation of collaborations in the physics community.

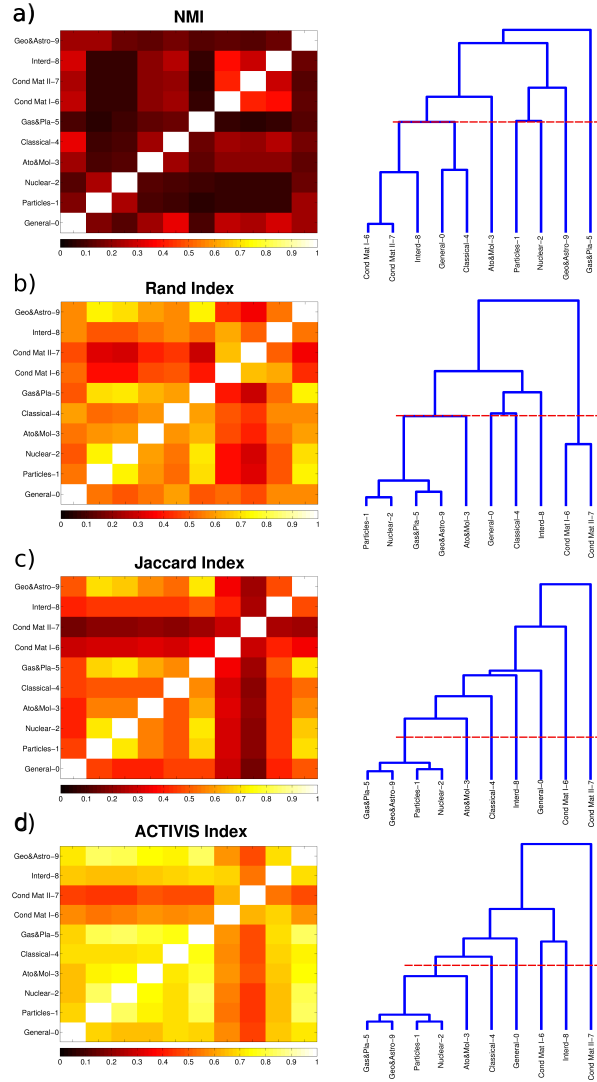


Figure 3.9: Other similarity measures used to hierarchically cluster the $M_1 = 10$ layers of the APS Collaboration Multiplex Network at the first level of the PACS hierarchy. The similarity matrices and their respective dendrograms cutted at the partition optimising the modularity Q (red-dashed line) are shown for the Normalized Mutual Information (Panel a), Rand Index (Panel b), Jaccard Index (Panel c) and for the *ACTIVIS* Index (Panel d). Layer partitions are obtained using the Infomap community detection algorithm. None of the optimal partitions $\tilde{\Theta}$ corresponds to the one obtained using $\tilde{\Theta}^S$ to measure similarities.

3.3 Emergence of multiplex communities in social collaboration networks

Here we investigate the multiplex nature of communities in collaboration networks and we propose a simple model to explain the appearance, coexistence and co-evolution of

communities at the different layers of a multiplex. Our hypothesis is that the formation of communities in collaboration networks is an intrinsically multiplex process, which is the result of the interplay between intra-layer and inter-layer triadic closure. For instance, in the case of scientific collaborations, multiplex communities naturally arise from the fact that scientists may collaborate with other researchers in their principal field of investigation and with colleagues coming from other scientific disciplines. Analogously, actors can prefer either to specialise in a specific genre or instead to explore different (sometimes dissonant) genres, and these two opposite behaviours undoubtedly have an impact on the kind of meso-scale structures observed on each of the layers of the system. The generative model we propose here mimics two of the most basic processes that drive the evolution of collaborations in the real world, namely intra- and inter-layer triadic closure, and is able to explain the appearance of overlapping modular organisations in multi-layer systems. We will show that the model is able to reproduce the salient micro-, meso- and macro-scale structure of different real-world collaboration networks, including the multi-layer network of co-authorship in journals of the American Physical Society (APS) and the multiplex co-starring graph obtained from the Internet Movie Database (IMDb).

3.3.1 Empirical analysis of multiplex collaboration network datasets

We start by analysing the structure of two multiplex collaboration networks from the real world. The first multiplex is constructed from the APS co-authorship data set, and consists of four layers representing four sub-fields of physics (respectively, Nuclear physics, Particle physics, Condensed Matter I, and Interdisciplinary physics). In particular, we considered only scientists with at least one publication in each of the four sub-fields, and we connected two scientists at a certain layer if they had co-authored at least a paper in the corresponding sub-field. The second multiplex is constructed from the Internet Movie Database (IMDb) and consist of four layers respectively representing the co-starring networks of actors with at least one participation in four different genres, namely Action, Crime, Romance, and Thriller movies. The basic structural properties of each layer of the two multiplexes are summarised in Table 3-C (see Appendix A for details about the datasets).

Since we are interested in assessing the role of intra- and inter-layer triadic closure in the formation of meso-scale multiplex structures, we quantified the transitivity of each layer through the clustering coefficient C defined by Eq. 2.15, which takes values in the interval $[0, 1]$. We notice that the four layers of each data set have similar values of clustering, ranging respectively in $[0.24, 0.3]$ in the case of APS and in $[0.56, 0.61]$ for

APS	N	$\langle k \rangle$	C
Nuclear (N)	1238	4.75	0.27
Particle (P)	1238	4.66	0.30
Cond. Matt. I (CM)	1238	10.29	0.24
Interdisciplinary (I)	1238	7.37	0.26
IMDb	N	$\langle k \rangle$	C
Action (A)	55797	83.56	0.61
Crime (C)	55797	82.30	0.58
Romance (R)	55797	86.00	0.59
Thriller (T)	55797	77.75	0.56

Table 3-C: Basic properties of real-world multiplex collaboration networks. We report the number of nodes N , the average degree $\langle k \rangle$ and the clustering coefficient C for each layer of a subset of the APS and IMDb data sets. In particular, we focus on the multiplex collaboration network of all scientists active in Nuclear, Particle, Condensed Matter I and Interdisciplinary physics, and the multiplex collaboration network of all actors starring in Action, Crime, Romance and Thriller movies. All the layers of APS have a clustering coefficient C in the range $[0.24, 0.30]$. Conversely, the values of C of all the IMDb layers are in the range $[0.56, 0.61]$.

IMDb. As we will discuss in the following, by focusing on layers having comparable clustering we will be able to perform a comparison between the structure of these real-world multiplex networks and the proposed model in its simplest formulation.

The multiplex nature of communities in collaboration networks can be measured by means of the normalised mutual information (NMI) defined by Eq. 3.19, which quantifies the similarity between the partition in communities observed in two different layers of a multiplex. The normalised mutual information takes values in $[0, 1]$. In general, higher values of NMI correspond to more similar partitions. The values of NMI for each pair of layers in APS and IMDb are shown in Figure 3.10. It is interesting to notice that in general pairs of layers corresponding to related subjects or genres exhibit higher values of NMI. This is for instance the case of Nuclear Physics and Particle Physics in APS. Similarly, in the IMDb network we observe a higher similarity between the communities at the three layers representing respectively Thriller, Crime and Action genres. Conversely, the layer of Romance movies displays a different modular structure from Crime and Action. Notice also that the level of similarity between the communities of two layers can vary substantially, despite the four layers of each multiplex have roughly the same clustering coefficient.

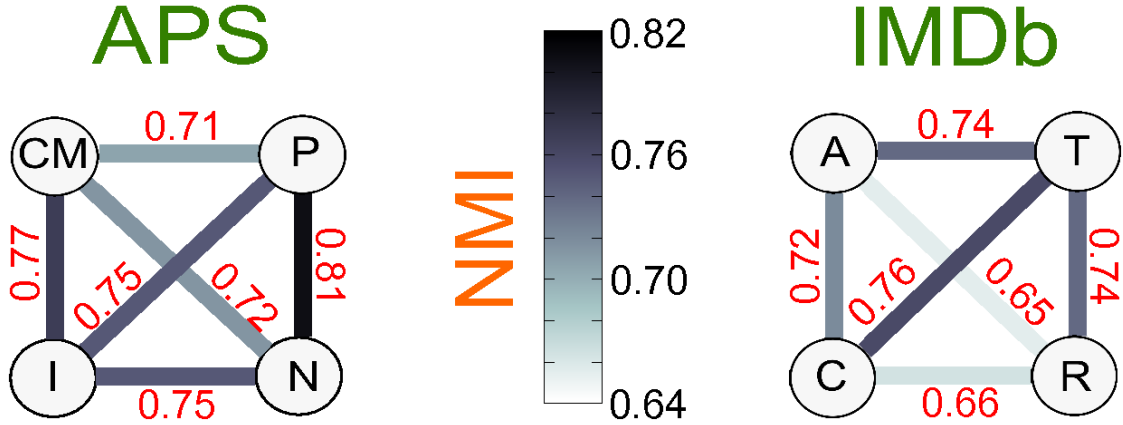


Figure 3.10: Similarity of communities at the different layers of real-world collaboration networks. In each of the two graphs nodes represent the layers of the multiplex (APS on the left and IMDb on the right) and the edges are coloured according to the value of the normalised mutual information for the community decompositions at the corresponding pairs of layers.

3.3.2 A model for evolving multiplex communities

In the following Section we introduce a model to grow collaboration networks with tunable multiplex community structure, able to reproduce the patterns observed in the considered real-world systems. Let us consider for simplicity the case of a multiplex with $M = 2$ layers, and assume that initially each layer consists of a clique of n_0 nodes. Then at each time step t a new node is added to the network, with $m^{[1]}$ edge stubs to be connected on layer 1 and $m^{[2]}$ other stubs to be connected on layer 2. The multiplex network grows according to the following rules:

- *Layer selection.* The newly arrived node i selects one of the two layers $\{1, 2\}$ uniformly at random. Let us label the first selected layer with the index a . The first edge of i is connected to one of the existing nodes on that layer, chosen uniformly at random, that we call n_a .
- *Intra-layer triadic closure (I).* The remaining $m^{[a]} - 1$ edges of node i on layer a are attached with probability $p^{[a]}$ to one of the first neighbours of n_a , chosen uniformly at random, and with probability $1 - p^{[a]}$ to one of the nodes of layer a , chosen uniformly at random.
- *Inter-layer triadic closure.* When all its $m^{[a]}$ edges on layer a have been created, node i starts connecting on the other layer b with $m^{[b]}$ edges. The first link in layer b is created with probability p^* to the same node n_a , and with probability $1 - p^*$ to one of the other nodes, chosen uniformly at random. The node to which this

first link is attached is called n_b .

- *Intra-layer triadic closure (II)*. The remaining $m^{[b]}-1$ links at layer b are attached with probability $p^{[b]}$ to one of the first neighbours of n_b chosen uniformly at random, and with probability $1 - p^{[b]}$ to one of the nodes at layer b , chosen uniformly at random.

This general model has five parameters to be tuned, namely the number of new edges $m^{[1]}$ and $m^{[2]}$ brought by each new node on each of the two layers, which determine the average degree on each layer, and the three probabilities $p^{[1]}$, $p^{[2]}$, and p^* , which are respectively responsible for the formation of intra- and inter-layer triangles. In fact, by varying the parameters $p^{[1]}$ and $p^{[2]}$ we can tune the strength of the intra-layer triadic closure mechanism, i.e the probability to form triangles on each of the two layers. In particular, larger values of $p^{[1]}$ and $p^{[2]}$ will foster the creation of a larger number of triangles in layer 1 and layer 2 respectively. Conversely, the parameter p^* tunes the inter-layer triadic closure mechanism, and in particular high values of p^* correspond to a higher probability that the neighbourhoods of node i at the two layers will exhibit a certain level of overlap. These two simple attachment rules, namely intra-layer and inter-layer triadic closure, aim to describe the real mechanisms characterising the evolution of collaboration networks. We argue that, for instance, scientists do not tend to collaborate with other scientists at random. Instead, they usually exploit the neighbourhoods of their collaborators in a specific field (*intra-layer* triadic closure). Similarly, when opening themselves to new scientific fields, a researcher usually takes into account the neighbourhoods of their past colleagues from previous collaborations in other fields (*inter-layer* triadic closure). A schematic representation of the model is depicted in Figure 3.11.

It has been recently shown [105] that in a single-layer network scenario the interplay between random attachment and triadic closure leads to a network growth in which the attachment probability (i.e., the probability for an existing node to receive one of the new edges) is a sub-linear function of the degree, and produces networks with non-trivial community structure, as long as the link density is not too high. In the multi-layer model we propose, the further addition of an inter-layer triadic closure mechanism allows to tune at will the overlap between the community structures at the different layers.

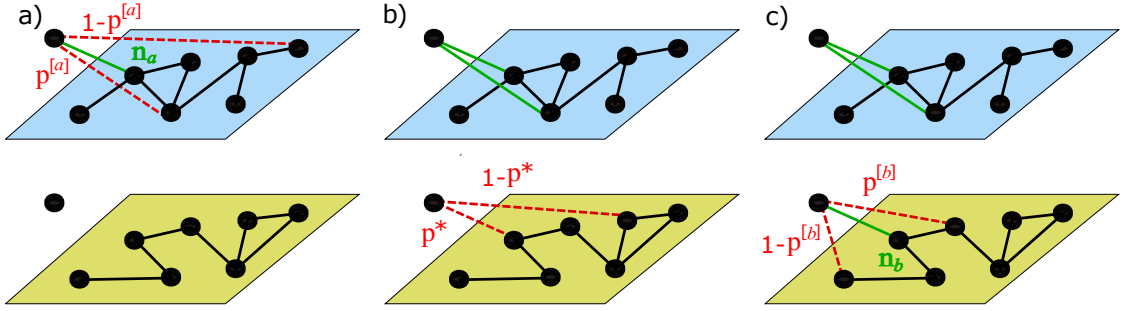


Figure 3.11: Schematic representation of network growth with intra-layer and inter-layer triadic closure. A newly arrived node i creates $m^{[1]}$ new edges on layer 1 and $m^{[2]}$ new edges on layer 2. The new node starts by choosing at random one of the two layers $\{1, 2\}$. We indicate the first chosen layer using the label a . a) The first link of the new node is connected to one of the nodes of layer a , chosen uniformly at random and called n_a (solid green line). Each of the remaining $m^{[a]} - 1$ links is attached with probability $p^{[a]}$ to a neighbour of the previously chosen node (intra-layer triadic closure) or with probability $1 - p^{[a]}$ to one of the nodes at layer a , chosen uniformly at random (dashed red lines). b) Afterwards, the new node starts connecting on the other layer b . The first link on layer b is created to node n_a with probability p^* , or to one of the other nodes at layer b at random with probability $1 - p^*$. We call n_b the first node to which i attaches on layer b . c) Each of the $m^{[b]} - 1$ remaining edges on layer b are attached with probability $p^{[b]}$ to one of the neighbours of n_b , and with probability $1 - p^{[b]}$ to one of the nodes on layer b , chosen uniformly at random.

Validation in a Simple Scenario

To assess the ability of the model to reproduce the organisation of communities in multiplex networks, we start by considering a simple scenario, i.e. the case in which the layers of the multiplex have the same density ($m^{[1]} = m^{[2]} = m$) and the same clustering coefficient ($p^{[1]} = p^{[2]} = p$). We show that this simplified version of the model is already able to reproduce both the different levels of similarity between community structures at different layers, and the microscopic patterns of intra-layer and inter-layer degree correlations observed in the real-world collaboration multiplexes of APS and IMDb.

In Figure 3.12(a), we report the values of the clustering coefficient C (which, by construction, does not depend on the parameter p^*) for several realisations of the model (see Methods). As expected, the clustering coefficient of each layer is a linearly increasing function of the parameter p , which tunes the strength of intra-layer triadic closure. This means that, if we consider a real-world multiplex network whose layers have approximately the same value of clustering coefficient C , we can set the value of the parameter p of the model accordingly. This is for instance the case of the four-layer multiplex networks of APS and IMDb constructed in the previous Section, where all the layers have

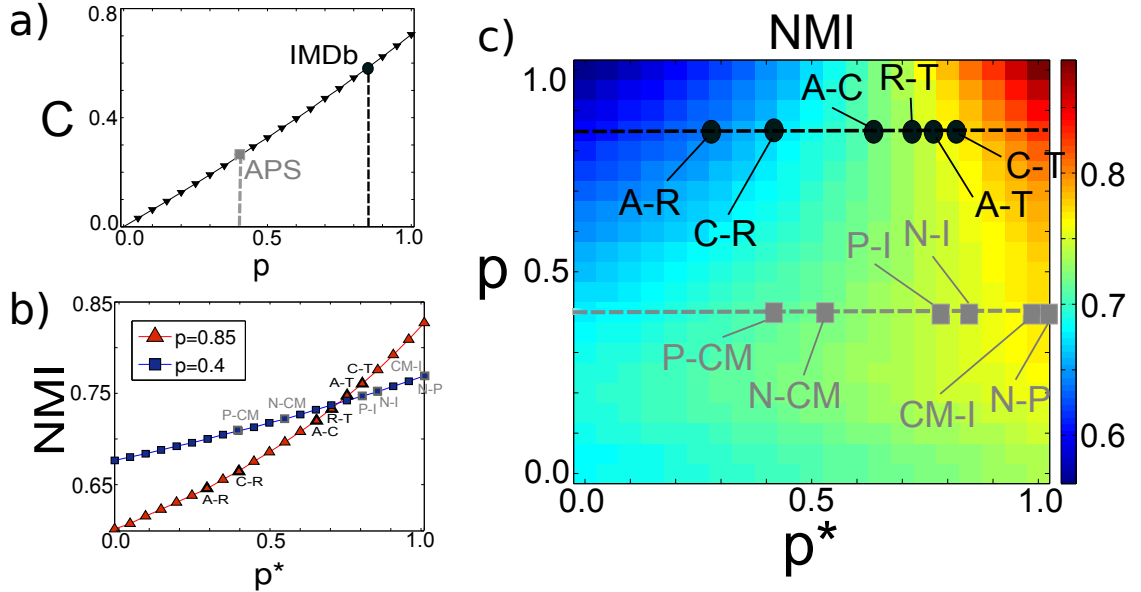


Figure 3.12: Model calibration in a simple scenario. We show the values of p and p^* extracted for the different pairs of layers of the four-layer collaboration networks of APS and IMDb. (a) The clustering coefficient C depends exclusively on the parameter p , which tunes intra-layer triadic closure. Since all the layers of those two multiplex networks have comparable clustering coefficients, we are able to determine the value of the parameter p in each of the two cases. (b) For each pair of layers, we can also determine the value of the inter-layer triadic closure parameter p^* by setting it equal to the value which yields an organisation in communities characterised by a value of NMI compatible with that observed in the real network.

comparable levels of clustering. We obtain $p = 0.40$ for APS and $p = 0.85$ for IMDb, respectively.

In Figure 3.12(c) we show, as a colour-map, the values of NMI of the networks obtained through the proposed model by using different combinations of the parameters p and p^* (see Methods). It is evident that, in spite of its simplicity, the model can yield a quite rich variety of multiplex networks. In agreement with intuition, when both p and p^* are large one obtains multiplexes with higher values of NMI. In fact, in this regime both the intra-layer and inter-layer triadic closure mechanisms are strongly affecting the network evolution and, as a consequence, it is likely that the new node joining the network will close a triad on both layers in the same region of the network. As a consequence, each layer will have a strong community structure (large p) which is pretty much correlated to the one present on the other layer, due to the large value of inter-layer triadic closure p^* . Conversely, if the inter-layer parameter p^* is small we will obtain layers whose partitions in communities are poorly correlated when p is large (blue region in the phase space of Figure 3.12, while the NMI is only marginally larger when

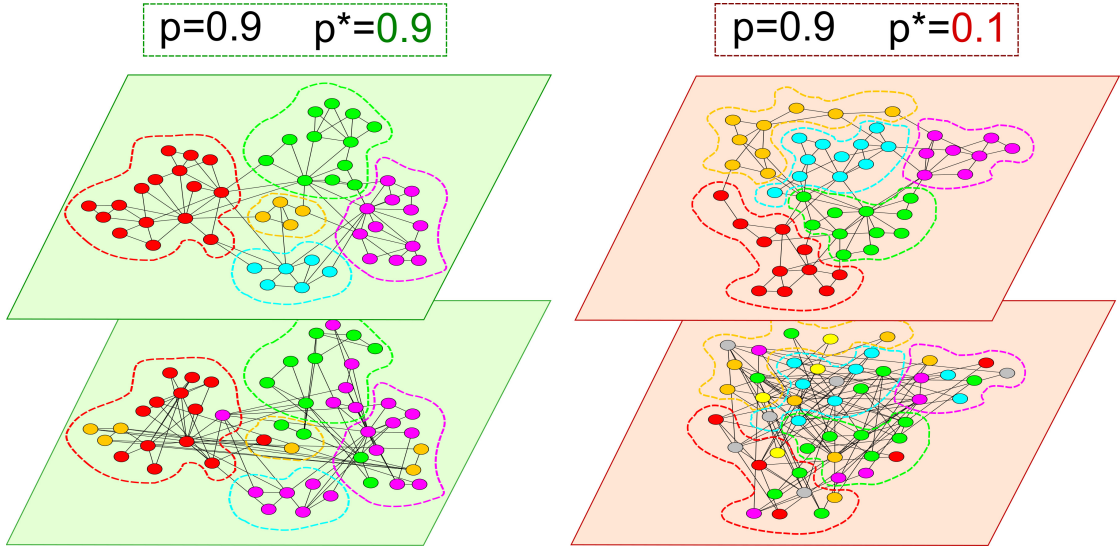


Figure 3.13: Layers with similar or dissimilar community structures. We show the effect of the value of the inter-layer triadic closure parameter p^* on the multiplex community structure. The two top layers show two typical realisations of the simplest version of the network model with $N = 50$, $m^{[1]} = m^{[2]} = 2$ and $p^{[1]} = p^{[2]} = 0.9$. Nodes belonging to the same community are given the same colour and are drawn close to each other. The two layers at the bottom of each multiplex are obtained by setting, respectively, $p^* = 0.9$ (left) and $p^* = 0.1$ (right). The nodes maintain the same placement in space on the second layer, but are coloured according to the community they belong in that layer (colours are chosen in order to maximise the number of nodes that have the same colour in the two layers). It is evident that the community structures of the two layers on the left, corresponding to $p^* = 0.9$, are very similar, while the partition into communities of the upper layer on the left panel is substantially different from the one observed in the bottom layer of that multiplex.

p is very small (bottom-left corner of the phase space).

In Figure 3.13 we report two realisations of the multiplex network model with $N = 50$, $m^{[1]} = m^{[2]} = 2$ and $p^{[1]} = p^{[2]} = 0.9$, respectively for $p^* = 0.9$ (left) and $p^* = 0.1$ (right). Nodes belonging to the same community are reported using the same colour, and the colour chosen for each community in the second layer (bottom) corresponds to the colour of the community in the first layer (top) for which the node overlap between the communities is maximum. These two examples help explain the role of the parameter p^* in shaping the inter-layer modular structure of the network. For $p^* = 0.9$ (left panel) the community structures of the two layers are closely matched (this situation corresponds to the high values of NMI found in the top-right region of the heat-map in Figure 3.12), while for $p^* = 0.1$ (right panel) the communities at the two layers are uncorrelated (low values of NMI in the top-left of the heat-map in Figure 3.12).

Differently from the clustering coefficient C , the values of the normalised mutual information NMI depend on both p and p^* . Having already determined a candidate value of p for each multiplex by fitting the clustering coefficient of its layers, we can determine the strength of the inter-layer triadic closure mechanism by fitting the NMI. Remarkably, for any fixed value of p , the simplest formulation of our model is able to reproduce all the values of NMI observed in the real-world networks by just tuning the parameter p^* , with the exception of the pair Nuclear-Particle physics which is slightly out of the plane with an NMI value of 0.81 (represented on the right border of the plane which corresponds to NMI=0.79). We would like to note here that the model is able to produce a remarkably wide range of values of NMI, which span the whole interval $[0.6, 0.9]$.

We further validate the model by showing that, using the inferred parameters (p, p^*) , we are able to reproduce quite well the patterns of degree-degree correlations observed in the real-world collaboration multiplexes.

Indeed, for each pair of layers α and β we analysed:

1. the intra-layer degree correlations, by looking at the average degree $\langle Knn^{[\alpha]} \rangle$ of the first neighbours on layer α of nodes having a certain degree $k^{[\alpha]}$, as a function of $k^{[\alpha]}$;
2. the inter-layer degree correlations, by looking at the average node degree $\langle k^{[\beta]} \rangle$ on β given the degree $k^{[\alpha]}$ on α ;
3. the mixed degree correlations by looking at the average mixed node degree $\langle Knn^{[\beta, \alpha]} \rangle$ given the degree $k^{[\alpha]}$ on α ;

(see Methods for details). The results are shown in Figure 3.14 for some significant examples. Dots represent the values measured on the real-world networks, while solid lines correspond to the values obtained in the corresponding multiplex models. Symbols with a hat ($\hat{}$) indicate that the value of the considered variables, for both the model and the data, have been normalised to the values of the corresponding configuration model to allow a comparison (see Methods). It is interesting to notice that the model reproduces quite well the three types of degree correlations in the IMDb multiplex, both in the case of high p and high p^* (Action, Thriller given Action) and the case of small p and small p^* (Action, Romance given Action). A quantitative comparison of the the power-law fits of the curves is reported in Table 3-D. As an example from APS we consider Condensed Matter I and Interdisciplinary physics (small p and high p^*). In this case we observe marked differences in the correlations measured in the real-world network and in the model network, for both $\langle Knn^{[\alpha]} \rangle$ and $\langle Knn^{[\beta, \alpha]} \rangle$. In particular, the model seems to

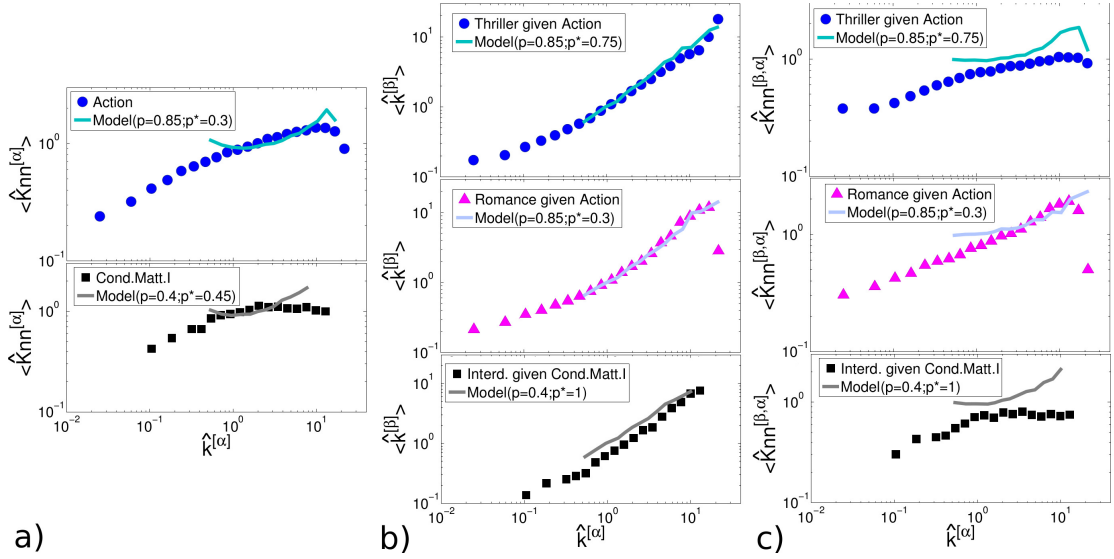


Figure 3.14: Intra-layer, inter-layer and mixed assortativity in collaboration networks. We show the intra-layer (a), inter-layer (b) and mixed (c) degree-correlations for couples of layers of the IMDb and APS collaboration networks. Real data (dots) are compared with the results of our model (solid lines) generated with the extracted values p and p^* . The symbols ($\hat{\cdot}$) indicate that the reported quantities (both for the model and the data) have been normalised to the values observed in the corresponding configuration model. As shown, the model is in general able to correctly capture the assortative trends of the three different types of correlations. Very good agreement with the data is attained in the case of the movie actor collaboration network. Less precise results are obtained for the APS network, where we deal with a system of considerably smaller size.

overestimate degree correlations. These discrepancies are probably due to the relatively small number of nodes (only 1238) in the considered data subset.

Although our intention was not to exactly reproduce all the features observed in real-world collaboration multiplex networks, it is interesting to observe that the two mechanisms of inter-layer and intra-layer triadic closure play an important role in determining the degree-degree correlations in such networks. We also notice that the degree distributions of the layers in the synthetic networks are compatible with the stretched exponential functional forms introduced and discussed in Ref. [105].

3.3.3 Model calibration for generic multiplex networks

We now discuss how to calibrate the model in the most general case in which the layers might possibly have different edge density, i.e. $m^{[1]} \neq m^{[2]}$, and different clustering, i.e. $p^{[1]} \neq p^{[2]}$. As an example, we consider the co-authorship networks of the same four sub-fields of physics (namely, Nuclear, Particle, Condensed Matter I and Interdisciplinary

Layers' pair	$\langle k^{[\beta]} \rangle$		$\langle Knn^{[\beta,\alpha]} \rangle$	
	γ_{data}	γ_{model}	γ_{data}	γ_{model}
Interd. given Cond. Matt. I	0.98	0.85	0.14	0.30
Thriller given Action	0.83	0.84	0.16	0.17
Romance given Action	0.89	0.87	0.30	0.27

Table 3-D: Quantitative comparison between the curves obtained from the model and the data for the inter-layer degree correlations and the mixed degree correlations. The curves have been fitted using a function of the form $f(x) \sim x^\gamma$; the γ parameter is reported for the corresponding curves in Figure 3.14.

physics) used to construct the four-layer APS multiplex (cf. Table 3-C and Figure 3.10). However, we focus here on all two-layer multiplex networks obtained by combining two networks at a time, so that, for instance, a node appears in the Nuclear-Particle (N-P) multiplex network if the corresponding author has published papers in both sub-fields. In general, the obtained multiplex networks are composed by layers with different edge density and different clustering coefficients, as shown in Table 3-E, thus we need to set separately the four parameters of the model $p^{[1]}$, $p^{[2]}$, $m^{[1]}$ and $m^{[2]}$.

We start by observing that the average degree of a synthetic layer is $\langle k \rangle \simeq 2m$, where m is the number of edge stubs connected by a newly arrived node, so that the parameters $m^{[1]}$, $m^{[2]}$ of the model can be set respectively equal to $\left\lfloor \frac{\langle k^{[1]} \rangle}{2} \right\rfloor$ and $\left\lfloor \frac{\langle k^{[2]} \rangle}{2} \right\rfloor$, where $\langle k^{[1]} \rangle$ and $\langle k^{[2]} \rangle$ are the measured average degrees of the two layers (numbers are approximated to the closest integers). Similarly, as we show in Figure 3.15(a), the clustering coefficient $C^{[\alpha]}$ of a layer α is univocally determined by $p^{[\alpha]}$, as soon as $m^{[\alpha]}$ is fixed. In Figure 3.15(a) we show how the values of $C^{[\alpha]}$ change as a function of $p^{[\alpha]}$, for different values of $m^{[\alpha]}$. Hence, the values of the intra-layer triadic closure parameters $p^{[1]}$ and $p^{[2]}$ can be set in order to match the values of clustering coefficient observed in each of the two layers. The only parameter yet to be determined is p^* . However, if we set the values of $m^{[1]}$, $m^{[2]}$, $p^{[1]}$, and $p^{[2]}$ to match the densities and clustering coefficients of the layers, we can then run the model for different values of p^* and look for the one which yields a value of NMI as close as possible to the one observed in the real two-layer multiplex. This procedure is sketched in Figure 3.15 (b) for the six two-layer multiplexes in APS.

In order to better understand the role of the different parameters, in Figure 3.15(c) we report the values of NMI obtained from different realisations of the model with $m^{[1]} = m^{[2]} = m$ and $p^{[1]} = p^{[2]} = p$ for m varying in $[2, 3, \dots, 10]$, and p varying in $[0, 0.1, \dots, 1]$ at different values of p^* , $[0.05, 0.5, 0.95]$, corresponding respectively to low, intermediate and high inter-layer triadic closure strength. We see that the effect of the increase in the link density m of the layers leads to a decrease in the similarity of their

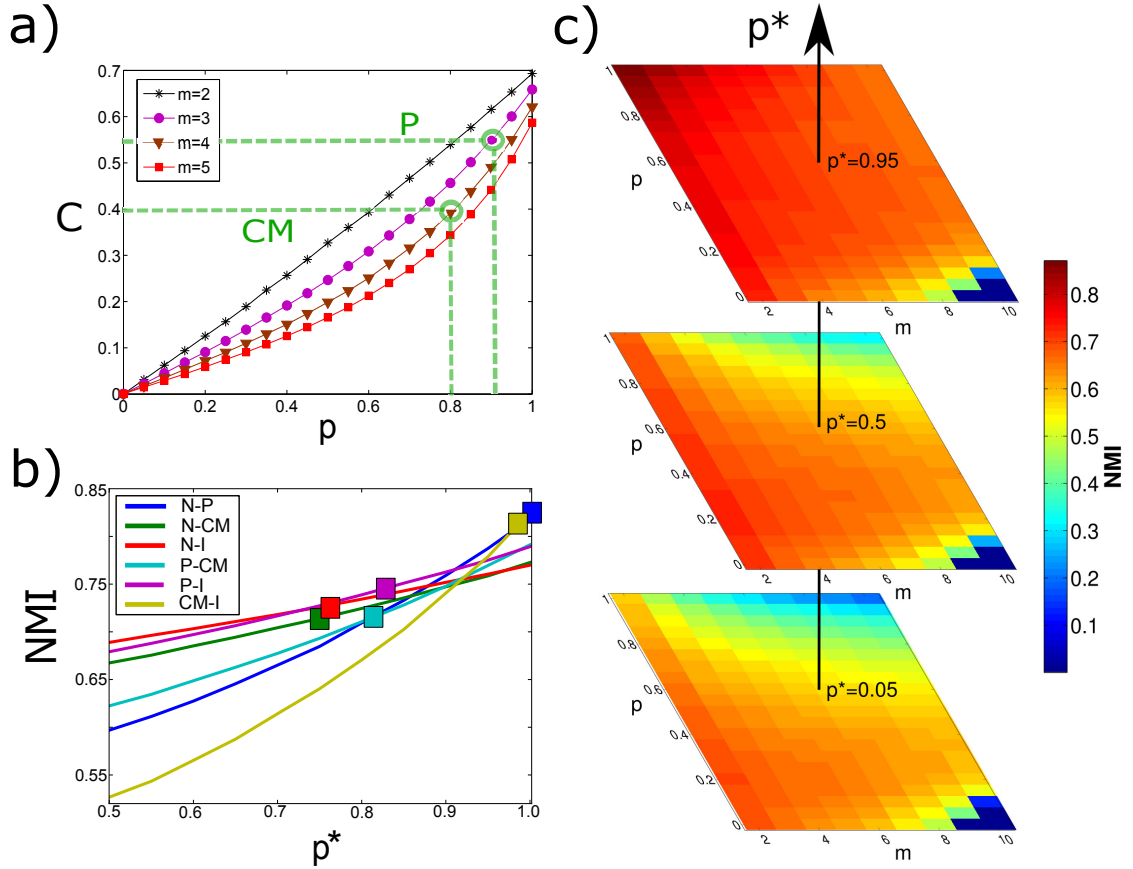


Figure 3.15: Model calibration. In panel a) we show the dependence of the clustering coefficient C on the intra-layer triadic closure parameter p for different values of the parameter m , which sets the layer's average degree. In the multiplex consisting of the layers Particle (P) and Condensed Matter I (CM), the average degree of each layer corresponds, respectively, to $m^{[1]} = 3$ and $m^{[2]} = 4$. The value of $p^{[1]}$ and $p^{[2]}$ are determined to match the clustering coefficients $C^{[1]}$ and $C^{[2]}$. In panel b), after having determined $m^{[1]}$, $m^{[2]}$, $p^{[1]}$ and $p^{[2]}$ for all the pairs of layers in the APS dataset, we run the model with such parameters for different value of p^* and infer, for each pair, the value of the inter-layer triadic closure parameter p^* yielding a value of NMI compatible with that observed (see Table 3-C for layers' acronyms). In panel c) we plot a heat-map of the NMI as a function of p and m , respectively for low (0.05), intermediate (0.50) and high (0.95) values of p^* in the model with $m^{[1]} = m^{[2]} = m$ and $p^{[1]} = p^{[2]} = p$. An increase in the link density of the layers produces a less correlated community structure in the two layers, even if the inter- and intra-layer triadic closure strengths are high.

community structures even for high values of p and p^* .

It is interesting to notice that, although the generic version of the model depends on five parameters, respectively accounting for layer density ($m^{[1]}$ and $m^{[2]}$), triadic closure ($p^{[1]}$ and $p^{[2]}$), and inter-layer overlap of communities (p^*), the values of those parameters can be easily set by measuring just the average degree and the average clustering

coefficient of each layer, and the normalised mutual information between the community structures at the two layers. Again, the good agreement between the synthetic networks and the real-world datasets extends also to other structural properties, such as intra-layer and inter-layer degree correlations, which were thought to have little or no direct relation at all with triadic closure. These results suggest that triadic closure plays an unexpectedly central role in determining the structural properties of real-world multiplex collaboration networks.

Layer 1	Layer 2	N	$\langle k^{[1]} \rangle$	$\langle k^{[2]} \rangle$	$C^{[1]}$	$C^{[2]}$	NMI
Nuclear	Particle	6572	6.88	7.46	0.56	0.56	0.83
Nuclear	Cond. Matt. I	3828	4.53	7.20	0.43	0.34	0.71
Nuclear	Interdisciplinary	2556	4.15	5.39	0.37	0.33	0.72
Particle	Cond. Matt. I	3774	5.70	7.82	0.53	0.40	0.71
Particle	Interdisciplinary	2502	4.82	5.66	0.49	0.39	0.74
Cond. Matt. I	Interdisciplinary	27257	10.34	7.05	0.55	0.64	0.82

Table 3-E: Basic properties of duplex networks in APS. We consider all the possible multiplex networks with $M = 2$ layers obtained from combinations of the APS collaboration networks corresponding to the four sub-fields Nuclear, Particle, Condensed Matter I and Interdisciplinary Physics. For each duplex, we report the number of nodes N , the average degree on the two layers $\langle k^{[1]} \rangle$ and $\langle k^{[2]} \rangle$, and the values of the clustering coefficients $C^{[1]}$ and $C^{[2]}$.

Chapter 4

Extracting network motifs from time series

4.1 Motifs in time series

As discussed in 1.1 one of the most interesting concepts that has emerged in Network Science is the one of network motifs [5]. These local topological features has proved to be very useful for classifying large networks in areas as biochemistry, neuroscience or ecology and for understanding the interplay between network's local structure and function [4, 5, 106, 107].

In the realm of time series analysis the concept of motif finds its natural counterpart in the idea of recurring pattern, a small subsequence of data showing a characteristic trend and occurring several times along the series [108, 109]. For example if we consider electroencephalogram (EEG) time series showing the activity of populations of neurons we can observe clinically meaningful patterns such as ‘alpha waves’ or ‘wicket spikes’ that are characteristic of the brain state (awakeness, sleep, lesions, etc.).

In [108] Lonardi et al. first defined the time series motifs as recurring subsequences (see Figure 4.1) and proposed an algorithm to extract them. Given a series $\{x_i\}_{i=1..N}$ and a subsequence $C = \{x_p, \dots, x_{p+n}\}_{n < N}$, one counts all the *match* subsequences of C , where a *match* M is a subsequence $M = \{x_q, \dots, x_{q+n}\}_{n < N}$ such that $D(C, M) < R$, D being a generic distance and R a threshold radius. Then the most significant motif in the series is the subsequence that has the highest count of non-trivial ($q \neq p$) matches and in general the k^{th} ranked motif has k matches.

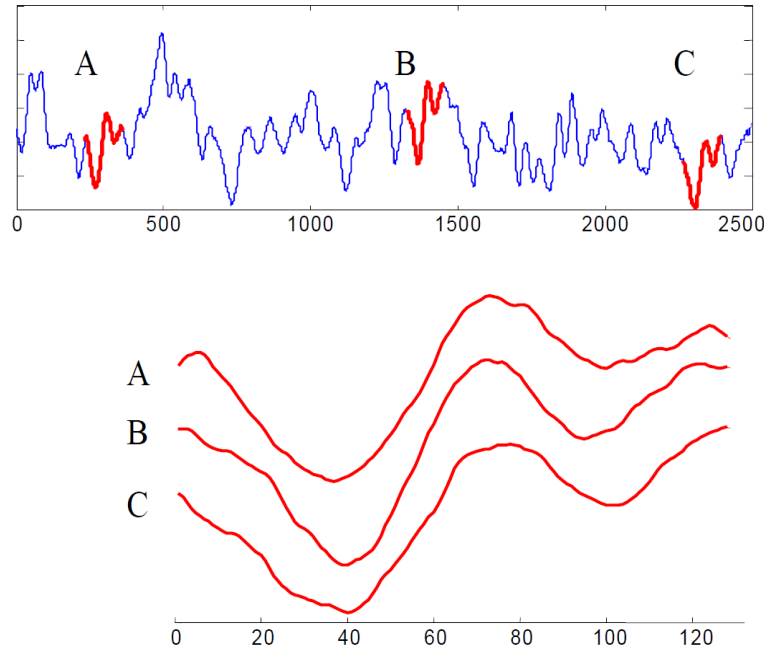


Figure 4.1: Time series motifs can be defined as recurrent characteristic patterns in the dynamics, as proposed in [108]. Figure taken from [108].

From a physical point of view the main problem of this approach, apart from the extraction algorithm which depends on an arbitrary parameter $R > 0$, is in the definition of motif itself. Let's imagine two subsequences of n data which show exactly the same trend but the values at each datum differ by a constant factor c . If $D = nc > R$ the algorithm will not match the two subsequences although they are probably the result of a very similar underlying dynamical process.

To make things rigorous one needs somehow to *symbolise* the series by using some alphabet that is able to capture relevant information from the data without dealing with the values *per se*. The most interesting approach in this sense is based on the study of the so called *ordinal patterns* [110], symbols that take into account only the relative rank of the data inside time series subsequences. Let's consider the case in which we want to symbolise a time series by considering the relative rank of data in subsequences of length two. Given the series $\{x_i\}_{i=1..N}$ we consider all the possible pairs of consecutive data $\{x_t, x_{t+1}\}$ and we write $\{0, 1\}$ if $x_t > x_{t+1}$ or $\{1, 0\}$ if $x_t < x_{t+1}$. We have thus at each time two possible ordinal symbols, namely $\{0, 1\}$ and $\{1, 0\}$, that we can use to map the original series into a symbols' series, and in general by considering a subsequence of length n our new alphabet will have $n!$ symbols. This method doesn't take into account the cases in which $x_t = x_{t+1}$, but the probability of finding two consecutive data of exactly the same value in real-valued time series is close to zero and for practical purposes a small amount of random noise can always be added to the series to get rid

of repeating data without altering its nature. Taking local rank numbers instead of the original values corresponds somehow to apply a kind of high-pass filter which removes a possible trend and very low frequencies from the time series, and the symbols defined in this way allow to derive analytically the probability of appearance of the ordinal patterns of size $n = 2, 3, 4, 5$ for a variety of different dynamical processes [111]. Moreover [110] the entropy of the of the n -length ordinal patterns s^n , given by $H^n = -\sum_{s^n} p(s) \log(p(s))$, where $p(s^n)$ are the frequency of appearance of the patterns, is a powerful indicator for discriminating random dynamics [110, 112]. Indeed for any random dynamics, one expects the distribution of the ordinal patterns to be uniform ($p(s^n) = 1/n!$) and to give maximal entropy $H^n = \log(n!)$, while a deterministic, regular behaviour in the series will give a small value of H .

Ordinal patterns are a clear example of how informative a process of symbolisation of the time series can become. In the following we will discuss the sequential visibility graph motifs [43], small subgraphs extracted from visibility graphs that allow for a time series symbolisation through network motifs. We will show that the statistics of the motif sequence is an informative feature to describe different types of complex dynamics such as chaotic and stochastic with different types of correlations and useful in the task of classifying empirical time series. For large classes of dynamical systems, it is possible to develop a theory to analytically compute the frequency of the visibility graph motifs.

4.2 Sequential visibility graph motifs

Let $\mathcal{S} = \{x(t)\}_{t=1}^T$ be a real-valued time series of T data. As previously discussed in 1.3 the natural visibility graph (VG) associated to \mathcal{S} is a planar graph of T nodes, such that (i) every datum $x(i)$ in the series is related to a node i in the graph (hence the graph nodes inherit a natural ordering), and (ii) two nodes i and j are connected by an edge if any other datum $x(k)$ where $i < k < j$ fulfils the following *convexity* criterion:

$$x_k < x_i + \frac{k-i}{j-i}[x_j - x_i], \quad \forall k : i < k < j$$

By construction, VGs are connected graphs with a natural Hamiltonian path given by the sequence of nodes $(1, 2, \dots, T)$, whose topology is invariant under a set of basic transformations in the series, including horizontal and vertical translations.

The horizontal visibility graph (HVG) associated to \mathcal{S} is defined as a subgraph of the VG, obtained by restricting the visibility criterion and imposing horizontal visibility instead. In this case, two nodes i and j are connected by an edge in the HVG if any other datum

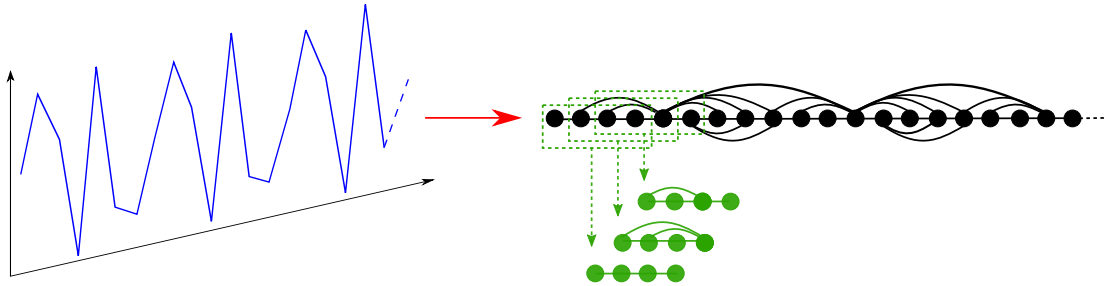


Figure 4.2: Schematic of visibility graph motif detection. A time series is converted into a visibility graph according to the visibility criterion (red arrow). A window of size $n = 4$ slides along the Hamiltonian path of the visibility graph and detects at each step a different VG motif.

$x(k)$ where $i < k < j$ fulfil the following *ordering* criterion:

$$x_k < \inf(x_i, x_j), \forall k : i < k < j$$

Such subgraph is indeed an outerplanar graph [113] and inherits some of the properties of VGs. We can now introduce a new topological property of VG/HVG.

Definition (*sequential VG/HVG n -node motifs*). Consider a VG/HVG of N nodes, associated to a time series of N data, and label the nodes according to the natural ordering induced by the arrow of time (i.e. the trivial Hamiltonian path). Set $n < N$ and consider, sequentially, all the subgraphs formed by the sequence of nodes $\{s, s + 1, \dots, s + n - 1\}$ (where s is an integer that takes values in $[1, N - n + 1]$) and the edges from the VG/HVG only connecting these nodes: these are defined as the sequential n -node motifs of the VG/HVG. This is akin to defining a sliding window of size n in graph space that initially covers the first n nodes and sequentially slides, in such a way that for each window, one can associate a motif by (only) considering the edges between the n nodes belonging to that window (see Figure 4.2).

Note that, importantly, this definition differs from the one of a standard network motif (which looks at the frequencies of appearance of all subgraphs of a given size, without imposing any restriction on the nodes forming a given subgraph), as here it is required that the labels of the nodes appearing in a motif are in strict sequential order -this is consistent with the vertex ordering of the natural Hamiltonian path induced by construction in the VGs/HVGs-. That is, in order to preserve in graph space the dynamical information of the series, the n nodes of an n -size motif are taken in sequential order, and only those edges that connect nodes from the motif are considered. For readability, from now on we will call these simply VG/HVG motifs but the reader should not get

confused and remind that these are not directly the standard notion of network motifs computed on a VG/HVG. Some basic properties of these motifs are:

- Trivially, there is a total of $N - n$ motifs (which can be the same motifs or not) within each VG/HVG.
- Each motif is a subgraph of the original VG/HVG. Moreover, HVG motifs are outerplanar and have a trivial Hamiltonian path, thus HVG motifs are also HVGs [113]. As a result, there are only 6 admissible motifs of size 4, and 2 admissible motifs of size 3 (see Table 4-A for an enumeration).
- Computational complexity: Computing motifs in both VG and HVG is extremely efficient. If instead of exploring the motif occurrence in the structure of the adjacency matrix, one directly examines the set of inequalities reported in Table 4-A, one directly has an algorithm that runs in *linear* time $O(N)$ for HVG motifs. A similar complexity is found for VG motifs [114].

As is done traditionally with network motifs [106], we can compare VG/HVGs associated to different time series and dynamics by comparing the relative occurrence of each motif inside a VG/HVG. In order to do that, we introduce the extension to the VG/HVG realm of a significance profile:

Definition (*VG/HVG motif profile \mathbf{Z}^n*). Let p be the total number of admissible VG/HVG motifs with n -nodes. Assign to each of these p motifs a label from 1 to p (that is, choose an ordering for the motifs). The motif assigned with the label i will be called a type- i motif. Then, we define the n -node VG/HVG motif significance profile \mathbf{Z}^n (or simply HVG motif profile) of a certain time series of size N as the vector function $\mathbf{Z}^n : n \in \mathbb{N} \rightarrow [\mathbb{P}_1^n, \dots, \mathbb{P}_p^n] \in [0, 1]^p$ whose output is a vector of p components, where the i -th component, \mathbb{P}_i^n , is the relative frequency of the type- i motif.

Several technical comments are in order:

- First, since \mathbf{Z}^n are n -dimensional real vectors, any L_p norm induces a natural similarity measure (distance) between two graphs.
- Second, \mathbf{Z}^n has, by construction, unit L_1 -norm, as $\sum_{i=1}^p |\mathbb{P}_i^n| = \sum_{i=1}^p \mathbb{P}_i^n = 1$.
- Third, note that if one considers dynamical processes instead of individual time series, then the estimated relative frequencies \mathbb{P}_i^n for an individual realization of the dynamical process converge for infinitely long series to the probabilities of

type- i motif associated to the process. For the motif profile to be a well-defined feature of a certain dynamical process, it needs to be self-averaging. We check this property by estimating \mathbf{Z}^n for an ensemble of realizations of the process, computing the mean $\langle \mathbb{P}_i^n \rangle$ and standard deviation $\sqrt{\langle [\mathbb{P}_i^n]^2 \rangle - \langle \mathbb{P}_i^n \rangle^2}$ over this ensemble, and checking that the standard deviation is small (meaning that a single realization provides a good description of the average behaviour). As we will show below, both VG and HVG motif profiles have very good self-averaging properties. In any case, for every dynamical process considered in this work, instead of \mathbb{P}_i^n we compute $\langle \mathbb{P}_i^n \rangle$ and $\sqrt{\langle [\mathbb{P}_i^n]^2 \rangle - \langle \mathbb{P}_i^n \rangle^2}$, but for readability, from now on we will drop the $\langle \cdot \rangle$ for the elements of the motif profile, as we found that for the size of the series used in the numerical analysis, $\sqrt{\langle [\mathbb{P}_i^n]^2 \rangle - \langle \mathbb{P}_i^n \rangle^2}$ was very small and hence $\mathbb{P}_i^n \approx \langle \mathbb{P}_i^n \rangle$.

- Fourth, note at this point that the definition of the VG/HVG motif profile is different from standard profiles (significance profile, subgraph ratio profile) defined in the literature [106], as in the latter case, they make use of a null model (ensemble of randomised networks) to appropriately normalise each frequency. The rationale for this normalization is that one wants to compare motif statistics across very different networks (with different sizes and degree sequences), so variations in the motif relative frequencies only due to size effects need to be removed to be able to correctly compare across different networks. In the context of VG/HVG the null model is not a randomised ensemble of the graph under study (which would not yield a VG/HVG with high probability), but on the contrary, it should be the VG/HVG of a randomisation of the *time series* under study. In other words, normalisation in the case of VG/HVG profiles should deal with the motif statistics of uncorrelated random series (i.i.d. white noise or surrogate series that preserve certain structures) with similar probability densities than the series under study. In the next Section we will prove that, in the case of HVGs (which will be the family of visibility graphs under study), such null model has a universal motif profile, independent of the probability density of the i.i.d. process. Therefore, it is not necessary in this case to normalise each profile accordingly as this would only yield a trivial, constant rescaling.

For illustration purposes, let $n = 4$, and consider two different dynamical processes: (i) white Gaussian noise described by the map $x_t = \xi$, where ξ are independent and identically distributed (i.i.d.) Gaussian random variables $\xi \sim \mathcal{N}[0, 1]$, and (ii) chaotic dynamics given by the fully chaotic logistic map $x_{t+1} = 4x_t(1 - x_t)$. In order to estimate the probability of appearance of each of the motifs, we have generated a time series of size $N = 10^4$ data for both processes (sample time series can be seen in the top panels of figure 4.3), and we have computed the relative frequencies of each motif. Results, averaged

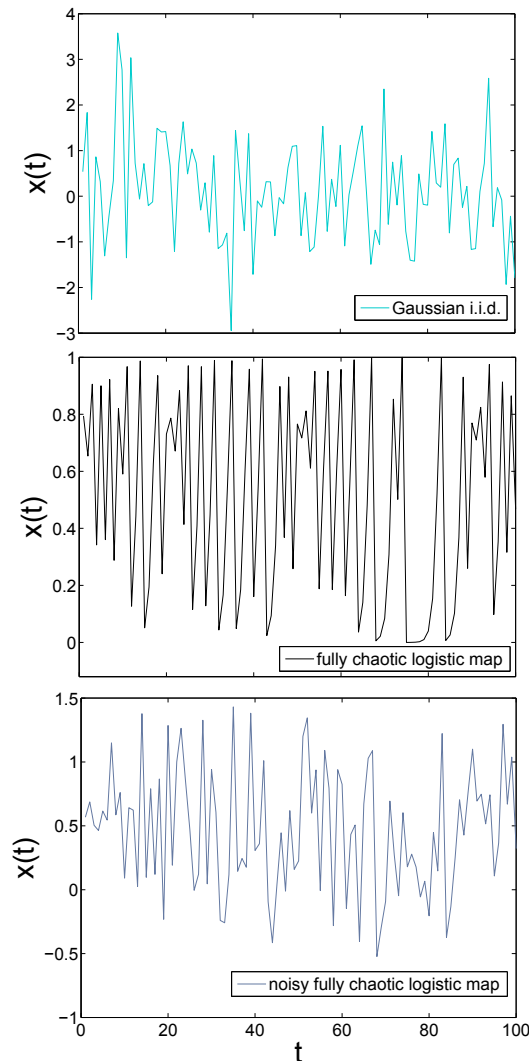


Figure 4.3: Sample time series from (a) i.i.d. Gaussian white noise, (b) fully chaotic logistic map, and (c) fully chaotic logistic map polluted with a certain amount of extrinsic white noise are shown for illustrative purpose. Visibility graph motifs can be extracted from these series to reveal differences in their intrinsic structure.

over an ensemble of 100 realizations, are shown in Figure 4.4 (error bars describing the ensemble standard deviation are contained inside the symbols); in panel (a) we plot the HVG motif profile, whereas in panel (b) we plot the VG profile. As we can see, in every case the type-II motif is absent. The simple reason is that this profile is absent for irregular (aperiodic) real-valued time series, by construction (see Table 4-A).

For the chaotic process, some other motifs are absent: this is related to forbidden patterns arising in chaotic dynamics. More importantly, in both panels, the average relative frequency of some motifs seems to be different for both dynamical processes,

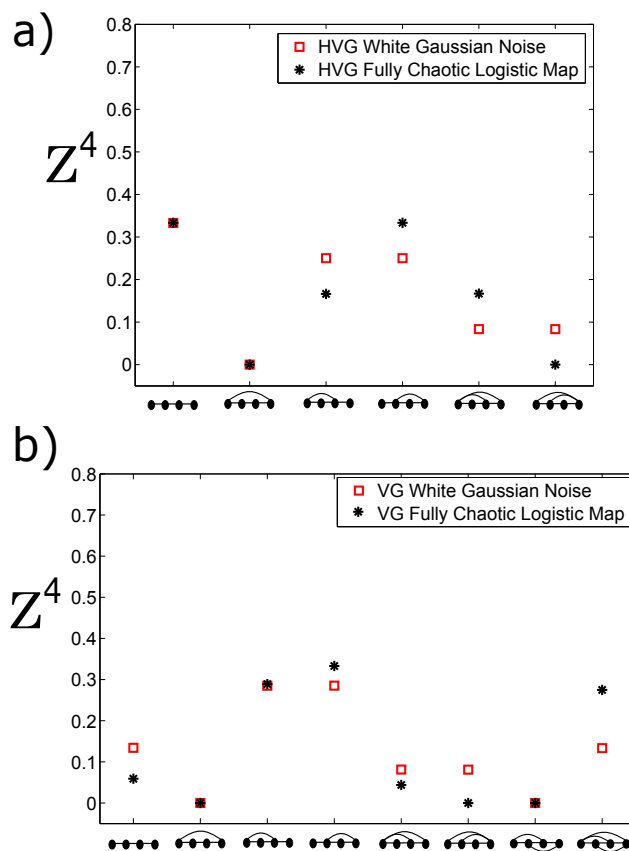


Figure 4.4: 4-node motif profiles Z^4 associated to Gaussian white noise (red squares) and to a fully chaotic logistic map (black stars) extracted respectively from HVG (panel a) and VG (panel b). Each dot represents the relative frequency of a given motif, averaged over an ensemble of 100 realizations of each process (time series of $N = 10^4$ data per realization). Standard deviations of each motif relative frequency over the ensemble are plotted as error bars, which are not visible as error bars fall inside the symbols. We conclude that these motifs can be used to distinguish between deterministic and stochastic dynamics.

enabling the possibility of using both HVG and VG motif profiles to distinguish amongst different dynamical origins. In the following Sections we advance a theory to compute the motif profile Z^n in an exact way for different classes of dynamical systems.

4.3 Theory for sequential HVG motifs

In order to numerically explore and compute the frequency of each HVG motif, one can generate the HVG associated to a given time series and count the presence of each motif directly from the adjacency matrix. Here, however, we will show that it is not

Motif label	Motif type	Inequality set
1		$\{\forall(x_0, x_2), x_1 > x_0\} \cup \{\forall x_0, x_1 < x_0, x_2 < x_1\}$
2		$\{\forall x_0, x_1 < x_0, x_2 > x_1\}$
1		$\{\forall(x_0, x_1), x_2 < x_1, x_3 < x_2\} \cup \{\forall(x_0, x_3), x_1 > x_0, x_2 > x_1\}$
2		$\{\forall x_0, x_1 < x_0, x_2 = x_1, x_3 > x_2\}$
3		$\{\forall x_0, x_1 < x_0, x_1 < x_2 < x_0, x_3 < x_2\} \cup \{\forall(x_0, x_3), x_1 < x_0, x_2 > x_0\}$
4		$\{\forall x_0, x_1 > x_0, x_2 < x_1, x_3 > x_2\} \cup \{\forall x_0, x_1 < x_0, x_2 < x_1, x_2 < x_3 < x_1\}$
5		$\{\forall x_0, x_1 < x_0, x_1 < x_2 < x_0, x_3 > x_2\}$
6		$\{\forall x_0, x_1 < x_0, x_2 < x_1, x_3 > x_1\}$

Table 4-A: Enumeration of all 3 and 4-node motifs. Each motif can be characterized according to a hierarchy of inequalities in the associated time series. Note that for real-valued aperiodic dynamics the type-II 4-node motif has a null probability of occurrence as the probability that two data in the time series repeat vanishes almost surely (if, on the other hand, the series only take values from a finite set then this motif has a finite probability). For the rest, the probability of each motif reduces to the measure of the set of inequalities.

necessary to do that as, via the zero-order terms of a diagrammatic expansion recently advanced [115], we can work out the motif occurrence directly from the exploration of the time series, that enables motif computation in linear time. This will allow us to build a theory by which the motif profiles can be computed exactly for a large set of classes of dynamics that fulfil certain properties.

Let us consider a dynamical process $\mathcal{H} : \mathbb{R} \rightarrow \mathbb{R}$ with a smooth invariant measure $f(x)$ that fulfils the Markov property, that means that conditional probabilities fulfil $f(x_n|x_{n-1}, x_{n-2}, \dots) = f(x_n|x_{n-1})$, where $f(x_n|x_{n-1})$ is the transition probability distribution. Note that this concept has a clear meaning in random dynamical systems, whereas for deterministic systems, say maps $x_{t+1} = \mathcal{H}(x_t)$, the Markov property is also trivially fulfilled with $f(x_2|x_1) = \delta(x_2 - \mathcal{H}(x_1))$, where $\delta(x)$ is the Dirac-delta distribution. The key element is that for these processes, each HVG motif has a probability of appearance as a subgraph that can *directly* be computed as the measure of a *set of ordering inequalities* that take place in the time series. For instance, for $n = 3$ and $n = 4$, probabilities associated to the appearance of a certain motif are based on integrals of the form:

$$\int f(x_0)dx_0 \int f(x_1|x_0)dx_1 \int f(x_2|x_1)dx_2 \quad (4.1)$$

for $n = 3$, and

$$\int f(x_0)dx_0 \int f(x_1|x_0)dx_1 \int f(x_2|x_1)dx_2 \int f(x_3|x_2)dx_3 \quad (4.2)$$

for $n = 4$.

The range of integration and the shape of the conditional probabilities are particular for each motif and each process, respectively. First, the range of integration fully deter-

mines the motif. In Table 4-A we depict the conditions in the time series that have to be fulfilled among n consecutive data x_0, x_1, \dots, x_{n-1} to yield a certain motif of size n in the HVG, for $n = 3, 4$ (extension to arbitrary n is easy but gets cumbersome as n increases). It can be proved quite easily that a given motif appears in an HVG if and only if these ordering restrictions are fulfilled in the time series. These restrictions directly translate in the integration range of the probabilities, we illustrate this principle in an example. The first motif, \mathbf{Z}_1^4 , according to Table 4-A is guaranteed when 4 consecutive values x_0, x_1, x_2 and x_3 are such that $\{\forall(x_0, x_1), x_2 < x_1, x_3 < x_2\} \cup \{\forall(x_0, x_3), x_1 > x_0, x_2 > x_1\}$. Accordingly, if $x \in [a, b] \subset \mathbb{R}$, the probability of this event is

$$\begin{aligned} \mathbf{Z}_1^4 \equiv \mathbb{P}_1^4 = & \int_a^b f(x_0) dx_0 \int_a^b f(x_1|x_0) dx_1 \int_a^{x_1} f(x_2|x_1) dx_2 \int_a^{x_2} f(x_3|x_2) dx_3 + \\ & \int_a^b f(x_0) dx_0 \int_{x_0}^b f(x_1|x_0) dx_1 \int_{x_1}^b f(x_2|x_1) dx_2 \int_a^b f(x_3|x_2) dx_3. \end{aligned} \quad (4.3)$$

Analogous expressions can be found for the rest of the probabilities that form the motif profile \mathbf{Z} . These terms are nothing but the contributions to the degree distribution at zero-order from a diagrammatic expansion in the number of hidden nodes [115]. From a geometric point of view, the first motif will not appear in fast fluctuating signals and hence deals with the degree of smoothness of a time series at short (order n) scales, whereas the other motifs deal with certain fluctuation shapes. Accordingly, in those processes where the degree of smoothness can vary -such as in fractional Brownian motion, where the smoothness of the signal increases with the Hurst exponent- we would expect that the first motif is particularly informative, whereas for fast-fluctuating series we expect this motif to be less informative. Integrals accounting for the probabilities are easy to deal with; in several cases these are exactly solvable, and in general one can solve them up to arbitrary precision with any symbolic programming software. In what follows we determine the motif profiles for i.i.d. (white noise), coloured noise with exponentially decaying correlations, and deterministic chaos (fully chaotic logistic map). We show that \mathbf{Z}^4 capture enough information to easily distinguish different processes and thus represent excellent features for series classification.

Relation with ordinal patterns. At this point it is important to highlight the relation between the probability of occurrence of a given HVG motifs and the probability of occurrence of so called ordinal patterns [110, 111]. In the theory proposed by Bandt and Pompei [110] for the case of the embedding dimension equal to 4 one proceeds to map each local time series segment of size 4 into an ordering symbol of 4 letters from the alphabet $\{0, 1, 2, 3\}$ (where the largest value maps to the letter 0, the second largest

to 1, the third largest to 2, and the smallest to 3). There are $4! = 24$ permutations, defining 24 symbols (ordinal patterns) whose frequencies are then counted to measure the so-called permutation entropy that acts as a complexity measure of the series [110]. Interestingly, the probability of occurrence of each HVG motif indeed reduces to the probability of occurrence of a set of possible ordinal patterns (this is no longer the case for VG motifs [114]). For instance, \mathbf{Z}_1^4 is the probability of finding any of the ordinal patterns 0123, 1023, 1203, 1230, 2103, 2130, 2310, or 3210, and similarly the rest of the motif probabilities can be linked to the probability of appearance of different sets of ordinal patterns. Accordingly, HVG motifs indeed induce a particular partition of the set of ordinal patterns. The HVG motif profile is thus intimately linked with the so called permutation [112] that accounts for the histogram of ordinal patterns. On the other hand VGs are not in general invariant under monotonic transformations in the series [116] (note however they are invariant under linear ones), and they depend on the marginal probability distribution of the time series. Thus they are not an order statistic and, accordingly, there is no obvious correspondence between n -OPs and VG n -motifs and both approaches in principle represent two independent symbolization methods that encode temporal information in a different way. It has to be remarked that both VG and HVG motif analysis can be applied without requiring any further assumption to time series taking values from finite sets (namely when $P(x_t = x_{t+1}) \neq 0$), while the ordinal patterns approach -based uniquely on the ranking statistics- require further assumptions in that case.

4.3.1 Random dynamics: i.i.d.

Let us start by considering time series generate by i.i.d. uniform random variables $\xi \sim U[0, 1]$. In this case we have $a = 0, b = 1, f(x) = 1$ and $f(x|y) = f(x) \forall y$, and simply enough, probabilities defined by Eqs. 4.1 and 4.2 easily factorize. According to Table 4-A, after a little bit of calculus we find

$$\mathbf{Z}^3 = \left[\frac{2}{3}, \frac{1}{3} \right]; \quad \mathbf{Z}^4 = \left[\frac{8}{24}, 0, \frac{6}{24}, \frac{6}{24}, \frac{2}{24}, \frac{2}{24} \right] \quad (4.4)$$

Note that these results are in perfect quantitative agreement with numerics performed for finite size series (top panel of Figure 4.4); we will show in the next Subsection that results for finite series converge quite fast to the (asymptotic) theory as the series size increases. Interestingly, results indeed coincide despite the fact that the theoretical values were computed for *uniform* white noise ($f(x) = 1$), while the numerics in Figure 4.4 were performed on *Gaussian* white noise (where $f(\cdot)$ is the Gaussian function). This suggests that i.i.d. may have a universal HVG motif profile, indeed independent of $f(\cdot)$.

We now state and prove a theorem that actually guarantees this result.

Theorem 1. *Consider a bi-infinite series of i.i.d. random variables extracted from a continuous distribution $f(x)$ with support (a, b) , where $a, b \in \mathbb{R}$. Then the probability of finding n -node HVG motifs (with $n = 3, 4$) follows Eq. 4.4, independently of the shape of $f(x)$.*

Proof. The proof is a constructive one. We only give here the explicit proof for \mathbb{P}_1^4 , as the proof for the rest of probabilities follow analogously. We rely on the cumulative distribution function $F(x)$, defined as $\int_a^x f(x')dx' = F(x)$, with properties $F(a) = 0, F(b) = 1$ and

$$f(x)F^{n-1}(x) = \frac{dF^n(x)}{ndx}. \quad (4.5)$$

We have

$$\begin{aligned} \mathbb{P}_1^4 = & \int_a^b f(x_0)dx_0 \int_a^b f(x_1)dx_1 \int_a^{x_1} f(x_2)dx_2 \int_a^{x_2} f(x_3)dx_3 + \\ & \int_a^b f(x_0)dx_0 \int_{x_0}^b f(x_1)dx_1 \int_{x_1}^b f(x_2)dx_2 \int_a^b f(x_3)dx_3 \end{aligned}$$

Using the properties of $F(x)$, the first term above is then

$$\begin{aligned} & \int_a^b f(x_0)dx_0 \int_a^b f(x_1)dx_1 \int_a^{x_1} f(x_2)dx_2 \int_a^{x_2} f(x_3)dx_3 = \\ & \int_a^b f(x_0)dx_0 \int_a^b f(x_1)dx_1 \int_a^{x_1} f(x_2)F(x_2)dx_2 = \\ & \int_a^b f(x_0)dx_0 \int_a^b f(x_1) \frac{F^2(x_1)}{2} dx_1 = \\ & \int_a^b \frac{f(x_0)}{6} dx_0 = \frac{1}{6}, \end{aligned}$$

and analogously for the second term,

$$\begin{aligned}
& \int_a^b f(x_0)dx_0 \int_{x_0}^b f(x_1)dx_1 \int_{x_1}^b f(x_2)dx_2 \int_a^b f(x_3)dx_3 = \\
& \int_a^b f(x_0)dx_0 \int_{x_0}^b f(x_1)(1 - F(x_1))dx_1 = \\
& \int_a^b f(x_0) \left[\frac{1}{2} - F(x_0) + \frac{F^2(x_0)}{2} \right] dx_0 = \\
& \left. \frac{F(x_0)}{2} - \frac{F^2(x_0)}{2} + \frac{F^3(x_0)}{6} \right|_a^b = \frac{1}{6},
\end{aligned} \tag{4.6}$$

hence $\mathbb{P}_1^4 = 2/6 = 8/24$, coinciding with the result for uniform and Gaussian series, and being independent of $f(x)$. The rest of the elements in \mathbf{Z}^4 are computed analogously. ■

As a matter of fact, the independence from $f(x)$ can be trivially extended for an arbitrary size of the motif n . This is intuitive so we only give here the strategy of a proof. The main ingredient which is required for this independency to hold $\forall n$ is that the limits of the n -th integral are either the extremes of the distribution support a, b (where the cumulative distribution $F(x)$ take the constant values 0 and 1 respectively, and independently of $f(x)$), or other variables $x_0 \dots x_{n-1}$. In this latter case, one can use iteratively the property in Eq. 4.5 to solve these integrals up to the last one (in x_0), whose range is always (a, b) and where $F(a) = 0$, $F(b) = 1$ can be finally applied, to give a result which will not depend on the precise shape of $f(x)$.

According to theorem 1, Gaussian, uniform, power law, etc, uncorrelated random series all have the same HVG motif profiles. As a by-product, for any kind of sufficiently long time series $\{x_t\}_{t=1}^N$ where $x_t \in f(x)$ and $f(x)$ is continuous, if we randomize (shuffle) the time series, the motif profile of the randomized series is equal to Eq. 4.4. This is the reason why, at odds with the standard definition of a network's motif profile, for HVGs we don't need to rescale \mathbf{Z} in any way to be able to compare across different time series and dynamical process.

Another notable consequence of theorem 1 is that it guarantees that series for which \mathbf{Z}^4 differ (even in the case of sufficiently long time series) from Eq. 4.4 are not uncorrelated random series. This suggests a simple test for randomness [40]. For instance, one can use a Pearson's χ^2 hypothesis test, where the null hypothesis is that the observed time

series of N data is random and uncorrelated (white noise). The test statistic is then

$$\chi^2 = (N - n) \sum_{i=1}^p \frac{[\mathbb{P}_i^4(\text{observed}) - \mathbb{P}_i^4(\text{i.i.d.})]^2}{\mathbb{P}_i^4(\text{observed})} \quad (4.7)$$

χ^2 upper-critical values with $p - 1$ degrees of freedom, for $p = 6$ ($n = 4$) are 11.07 and 15.086 at the 95% and 99% significance level (meaning that values of the χ^2 larger than 11.07 suggest that the observed series is not random at the 95% significance level). More rigorously, as type-II motif is forbidden for aperiodic dynamics, we have only $p = 5$ different motifs of size $n = 4$, so the χ^2 upper-critical values should be considered for 4 degrees of freedom: 9.49 (95%) and 13.28 (99%).

4.3.2 Deterministic chaos: fully chaotic logistic map

As previously stated, deterministic maps $x_{t+1} = \mathcal{H}(x)$ are indeed Markovian, and for these situations the conditional probability is simply $f(x_2|x_1) = \delta(x_2 - \mathcal{H}(x_1))$, where $\delta(x)$ is the Dirac-delta distribution. Therefore Eqs. 4.1 and 4.2, combined with inequality sets given in Table 4-A can be used to compute the motif profiles for different deterministic processes. In these cases, one has to deal with simple integrals of the form

$$\int_p^q \delta(x - y) dx = \begin{cases} 1 & y \in [p, q] \\ 0 & \text{otherwise} \end{cases} \quad (4.8)$$

While in principle any deterministic process can be studied, we are interested in complex signals, so we focus on irregular, aperiodic dynamics. As a paradigmatic case, we tackle the fully chaotic logistic map

$$\mathcal{H}(x) = 4x(1 - x), \quad x \in [0, 1], \quad f(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

In this case, $f(x)$ is the invariant measure that describes in a probabilistic way the average time spent by a chaotic trajectory in each region of the attractor. Let us start by considering $\mathbf{Z}^3 := (\mathbb{P}_1^3, \mathbb{P}_2^3)$, for which

$$\begin{aligned} \mathbb{P}_1^3 &= \int_0^1 f(x_0) dx_0 \int_{x_0}^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2, \\ \mathbb{P}_2^3 &= \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_1}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2. \end{aligned}$$

According to property in Eq. 4.8, the Dirac-delta integrals only have the effect of shrinking the range of integration of x_0 . For instance, for \mathbb{P}_1^3 , the integral in x_1 requires

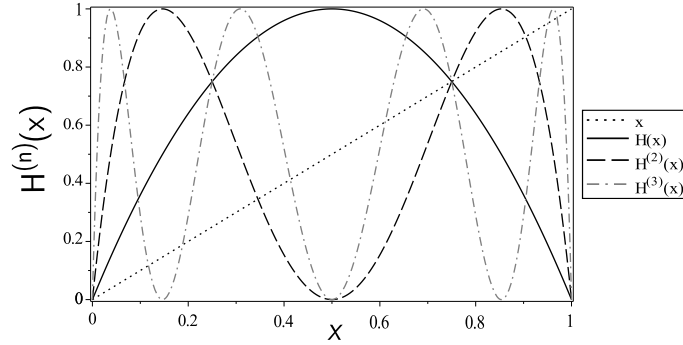


Figure 4.5: Cobweb plot of the iterates of the fully chaotic logistic map $\mathcal{H}(x) = 4x(1-x)$.

$\mathcal{H}(x_0) > x_0$, whereas the integral in x_2 simply requires $\mathcal{H}^2(x_0) \in [0, 1]$. While the latter inequality is fulfilled for all $x_0 \in [0, 1]$ (and thus has no effect), the former one requires $x_0 \in [0, 3/4]$. This can be easily seen from the cobweb plot of $\mathcal{H}(x)$ and its iterates (see Figure 4.5): $\mathcal{H}(x) > x$ for $x \in [0, 3/4]$. Altogether,

$$\mathbb{P}_1^3 = \int_0^{3/4} f(x_0) dx_0 = 2/3$$

On the other hand, motif normalization imposes $\mathbb{P}_2^3 = 1/3$. The same result is obviously found if we compute \mathbb{P}_2^3 explicitly: in this case the integral in x_1 requires $\mathcal{H}(x_0) < x_0$, which holds when $x_0 \in [3/4, 1]$, and the integral in x_2 requires $\mathcal{H}^2(x_0) > x_1 \leftrightarrow \mathcal{H}^2(x_0) > \mathcal{H}(x_0)$. Looking at the cobweb plots, this final condition is met in two subintervals, so the intersection with the first condition yields a final interval $x_0 \in [3/4, 1]$, for which

$$\mathbb{P}_2^3 = \int_{3/4}^1 f(x_0) dx_0 = 1/3,$$

as expected. These results coincide with those found for i.i.d. series, meaning that \mathbf{Z}^3 doesn't capture enough structure to distinguish both processes. Let us proceed in an equivalent way to compute $\mathbf{Z}^4 = (\mathbb{P}_1^4, \dots, \mathbb{P}_6^4)$. It becomes evident that integrals associated to x_n deal with the cobweb plots of $\mathcal{H}(x), \mathcal{H}^2(x), \dots, \mathcal{H}^n(x)$. Accordingly, these integrals are ultimately related with the structure of fixed points of $\mathcal{H}^n(x)$, and with the solutions of equations of the form $\mathcal{H}^r(x) = \mathcal{H}^s(x)$ for some r and s . We only have algebraic closed expressions for the fixed points of $\mathcal{H}(x) \rightarrow \{0, 3/4\}$ and $\mathcal{H}^2(x) \rightarrow \{0, \frac{5-\sqrt{5}}{8}, 3/4, \frac{5+\sqrt{5}}{8}\}$ (for $n \geq 3$, $\mathcal{H}^n(x)$ is a polynomial of order larger or equal to 6 and according to Abel-Ruffini's theorem, the set of fixed points does not have in general an algebraic expression, however we can compute them up to arbitrary precision). Other values of interest include the roots of $\mathcal{H}^3(x) = \mathcal{H}^2(x)$, and specially the largest one $x = 1/2 + \sqrt{3}/4$.

Let us show how to compute one of these motif probabilities. For instance,

$$\mathbb{P}_5^4 = \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_1}^{x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{x_2}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3 \quad (4.9)$$

which reduces to

$$\mathbb{P}_5^4 = \int_p^q f(x_0) dx_0,$$

where $[p, q]$ can be hierarchically obtained as:

$$\mathcal{H}(x_0) < x_0 \cap [0, 1] \Rightarrow x_0 \in [3/4, 1];$$

$$\mathcal{H}^2(x_0) < x_0 \cap \mathcal{H}^2(x_0) > \mathcal{H}(x_0) \cap [3/4, 1] \Rightarrow x_0 \in [\frac{5+\sqrt{5}}{8}, 1];$$

$$\mathcal{H}^3(x_0) > \mathcal{H}^2(x_0) \cap [\frac{5+\sqrt{5}}{8}, 1] \Rightarrow x_0 \in [x_p, 1], \text{ where } x_p \text{ is the second largest root fulfilling } \mathcal{H}^3(x_p) = \mathcal{H}^2(x_p), \text{ i.e. } x_p = 1/2 + \sqrt{3}/4. \text{ Altogether,}$$

$$\mathbb{P}_5^4 = \int_{1/2+\sqrt{3}/4}^1 \frac{1}{\pi \sqrt{x_0(1-x_0)}} dx_0 = \frac{1}{\pi} B_{[\frac{1}{2}+\frac{\sqrt{3}}{4}, 1]} \left(\frac{1}{2}, \frac{1}{2} \right) = \frac{1}{6} (= 4/24)$$

(where B is the incomplete Beta function), which is indeed quite different from the result found for i.i.d., $\mathbb{P}_5^4(\text{i.i.d.}) = 2/24$.

Similar arguments can be used to obtain analytically the rest of probabilities (explicit computations can be found in Appendix B), finding

$$\mathbf{Z}^3 = \left[\frac{2}{3}, \frac{1}{3} \right]; \quad \mathbf{Z}^4 = \left[\frac{8}{24}, 0, \frac{4}{24}, \frac{8}{24}, \frac{4}{24}, 0 \right] \quad (4.10)$$

Comparing this set of motif probabilities with the result for i.i.d. (Eq. 4.4), we can conclude that \mathbf{Z}^4 distinguishes the fully chaotic logistic map from a purely uncorrelated stochastic process. Note, of course, that a similar derivation can be performed in other deterministic maps; in this sense the methodology is general (however one encounters problems when the attractor has a fractal dimension, and one needs to carefully choose a proper integration theory). These exact results are also in excellent quantitative agreement with numerics performed in finite series (top panel of Figure 4.4), so convergence to the theory with series size is quite fast, enabling its use in empirical cases. To be more precise, in the next Subsection we make a study of how fast results for short time series converge to the asymptotic theory as series size increases.

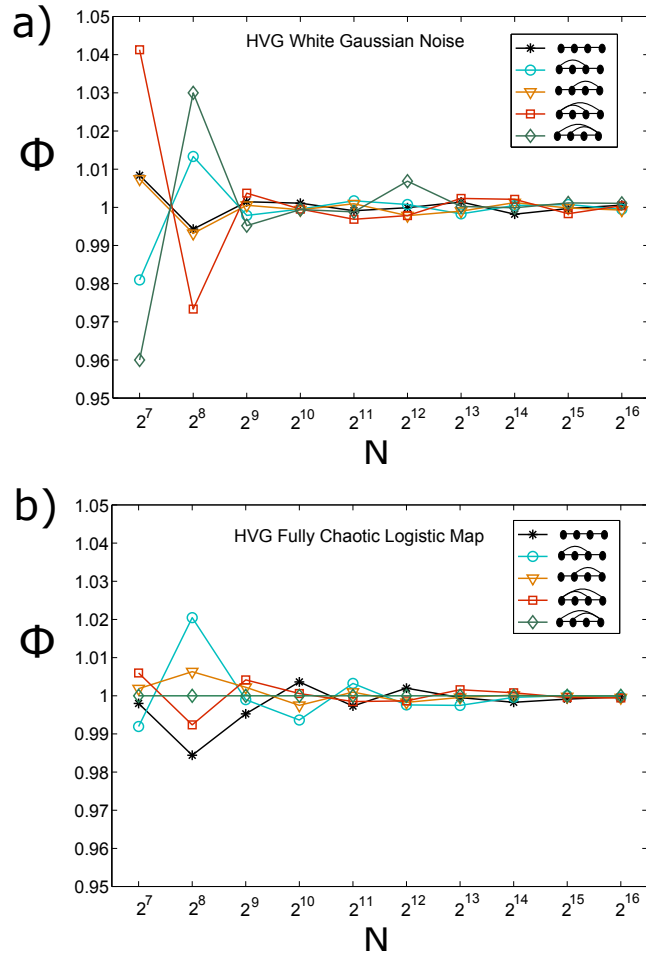


Figure 4.6: The measured frequency of appearance rescaled by its theoretical value Φ is plotted for each motif associated to Gaussian white noise (panel a) and to a fully chaotic logistic map (panel b) in function of the time series size N ; results are averaged over 100 realisations. The curves oscillate with fast decreasing amplitude around the value 1 (for 2^9 the amplitude is less than 10^{-2}) indicating fast asymptotic convergence of the measured motif profile to the theoretical profile in both cases.

4.3.3 Convergence of finite series

In order to be more precise about the convergence speed of finite-size numerics to the theory (which in rigour only holds for bi-infinite time series), we have computed for series of size N the numeral estimate $\mathbf{Z}^4(N)$ for both i.i.d. and the fully chaotic logistic map, and compare it with the asymptotic values \mathbf{Z}^4 . Results are plotted in Figure 4.6, where we plot $\Phi(N) = \langle \mathbf{Z}^4(N) \rangle / \mathbf{Z}^4$ as a function of the series size N (the average is with respect to realizations). Results indicate that convergence to the asymptotic theory is already reached for $N \ll 10^4$ (which is the conservative size that is used all over this work).

4.3.4 Stochastic processes with correlations

To round off the theory section, and to explore how results deviate from i.i.d. for correlated stochastic processes, we consider coloured noise with exponentially decaying correlations as described by the AR(1) process:

$$\begin{cases} x_0 = \xi_0 \\ x_t = rx_{t-1} + \sqrt{(1-r^2)}\xi_t, \quad t \geq 1 \end{cases} \quad (4.11)$$

where $\xi_t \sim \mathcal{N}(0, 1)$ is Gaussian white, and r , $0 < r < 1$ is a parameter that tunes the correlation. The auto-correlation function $C(t)$, which describes the correlation of the position at x_{t_0} and x_{t_0+t} decays exponentially $C(t) = e^{-t/\tau}$, where the characteristic time $\tau = 1/\ln(r)$. In the limit $r \rightarrow 0$, the correlations vanish and the process reduces to a white noise signal. The limit $r \rightarrow 1$ is more delicate, but intuitively in this limit the process gets completely correlated and tends to be constant $x_{t+1} = x_t \forall t$.

This is a family of models parametrized by the coefficient r . For $0 < r < 1$, these models are indeed Gaussian, Markovian and stationary, with a probability density $f(x)$ and transition probability $f(x_2|x_1)$ are

$$\begin{cases} f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \\ f(x_2|x_1) = \frac{\exp[-(x_2-rx_1)^2/(2(1-r^2))]}{\sqrt{2\pi(1-r^2)}} \end{cases}$$

respectively. Since x are Gaussian variables they can vary in $(-\infty, \infty)$. We focus on \mathbf{Z}^4 that we know gave good discriminatory results between i.i.d. and chaos. For illustration, the first element reads

$$\begin{aligned} \mathbb{P}_1^4 = & \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{(x_1-rx_0)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_1 \int_{-\infty}^{x_1} \frac{e^{-\frac{(x_2-rx_1)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_2 \int_{-\infty}^{x_2} \frac{e^{-\frac{(x_3-rx_2)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_3 + \\ & \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{x_0}^{\infty} \frac{e^{-\frac{(x_1-rx_0)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_1 \int_{x_1}^b \frac{e^{-\frac{(x_2-rx_1)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_2 \int_{-\infty}^{\infty} \frac{e^{-\frac{(x_3-rx_2)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_3 \end{aligned} \quad (4.12)$$

For any particular value of r , these integrals can be evaluated up to arbitrary precision using Mathematica [117]. In table Table 4-B we report the theoretical values of $\mathbf{Z}^4(r)$ for $r \in [0.02 - 0.99]$. These are in perfect agreement with numerical simulations performed on finite series of size $N = 10^4$ (ensemble averaged over 100 realizations) for $r = \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$, as shown in Figure 4.7. As $r > 0$ the profiles deviate from i.i.d. and thus, again, these features can easily distinguish between exponentially

r	P_1^4	P_2^4	P_3^4, P_4^4	P_5^4, P_6^4
0.02	0.3370	0	0.2482	0.0833
0.04	0.3406	0	0.2464	0.0833
0.06	0.3443	0	0.2446	0.0832
0.08	0.3478	0	0.2429	0.0831
0.1	0.3514	0	0.2412	0.0830
0.2	0.3690	0	0.2333	0.0822
0.3	0.3862	0	0.2260	0.0809
0.4	0.4030	0	0.2192	0.0793
0.5	0.4196	0	0.2130	0.0772
0.6	0.4359	0	0.2072	0.0748
0.7	0.4521	0	0.2018	0.0722
0.8	0.4681	0	0.1967	0.0692
0.9	0.4841	0	0.1920	0.0660
0.95	0.4919	0	0.1897	0.0643
0.97	0.4945	0	0.1888	0.0636
0.99	0.4973	0	0.1879	0.1879

Table 4-B: Theoretical values of $Z^4(r)$ for the AR(1) process evaluated at different values of the coefficient r .

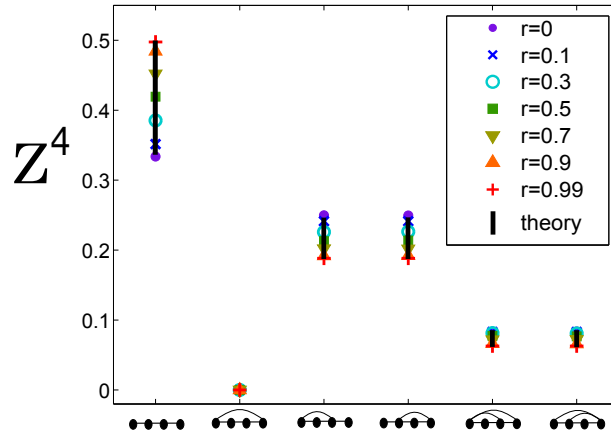


Figure 4.7: HVG significance profile Z^4 for AR(1) processes described by Eq. 4.11, for different values of the correlation coefficient r . When r increases the appearance probability of motif of type-I increases while the rest of probabilities decrease. This is simply due to the fact that finding constant sequences $x_{t+3} = x_{t+2} = x_{t+1} = x_t$ becomes more probable as r increases.

coloured and white noise.

4.4 Theory for sequential VG motifs

We saw that the set of sequential VG motifs of size n , $n \in [2, 3, \dots, N]$ is defined as the set of all the M_n possible sub-graphs with n consecutive vertices along the Hamiltonian path of a VG (similarly, the set of HVG motifs of size n is the set of all the M_n^h admissible sub-graphs with n consecutive vertices along the Hamiltonian path of a HVG). For $n = 4$, there are in principle a total of $M_4 = 8$ possible motifs (see table 4-C for an enumeration), although as we will show below the number of admissible ones is just six. We use the notation Φ_m to indicate the probability of appearance of a certain VG motif m (to make a distinction from the HVG motifs probabilities), and accordingly the VG n -motif profile is $Z^n = [\Phi_1, \dots, \Phi_{M_n}]$.

Let us consider again a time-discrete (deterministic or stochastic) dynamical process $x_{t+1} = \mathcal{H}(x_t, \xi)$, where ξ is a generic stochastic term, that fulfils the Markov property: $\forall l f(x_l|x_{l-1}, x_{l-2}, \dots) = f(x_l|x_{l-1})$, where $f(x_l|x_{l-1})$ is the transition probability distribution and $x \in (a, b)$. For deterministic processes $f(x_l|x_{l-1}) = \delta(x_l - \mathcal{H}(x_{l-1}))$ where $\delta(x)$ is the Dirac-delta distribution where $\delta(x)$ is the Dirac-delta distribution:

$$\int_p^q \delta(x - y) dx = \begin{cases} 1 & y \in [p, q] \\ 0 & \text{otherwise} \end{cases} \quad (4.13)$$

and $f(x)$ is a smooth invariant measure of the process $\mathcal{H}(x)$, whereas for stochastic processes $f(x)$ is simply the underlying probability density, i.e. the marginal distribution of the process. Our theory addresses the motif profile Z^4 , in what follows we split this analysis in two cases, depend whether x is bounded or unbounded. In both cases, each probability Φ_m^4 is computed formally using concatenated integrals which are formally equivalent to Eq. 4.2, where the ranges of each integral are given according to the convexity criteria defining the visibility rule, as opposed to the HVG case where these were simply ordering criteria. Whereas in the case of unbounded variables the inequality set will only take into account the visibility criteria within the motifs, in the case of bounded variables the additional restriction of variables needing to be bounded adds a layer of complexity as we will see. From now on, let $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$ be four arbitrary consecutive data ($l \in [1, N - 3]$).

Label	VG motif type	Inequality set
1		$\{\forall(x_l, x_{l+1}), x_{l+2} < 2x_{l+1} - x_l, x_{l+3} < 2x_{l+2} - x_{l+1}\}$
2		$\{\emptyset\}$
3		$\{\forall(x_l, x_{l+1}), x_{l+2} > 2x_{l+1} - x_l, x_{l+3} < \frac{3}{2}x_{l+2} - \frac{1}{2}x_l\}$
4		$\{\forall(x_l, x_{l+1}), x_{l+2} < 2x_{l+1} - x_l, 2x_{l+2} - x_{l+1} < x_{l+3} < 3x_{l+1} - 2x_l\}$
5		$\{\forall(x_l, x_{l+1}), x_{l+2} > 2x_{l+1} - x_l, \frac{3}{2}x_{l+2} - \frac{1}{2}x_l < x_{l+3} < 2x_{l+2} - x_{l+1}\}$
6		$\{\forall(x_l, x_{l+1}), x_{l+2} < 2x_{l+1} - x_l, x_{l+3} > 3x_{l+1} - 2x_l\}$
7		$\{\forall(x_l, x_{l+1}), x_{l+2} > 2x_{l+1} - x_l, x_{l+3} > 2x_{l+2} - x_{l+1}\}$
8		$\{\emptyset\}$

Table 4-C: The set of size-4 VG motifs are defined according to a set of relations between 4 consecutive data $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$, $l \in [1, N - 3]$ in the time series.

4.4.1 Unbounded variable $x \in (-\infty, \infty)$

In the case of unbounded variables, it is easy to prove [43, 115] that in general

$$\Phi_m^n = \int_{\mathbb{R}} f(x_l) dx_l \int_{d_1^m(x_l)}^{c_1^m(x_l)} f(x_{l+1}|x_l) dx_{l+1} \dots \int_{d_{n-1}^m(x_l, \dots, x_{l+n-2})}^{c_{n-1}^m(x_l, \dots, x_{l+n-2})} f(x_{l+n-1}|x_{l+n-2}) dx_{l+n-1} \quad (4.14)$$

where $\{c_i^m(\cdot)\}_{i=1, \dots, n-1}$ and $\{d_i^m(\cdot)\}_{i=1, \dots, n-1}$ are the set of functions which specify respectively the upper bound condition and the lower bound condition for the i -th integral. As advanced, these conditions are directly related to the visibility criterion (which in the VG case is a convexity relation) and a summary of those are explicitly reported in table 4-C. For example, if we want to build motif 1 using 4 consecutive equispaced data, we need that, given the first two generic values x_l, x_{l+1} (node l always connected with node $l+1$), the third datum x_{l+2} has to satisfy the relation $x_{l+2} - x_{l+1} < x_{l+1} - x_l \rightarrow x_{l+2} < 2x_{l+1} - x_l$ (node $l+2$ connected with node $l+1$ but not with node l) and that the fourth datum x_{l+3} has to satisfy the relation $x_{l+3} - x_{l+2} < x_{l+2} - x_{l+1} \rightarrow x_{l+3} < 2x_{l+2} - x_{l+1}$ (node $l+3$ connected only with node $l+2$). The equivalence between each motif and its associated inequality set can be proved rigorously also for the other motifs in an analogous way. First, note that for the motifs 2 and 8 the inequality set is empty. This means that these motifs are actually not admissible under the VG algorithm. In the case of motif number 2, note that this motif was an admissible one for HVGs associated to discrete-valued series where the probability of finding equal consecutive data is finite. For VG, it is easy to prove that if the bounding nodes share an edge, then either the left edge or the right edge will necessary share an edge with one of the inner nodes, thus motif 2 is not a VG and is therefore not occurring. Similarly, it is easy to prove that if a time series gives rise to a motif of type 8, then an edge would necessarily appear between the two bounding nodes, reducing this to type 7. Accordingly, the number of admissible motifs is not 8 but 6, and thus the

effective number of degrees of freedom associated to $n = 4$ VG motifs is $M_n - 2 - 1 = 5$. To better understand the application of the inequality set in the case of unbounded variables, consider a white Gaussian process ($x_i \in (-\infty, \infty)$) with

$$f(x_i) = \frac{\exp(-x_i^2/2)}{\sqrt{2\pi}} \quad \text{and} \quad f(x_{i+1}|x_i) = f(x_{i+1})$$

the probability of appearance of motif 1 in Table 4-C can be written explicitly as

$$\Phi_1^4 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} dx_1 \int_{-\infty}^{2x_1-x_0} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} dx_2 \int_{-\infty}^{2x_2-x_1} \frac{e^{-\frac{x_3^2}{2}}}{\sqrt{2\pi}} dx_3. \quad (4.15)$$

This integral cannot be solved in closed form but can be easily evaluated up to arbitrary precision using symbolic computation, to obtain $\Phi_1^4 \simeq 0.13386$.

4.4.2 Bounded variables $x \in [a, b]$; $a, b \in \mathbb{R}$; $a, b < \infty$

In the case where $x \in (a, b)$ where the bounds $a, b \in \mathbb{R}$ are finite, these restrictions in turn induce further conditions on the lower and upper bounds of the integrals in Eq.4.14 which have the effect of splitting the overall integral in a sum of different integrals. For illustrative purposes we start by considering a particular example. Consider a series of i.i.d. uniform random variables $x_i \sim \mathcal{U}[a, b]$ with $f(x) = (b-a)^{-1}$ and $f(x_{i+1}|x_i) = f(x_{i+1})$, and let us consider again Φ_1^4 . According to table 4-C, in principle the conditions for the first two variables x_0, x_1 are $\forall(x_0, x_1) \in [a, b]$; for the third variable x_2 the lower bound condition becomes $x_2 > a$ but the upper bound condition will depend on the function $2x_1 - x_0$ which can take values in $[2a - b, 2b - a]$ and thus we need to consider three different cases:

$$\left\{ \begin{array}{l} 2x_1 - x_0 > b \implies x_1 \in (\frac{(x_0+b)}{2}, b], \quad x_2 \in [a, b] \\ a < 2x_1 - x_0 < b \implies x_1 \in (\frac{(x_0+a)}{2}, \frac{(x_0+b)}{2}], \quad x_2 \in [a, 2x_1 - x_0] \\ 2x_1 - x_0 < a \quad \{\emptyset\} \end{array} \right. \quad (4.16)$$

where the last case doesn't contribute (as $x_2 > a$ is always fulfilled). Similarly for each admissible choice of the variable x_2 the bound conditions for x_3 will produce an

additional split:

$$\left\{ \begin{array}{l} 2x_2 - x_1 > b \implies x_2 \in \left(\frac{(x_1+b)}{2}, b\right], \quad x_3 \in [a, b] \\ a < 2x_2 - x_1 < b \implies x_2 \in \left(\frac{(x_1+a)}{2}, \frac{(x_1+b)}{2}\right], \quad x_3 \in [a, 2x_2 - x_1] \\ 2x_2 - x_1 < a \quad \{\emptyset\} \end{array} \right. \quad (4.17)$$

After a bit of algebra one finds

$$\begin{aligned} \Phi_1^4 = & \frac{1}{(b-a)^4} \left[\int_a^b dx_0 \int_{\frac{(x_0+b)}{2}}^b dx_1 \int_{\frac{(x_1+b)}{2}}^b dx_2 \int_a^b dx_3 + \right. \\ & + \int_a^b dx_0 \int_{\frac{(x_0+b)}{2}}^b dx_1 \int_{\frac{(x_1+a)}{2}}^{\frac{(x_1+b)}{2}} dx_2 \int_a^{2x_2-x_1} dx_3 + \\ & + \int_a^b dx_0 \int_{\frac{(2x_0+b)}{3}}^{\frac{(x_0+b)}{2}} dx_1 \int_{\frac{(x_1+b)}{2}}^{2x_1-x_0} dx_2 \int_a^b dx_3 + \\ & + \int_a^b dx_0 \int_{\frac{(2x_0+a)}{3}}^{\frac{(2x_0+b)}{3}} dx_1 \int_{\frac{(2x_1+a)}{2}}^{2x_1-x_0} dx_2 \int_a^{2x_2-x_1} dx_3 + \\ & \left. + \int_a^b dx_0 \int_{\frac{(2x_0+b)}{3}}^{\frac{(x_0+b)}{2}} dx_1 \int_{\frac{(x_1+a)}{2}}^{\frac{(x_1+b)}{2}} dx_2 \int_a^{2x_2-x_1} dx_3 \right] = \frac{5}{36} \simeq 0.1389. \end{aligned} \quad (4.18)$$

Two comments are in order. First, note that this result is different from the value for Φ_1^4 for Gaussian white noise, that is, the results for white noise seems to be dependent on the marginal distribution of the noise. This lack of invariance was expected as VG is not an order statistic, and differs from the phenomenology found for HVG, where results for white noise were universal (independent from the marginal distribution). This evidence will be confirmed in the next sections. Second, this motif profile turns to be independent from the bounding values a and b (where $x \in [a, b]$). This is due to the invariance properties of VGs: if the random variable $\xi \sim \mathcal{U}[0, 1]$, then $a + (b-a)\xi \sim \mathcal{U}[a, b]$, and the transformation $\xi \rightarrow a + (b-a)\xi$, when $b > a$, leaves the VG (and hence the VG motifs) unaltered. This is indeed a peculiarity of the uniform distribution, in other words the VG motif profile of white noise extracted from a bounded distribution *generally* depends on the bounds of the distribution.

After a bit of algebra, we are able to translate the visibility and bounded variable restrictions inside each motif into another set of inequalities, which we have reported in table 4-D. In what follows we make use of this theory to compute the theoretical VG motif profile for several dynamical processes.

VG motif Inequality set

••••	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [\frac{(x_{l+1}+b)}{2}, b], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [\frac{(x_{l+1}+b)}{2}, 2x_{l+1} - x_l], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, 2x_{l+1} - x_l], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$
⌒	$\{\emptyset\}$
⌒	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+b)}{2}], x_{l+2} \in [\frac{(x_l+2b)}{3}, b], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_l+2a)}{3}, \frac{(x_l+2b)}{3}], x_{l+3} \in [a, \frac{3x_{l+2}-x_l}{2}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [2x_{l+1} - x_l, b], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_l+2b)}{3}, b], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_l+2b)}{3}], x_{l+3} \in [a, 3x_{l+1} - 2x_l]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(x_l+2a)}{3}, \frac{(x_l+2b)}{3}], x_{l+3} \in [a, 3x_{l+1} - 2x_l]\}$
⌒	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+b)}{2}, b], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, 2x_{l+1} - x_l], x_{l+3} \in [2x_{l+2} - x_{l+1}, 3x_{l+1} - 2x_l]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, b], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, 3x_{l+1} - 2x_l]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+b)}{3}, \frac{(x_l+b)}{2}], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$
⌒	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_{l+1}+2a)}{3}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(2x_l+a)}{3}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_l+2a)}{3}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_l+2a)}{3}], x_{l+3} \in [a, 2x_{l+2} - x_{l+1}]\}$
	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(2x_l+b)}{3}], x_{l+2} \in [\frac{(x_{l+1}+b)}{2}, \frac{(x_l+2b)}{3}], x_{l+3} \in [\frac{3x_{l+2}-x_l}{2}, b]\}$
⌒	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [a, 2x_{l+1} - x_l], x_{l+3} \in [3x_{l+1} - 2x_l, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [a, 2x_{l+1} - x_l], x_{l+3} \in [a, b]\}$
⌒	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [a, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [a, \frac{(x_l+a)}{2}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_{l+1}+a)}{2}], x_{l+3} \in [a, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(x_l+a)}{2}, \frac{(2x_l+a)}{3}], x_{l+2} \in [\frac{(x_{l+1}+a)}{2}, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, b]\}$
	$\{x_l \in [a, b], x_{l+1} \in [\frac{(2x_l+a)}{3}, \frac{(2x_l+b)}{3}], x_{l+2} \in [2x_{l+1} - x_l, \frac{(x_{l+1}+b)}{2}], x_{l+3} \in [2x_{l+2} - x_{l+1}, \frac{(x_{l+1}+b)}{2}]\}$
⌒	$\{\emptyset\}$

Table 4-D: Sets of inequalities between 4 consecutive data $\{x_l, x_{l+1}, x_{l+2}, x_{l+3}\}$, $l \in [1, N - 3]$ in a time series of length N which define the VG motifs of size 4 in the case of bounded variables $x_i \in [a, b]$.

4.4.3 Fully chaotic logistic map

We start by considering the fully chaotic logistic map $\mathcal{H}(x) = 4x(1 - x)$, $x \in [0, 1]$, with invariant density $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$. As this process is deterministic, it fulfils a trivial Markov property such that $f(x_2|x_1) = \delta(x_2 - \mathcal{H}(x_1))$. The HVG motif profile for this process was computed exactly in [43], here we compute the VG motif profile. Before proceeding to compute each probability contribution, it is important to highlight a subtle point. Since for this process $x \in [0, 1]$ is bounded, in principle one should use the inequality set depicted for bounded variables in Table 4-D. However, in this particular

case it is actually not necessary to explicitly consider the restriction $x \in [0, 1]$. As we will see in a moment, this is already taken into account implicitly in the computation of each integral and therefore one can use the (simpler) inequality set for unbounded variables given in table 4-C.

We start by computing Φ_1^4 :

$$\Phi_1^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{2x_1 - x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^{2x_2 - x_1} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) < 2\mathcal{H}^2(x_0) - \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) < 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.1743, 0.25]$. Note at this point that the latter conditions are also satisfied in other ranges, but we only consider those ranges that belong to $[0, 1]$, and this is indeed the reason why we don't need to use in this case the inequality set for bounded variables. We thus have

$$\Phi_1^4 \simeq \frac{1}{\pi} B_{[0.1743, 0.25]} \left(\frac{1}{2}, \frac{1}{2} \right) \simeq 0.0591$$

where B is the incomplete Beta function. As $\Phi_2^4 = 0$ by construction, we proceed by calculating Φ_3^4 :

$$\Phi_3^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{2x_1 - x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^{\frac{3}{2}x_2 - \frac{1}{2}x_0} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) < \frac{3}{2}\mathcal{H}^2(x_0) - \frac{1}{2}\mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) > 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.0522, 0.1743] \cup [0.75, 0.929]$. Therefore:

$$\Phi_3^4 \simeq \frac{1}{\pi} (B_{[0.0522, 0.1743]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[0.75, 0.929]} \left(\frac{1}{2}, \frac{1}{2} \right)) \simeq 0.289.$$

Similarly for Φ_4^4 we have

$$\Phi_4^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{2x_1 - x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{2x_2 - x_1}^{3x_1 - 2x_0} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$2\mathcal{H}^2(x_0) - \mathcal{H}(x_0) < \mathcal{H}^3(x_0) < 3\mathcal{H}^2(x_0) - 2x_0$$

$$\mathcal{H}^2(x_0) < 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.25, 0.75]$, and therefore

$$\Phi_4^4 \simeq \frac{1}{\pi} B_{[0.25, 0.75]} \left(\frac{1}{2}, \frac{1}{2} \right) \simeq 0.3333$$

For Φ_5^4 we have

$$\Phi_5^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{2x_1 - x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{\frac{3}{2}x_2 - \frac{1}{2}x_0}^{2x_2 - x_1} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\frac{3}{2}\mathcal{H}^2(x_0) - \frac{1}{2}x_0 < \mathcal{H}^3(x_0) < 2\mathcal{H}^2(x_0) - \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) > 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0.927, 0.954]$, and thus

$$\Phi_5^4 \simeq \frac{1}{\pi} (B_{[0.04568, 0.05224]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[0.9239, 0.9543]} \left(\frac{1}{2}, \frac{1}{2} \right)) \simeq 0.0439$$

For Φ_6^4 we have

$$\Phi_6^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{2x_1 - x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{3x_1 - 2x_0}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) > 3\mathcal{H}(x_0) - 2x_0$$

$$\mathcal{H}^2(x_0) < 2\mathcal{H}(x_0) - x_0$$

which are never satisfied and thus

$$\Phi_6^4 = 0.$$

For Φ_7^4 we have

$$\Phi_7^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{2x_1 - x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{2x_2 - x_1}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

which gives the following conditions:

$$\mathcal{H}^3(x_0) > 2\mathcal{H}^2(x_0) - \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) > 2\mathcal{H}(x_0) - x_0$$

which are satisfied for $x_0 \in [0, 0.0457] \cup [0.9544, 1]$, and thus

$$\Phi_7^4 \simeq \frac{1}{\pi} (B_{[0, 0.046]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[0.95, 1]} \left(\frac{1}{2}, \frac{1}{2} \right)) \simeq 0.2741$$

Finally, by construction $\Phi_8^4 = 0$. Altogether, we find the VG motif profile of a fully

chaotic logistic map

$$\mathbf{Z}^4 = \left[0.0591, 0, 0.289, 0.3333, 0.0439, 0, 0.2741, 0 \right] \quad (4.19)$$

Note that while the result is in this case an approximation, our theory allows for numerical estimates with arbitrary precision (the result is not exact because the location of fixed points of the map is only approximate, although this approximation is arbitrarily close to the true values).

4.4.4 Uniform white noise

For white uniform noise $x(t) = \xi$, $\xi \sim U[a, b]$ we have a probability density $f(x)$ and transition probability $f(x_2|x_1)$ given by

$$f(x) = \frac{1}{b-a}, \text{ and } f(x_2|x_1) = f(x_2) \quad (4.20)$$

In this case the computations are more cumbersome since we need to make use of the inequality set for bounded variables described in Table 4-D. The m th component of \mathbf{Z}^4 is given by

$$\Phi_m^4 = \sum_s \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{d_{s1}^m}^{c_{s1}^m} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} dx_1 \int_{d_{s2}^m}^{c_{s2}^m} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} dx_2 \int_{d_{s3}^m}^{c_{s3}^m} \frac{e^{-\frac{x_3^2}{2}}}{\sqrt{2\pi}} dx_3 \quad (4.21)$$

where the sum runs over all the set s of conditions d_{si}^m, c_{si}^m which contribute to evaluate the probability for motif m in Table 4-D. All the integrals can nonetheless be solved analytically in closed form and give the following motif profile

$$\mathbf{Z}^4 = \left[\frac{5}{36}, 0, \frac{31}{108}, \frac{31}{108}, \frac{2}{27}, \frac{2}{27}, \frac{5}{36}, 0 \right] \quad (4.22)$$

which differs from the one found for the chaotic logistic map.

4.4.5 Gaussian white noise

For standard white Gaussian noise $x(t) = \xi$, $\xi \sim \mathcal{N}(0, 1)$ the probability density $f(x)$ and transition probability $f(x_2|x_1)$ are given by

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}, \text{ and } f(x_2|x_1) = f(x_2) \quad (4.23)$$

The m component of \mathbf{Z}^4 is given by

$$\Phi_m^4 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{x_1^2}{2}}}{\sqrt{2\pi}} dx_1 \int_{d_2^m}^{c_2^m} \frac{e^{-\frac{x_2^2}{2}}}{\sqrt{2\pi}} dx_2 \int_{d_3^m}^{c_3^m} \frac{e^{-\frac{x_3^2}{2}}}{\sqrt{2\pi}} dx_3 \quad (4.24)$$

where (d_i^m, c_i^m) are now the top and bottom conditions for the variable x_{l+i} in motif m reported in Table for unbounded variables 4-C. The integrals can be evaluated numerically up to arbitrary precision and they give the following results

$$\mathbf{Z}^4 = \left[0.13386, 0, 0.2850, 0.2850, 0.0811, 0.0811, 0.13386, 0 \right] \quad (4.25)$$

At odds with what happens for HVG motifs [43], this result is different from the benchmark result for uniformly distributed white noise, thus there is not a universal VG motif profile for white noise as previously anticipated.

4.4.6 Gaussian red noise

Gaussian colored (red) noise with exponentially decaying correlations [118] can be simulated using an $AR(1)$ process:

$$x_t = rx_{t-1} + \xi \quad (4.26)$$

where $\xi \sim \mathcal{N}(0, 1)$ is Gaussian white, and $0 < r < 1$ is a parameter that tunes the correlation. The auto-correlation function $C(t)$ decays exponentially $C(t) = e^{-t/\tau}$, where the characteristic time $\tau = 1/\ln(r)$. This model is Markovian and stationary, with a probability density $f(x)$ and transition probability $f(x_2|x_1)$ given by [118]

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}, \text{ and } f(x_2|x_1) = \frac{\exp[-(x_2 - rx_1)^2/(2(1-r^2))]}{\sqrt{2\pi(1-r^2)}} \quad (4.27)$$

The m component of \mathbf{Z}^4 is given by

$$\Phi_m^4 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x_0^2}{2}}}{\sqrt{2\pi}} dx_0 \int_{-\infty}^{\infty} \frac{e^{-\frac{(x_1 - rx_0)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_1 \int_{d_2^m}^{c_2^m} \frac{e^{-\frac{(x_2 - rx_1)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_2 \int_{d_3^m}^{c_3^m} \frac{e^{-\frac{(x_3 - rx_2)^2}{2(1-r^2)}}}{\sqrt{2\pi(1-r^2)}} dx_3 \quad (4.28)$$

where, again, (d_i^m, c_i^m) are respectively the bottom and the top conditions for the variable x_{l+i} in motif m reported in table Table 4-C. Once we set the precise values of the parameter r , the profile can be evaluated numerically up to arbitrary precision; here we

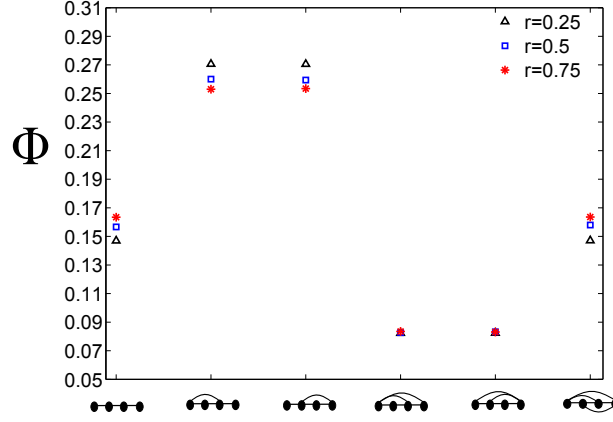


Figure 4.8: Average (50 realisations) frequency of appearance Φ of VG motifs extracted from the AR(1) processes described by Eq. 4.26, for different values of the correlation coefficient $r = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}]$ (only motifs with a non-null probability are shown). Error bars are contained in the symbols; results are in very good agreement with the theoretical expected value.

give the profile for three possible values $r = 1/4, 1/2$ and $3/4$)

$$\begin{aligned}
 \mathbf{Z}_{r=\frac{1}{4}}^4 &= \left[0.14713, 0, 0.27028, 0.27028, 0.08259, 0.08259, 0.14713, 0 \right] \\
 \mathbf{Z}_{r=\frac{1}{2}}^4 &= \left[0.15731, 0, 0.2595, 0.2595, 0.08316, 0.08316, 0.15731, 0 \right] \\
 \mathbf{Z}_{r=\frac{3}{4}}^4 &= \left[0.16410, 0, 0.25258, 0.25258, 0.08332, 0.08332, 0.16410, 0 \right]
 \end{aligned} \tag{4.29}$$

In all these examples, theoretical results are in very good agreement with the results obtained from numerical simulations reported in Figure 4.8.

4.4.7 Noise characterisation

Differently from the HVG motifs, VG motifs statistics does not depend uniquely on the ranking statistics of the data and therefore the VG motif profile could be able in principle to discriminate white noises with different marginal distributions. In the latter sections we have been able to distinguish between Gaussian and uniform white noise. In Figure 4.9 we summarize the motif frequencies Φ_m^4 of VG motifs forming Z^4 , extracted

from i.i.d. series with different marginals:

$$\begin{cases} \text{Uniform} \rightarrow x_i \in [0, 1]; & f(x_i) \sim 1 \\ \text{Gaussian} \rightarrow x_i \in (-\infty, \infty); & f(x_i) \sim \frac{\exp(-x_i^2/2)}{\sqrt{2\pi}} \\ \text{Power-law} \rightarrow x_i \in [1, \infty); & f(x_i) \sim x_i^{-k}, \quad k = 2.5 \\ \text{Exponential} \rightarrow x_i \in [0, \infty); & f(x_i) \sim \exp(-kx_i), \quad k = 2.5 \end{cases} \quad (4.30)$$

In every case we extract series of 10^5 data. The universal profile obtained for HVG is also plotted for comparison. As expected, motif profiles are different for different marginals. Motifs which are symmetric respect to reflection (3 and 4, 5 and 6) occur with equal probabilities, differently from the case of chaotic time series such as the logistic map (Eq. 4.19) (indeed an i.i.d time series and the same series reversed respect to time are both random and share the same marginal distribution).

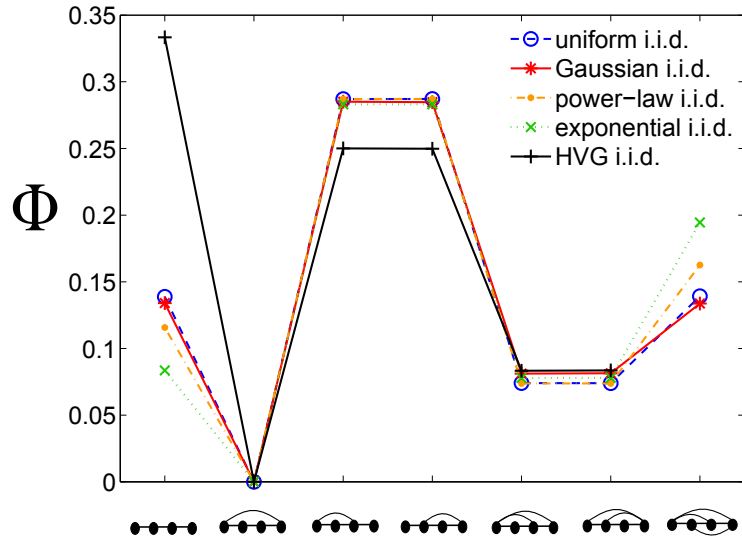


Figure 4.9: (Average frequency of appearance Φ of VG and HVG motifs extracted from i.i.d. series of $N = 10^5$ with different marginal distributions; VG motifs are not related with the ranking statistics of the data and able to discriminate the different types of noise.

According to the values obtained for the components of Z^4 in the preceding cases, one can extract some preliminary heuristic conclusions:

- Φ_1, Φ_7 seems to encode information on the marginal distribution of the process as well as its autocorrelation structure.
- Φ_2 is null as this motif is not a VG. This is at odds with the HVG case, where this is an admissible motif provided the probability of finding consecutive equal data

in the series is finite (e.g. for discrete-valued series).

- The motifs associated to the pairs (Φ_3, Φ_4) (Φ_5, Φ_6) have an obvious mirror symmetry: the motifs associated to Φ_3 and Φ_4 are isomorphic, the correct permutation being $1 - 2 - 3 - 4 \rightarrow 4 - 3 - 2 - 1$ (the same holds for Φ_5 and Φ_6). As the node labelling encapsulates time ordering, for any process which is statistically time reversible [116], we expect these probabilities to be equal. Reversible processes include linear stochastic processes (and both white and red noise belong to this family), while non-invertible chaotic processes are usually time irreversible (the fully chaotic logistic map is an example). Time irreversibility of the process is therefore encoded in these terms.
- $\Phi_8 = 0$ as this is not a VG and therefore does not appear (not admissible).
- As we can notice no set of relation exists for the VG motif 2 and 8, meaning that the two motifs are not allowed by natural visibility. However we have chosen to present the motif profile as an 8-dimensional vector. This choice has been done for the sake of the theoretical discussion, considering that it is not trivial to say *a priori* if a certain motif is admissible by the natural visibility criterion. Also we have included the null components of the VG motif profile for allowing a better comparison with the HVG motif profile.

4.5 Time series classification via visibility graphs motifs

4.5.1 Robustness

In the preceding Section we have developed a general theory to compute explicitly the motif profile of HVGs associated to a given type of dynamics. We have applied this theory to find theoretical expressions in the case of white and coloured noise as well as chaotic dynamics, and have shown that these predictions perfectly match the results found in numerical simulations for reasonably short time series. The theory (which is exact in the limit of infinite size series) is thus correct also in the case of short time series. These are nonetheless only idealised models: empirical time series, however, even if they comply to a particular dynamical system are usually polluted with measurement noise. Therefore, before being able to apply this new technique to real world phenomena, we need to assess its robustness and reliability against noise contamination. To do that, we consider a situation where a chaotic time series is contaminated with different amounts of white noise, and explore the ability of \mathbf{Z}^4 to detect the chaotic signal. Formally, we

pollute a chaotic signal $x(t)$ with uniform white noise $\xi(a)$ and thus construct a noisy chaotic signal $Y(t)$ such that

$$\begin{cases} Y_t = x_t + \xi \\ x_t = 4x_{t-1}(1 - x_{t-1}) \\ \xi \sim U[0, a], \quad 0 \leq a \leq 1, \end{cases} \quad (4.31)$$

where a tunes the noise power. The noise-to-signal ratio of the signal Y_t is defined as $NSR = \sigma_\xi^2 / \sigma_Y^2$ (where σ^2 denotes the variance of signal \cdot), thus NSR will increase monotonically with a . For $NSR \ll 1$, the noise contamination is small. Any technique that is able to distinguish $Y(t)$ and $\xi(t)$ for increasing values of NSR is said to be robust to noise. For $NSR = 1$ the levels of the signal and the noise contamination are comparable and for $NSR > 1$ the underlying chaotic signal is effectively hidden. Of course, when a reaches a certain value it won't be possible any more to distinguish the underlying chaotic nature of the time series by looking at the motif profile. To estimate this threshold we can use two different tests:

- The first test makes use of the (L_1) distance in motif space between the signal and the noise $d(a) = |\mathbf{Z}^4(Y) - \mathbf{Z}^4(\text{iid})|$. This is just a simple, motif-based similarity metric between two graphs, that we use here to measure the similarity between two series. Ideally, the threshold of distinguishability is the smallest value of a for which $d(a) = 0$. However, in practice, as we are dealing with finite size series, there will always be a small uncertainty associated to small finite-size deviations from the theory. That is, if one estimates the $\mathbf{Z}^4(\text{iid})$ with an ensemble average of m realizations of a finite random time series of N data, then for each element in the profile, the standard deviation of the estimate \mathbb{P}_i^4 will be a finite value (that converges to zero as N and m increases). We define $\sigma(\mathbf{Z}^4(\text{iid}))$ as the vector where the i -th term is such standard deviation, for the same values of N and m used in the analysis of $Y(t)$. Then, we define the *uncertainty threshold* a^* as the smallest value of a such that $d(a) \leq |\sigma(\mathbf{Z}^4(\text{iid}))|$ (intuitively, a^* is the smallest value for which we don't know if the difference in the motif profile between the empirical results and the theory are due to the fact that there is a chaotic signal underlying the process, or just due to finite size effects).
- The second possibility is to use a Pearson's χ^2 hypothesis test such as Eq. 4.7 with 4 degrees of freedom, where the null hypothesis is that $Y(t)$ (the observed series) is just white noise (no hidden signal). In this latter case, we are not taking into account the deviations associated to finite size effects in the profile of i.i.d., though.

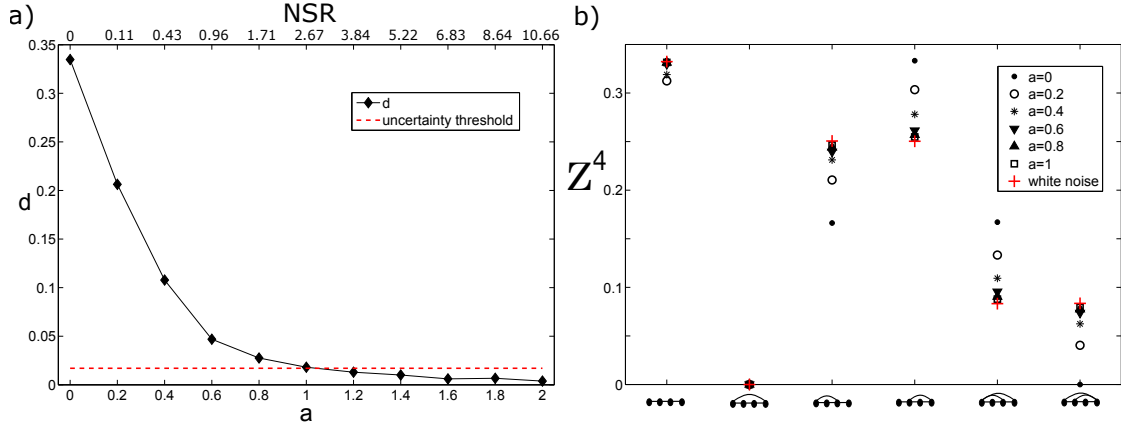


Figure 4.10: Robustness of motif profiles for chaotic series (fully chaotic logistic map) polluted with white noise. Panel a): by increasing the amount of extrinsic noise (parametrised by a) the distance in motif space between the noisy chaotic signal and white noise decreases (see the text). The method is extremely robust as one can distinguish the noisy chaotic signal from pure white noise up to a noise-to-signal ratio $NSR \approx 2.67$. Panel b): the 4-node motif profile of the noisy chaotic signal Y_t for different degrees of noise contamination (a). Motifs III, IV, V and VI are the most informative as they concentrate most of the profile variability.

If $\chi^2 < 9.49$, then we can't reject the null hypothesis at the 95% significance level: this is the limit of what we could call certain distinguishability.

For each value of the parameter a , we have simulated a time series of $N = 10^4$ steps from the process $Y(t)$, and results were ensemble averaged over $m = 100$ realisations. In panel (b) of figure 4.10, we plot the motif profile as a function of a . It is interesting to observe that the probabilities which vary most with a are related to types III, IV, V and VI, while type-I seems to maintain approximately the same rate of appearance (we will show later that this is not always the case). In the panel (a) of the same figure we plot $d(a)$. As expected, $d(a)$ is a monotonically decreasing function of a , and we find $a^* \approx 1$. Remarkably, this corresponds to a value of the noise to signal ratio $NSR \approx 2.67$. This is indeed confirmed by the Pearson χ^2 test, where we found that the limit for confidently rejecting the null hypothesis -certain distinguishability- is $a \approx 1$ (i.e. $NSR \approx 2.67$). These results prove that Z^4 is indeed an extremely robust feature with respect to measurement noise contamination, hence useful for applications.

4.5.2 Principal component analysis

According to the last sections, we can conclude that the HVG Z^4 is an informative feature of complex dynamics. Here we summarise and gather the findings on i.i.d., fully

chaotic logistic maps (with and without noise contamination) and coloured noise, and we complement those with additional chaotic maps (Ricker's map, Cubic map, Sine map). Each process is described by the six dimensional vector \mathbf{Z}^4 (although in practice this space is 5-dimensional as $\mathbb{P}_2^4 = 0$). As this representation is obviously not very convenient for readability, we have projected each point into a 2-dimensional space spanned by the principal components of the data. We recall that Principal Component Analysis (PCA) [119] is a common statistical procedure to perform dimensionality reduction on data. It uses an orthogonal transformation to project our set of observations, originally described in \mathbb{R}^6 -where each direction describes the probability of occurrence of a given motif, this being possibly correlated among observations- into a lower dimensional subspace spanned by the so called principal components, obtained from the eigenvectors of the dataset covariance matrix. These particular directions are such that (i) they are orthogonal, (ii) the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. If the data can be efficiently projected in a lower dimensional space, then the eigenvalues associated to each of the principal components sum up a large percentage of the data variability. In that case, the projection is said to be faithful, and constitutes an accurate description of the data.

To summarise, the following processes have been considered (for all of them, we have estimated \mathbf{Z}^4 from a time series of $N = 10^4$ points, and have averaged this over 100 realisations):

- White noise (i.i.d.) with Gaussian, exponential, uniform and power-law probability densities.
- Chaotic maps, in particular: Fully chaotic logistic map $x_{t+1} = 4x_t(1 - x_t)$, Ricker's map $x_{t+1} = 20x_t e^{-x_t}$, Cubic map $x_{t+1} = 3x_t(1 - x_t^2)$, Sine map $x_{t+1} = \sin(\pi x_t)$
- Noisy logistic map with $a = \{0.2, 0.4, 0.6, 0.8, 1.0\}$
- Coloured noise for $r = \{0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$

The projection into the space spanned by the first two principal components is shown in Figure 4.11. Interestingly, these first two components capture about 98.3% of the variability of the set of variables $\{\mathbf{Z}^4\}$. This means that motif probabilities are indeed highly correlated, and as few as two real numbers per time series seem already enough to describe them. The patterns related to the different processes in this 2-dimensional

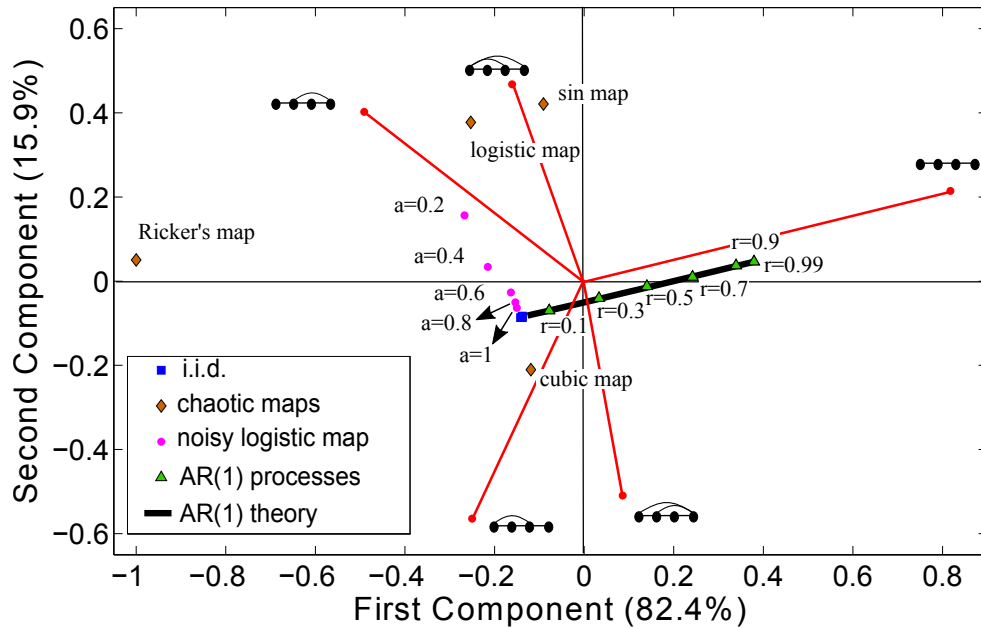


Figure 4.11: 2-dimensional projection obtained via Principal Component Analysis on \mathbf{Z}^4 for time series generated from different deterministic and stochastic processes: different white noise series respectively with Gaussian, exponential, uniform and power law (blue squares), chaotic maps (brown diamonds), noisy logistic map for different levels of contamination (purple dots) and different stochastic correlated AR(1) processes (green triangles). The relative weight of each motif in this projection principal components is also plotted using red solid axes.

component space help visualize some of the results previously found and make interesting considerations:

- All the i.i.d. processes have the same coordinates in the 2-dimensional space which do not correspond to the coordinates of any other class of processes considered. Indeed according to the theory, i.i.d. processes share the same \mathbf{Z}^4 .
- Red solid axes describe the projection of each motif in this new basis (see also table 4-E) and give an idea of which motif types are more related to different processes, thus helping to interpret a particular trajectory in this space, as a given process changes. For instance coloured noise which interpolates between white noise ($r \rightarrow 0$) and a constant series ($r \rightarrow 1$) projects into a straight line-like trajectory, departing from the i.i.d. coordinates and following the direction where type-I motif increases as r increases. Analogously, as the noise level a increases the noisy logistic map interpolates between the fully chaotic logistic coordinate and i.i.d. following a specific path.
- The distance in this space between i.i.d. and the ($a = 1$)-noisy logistic map gives







	First Component	Second Component
	0.814	0.2114
	$-8.6e^{-18}$	$1.5e^{-16}$
	-0.2506	-0.5670
	-0.4926	0.4030
	-0.1569	0.4608
	0.0852	-0.5090

Table 4-E: Weights of each motif in the 2-dimensional projection of the set of all dynamical processes analysed (i.i.d. white noise, coloured noise with exponentially decaying correlations, chaotic maps, noisy chaotic logistic map).

us a rough idea of the distinguishability or coarse-graining distance, a lower bound below which any two processes cannot be distinguished.

We conclude that \mathbf{Z}^4 is a highly informative and robust feature, which in principle could be used to assess similarities and differences across empirical complex signals. To test this hypothesis, in the final section we will explore this idea and will show that clustering of complex physiological processes is possible with this simple feature.

4.6 Unsupervised learning: disentangling meditative from other relaxation states using HVG motif profiles from heart rate time series

It is well-known that meditation has a measurable effect on well-being. In particular, neuroscience has shown that meditation promotes EEG high-amplitude gamma synchronisation [120], or increases sustained attention [121] among others effects on the brain [122]. In this final section we explore, via a HVG motif profile analysis, if one can distinguish purely meditative states from general states of relaxation by only looking at a single physiological indicator: the heart rate series [123, 124]. This analysis is based on experiments performed in a former publication [125]. Data are freely available online [126].

4.6.1 Data

Data are collected for five different groups of healthy subjects [125]:

- The first group of 4 subjects (two women and two men in the age range 20-52) were *expert* Kundalini Yoga meditators. Their heart rate was recorded for approx-

imately fifteen minutes before the Yoga practice (pre-meditative state) and for approximately one hour during the breathing and chanting exercises (meditative state) (a total of 8 time series);

- The second group comprised 8 Chinese Chi Meditation practitioners, (five women and three men in the age range 26-35) *relatively novice* in the practice. The heart rate of the subjects was recorded for approximately five hours during the pre-meditation (pre-meditative state) and for approximately one hour during the meditation session (meditative state)(a total of 16 time series).

To better compare the pre-meditation and meditation states, three healthy, non-meditating control groups were considered from a database of retrospective electrocardiogram (ECG) signals:

- a spontaneous breathing group of 13 subjects (eight women and five men in the age range 25-35) during sleeping hours (general relaxation state) (a total of 13 time series);
- a group of 9 elite triathlon athletes (six women three men, age range 21-55) in the pre-race period during sleeping hours (general relaxation state) (a total of 9 time series);
- a group of 14 subjects (nine women and five men, age range 20-35) during supine metronomic breathing at 0.25 Hz (a total of 14 time series);

Sample time series from each group are plotted in Figure 4.12. In the original study the authors addressed the frequency spectra and observed prominent heart rate oscillations in the time series recorded during the two meditation practices with a peak in the range 0.025-0.35 Hz, and an overall variability of these series with respect to those from non-meditative states.

4.6.2 Unsupervised clustering based on HVG motif profiles

The total dataset is made of a total of 60 time series (60 observations). A priori, we assume that each series is a different process. For each subject and state, we extract from the heart beat series the corresponding \mathbf{Z}^4 (detailed results are put in an appendix).

As a first analysis, we only consider the *expert* meditators (first group) performing two different tasks and we explore if \mathbf{Z}^4 can disentangle the two tasks. Results are shown in panel (a) of figure 4.13. In PCA space, we have 8 points scattered over the subspace

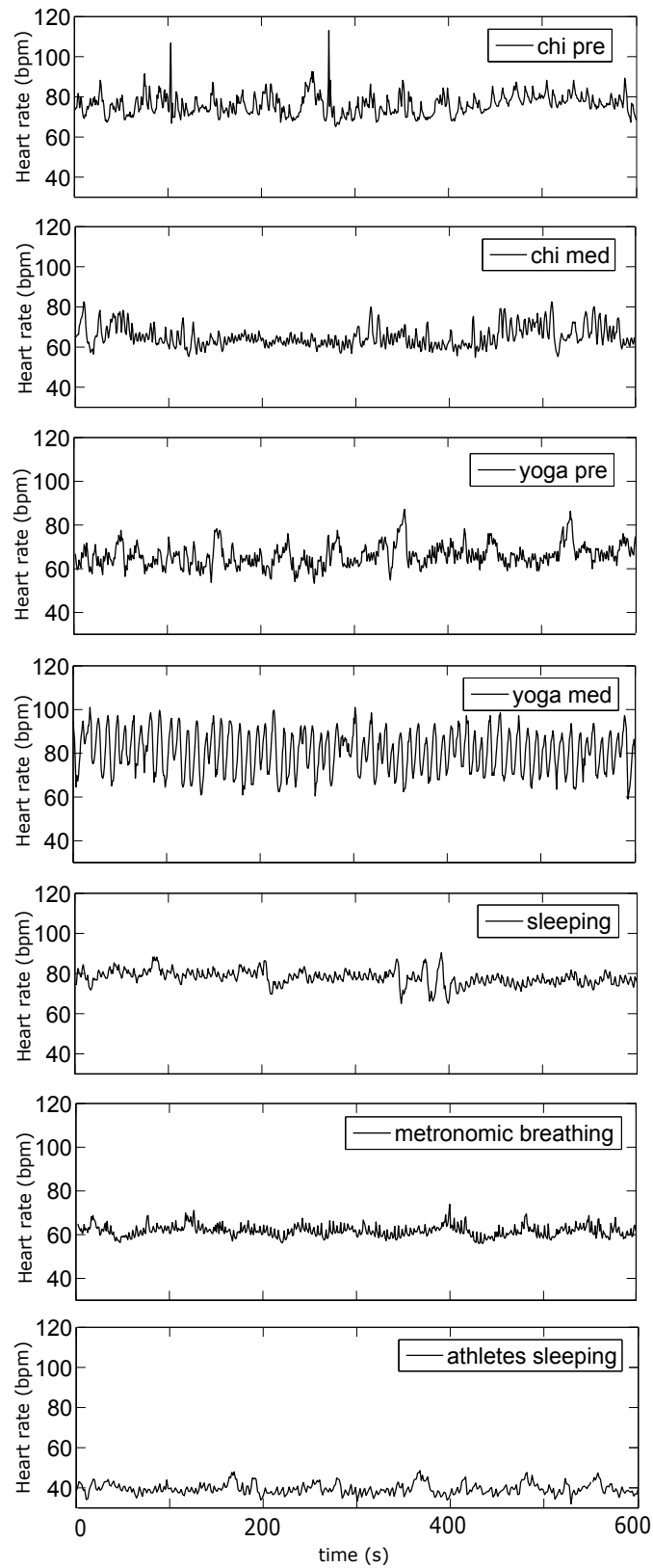


Figure 4.12: Sample heart rate time series from patients in meditative and non-meditative states.

spanned by the first two principal components. These aggregate more than 99% of the data variance and is thus a faithful projection. Interestingly, already a visual inspection clusters the 4 subjects in the meditative state (red circles, right hand side of the plane) from those in the pre-meditative state (green squares). A simple k -means algorithm [119] with $k = 2$ correctly distinguishes the two states by assigning different clusters to both states (a black dotted oval is depicted with the purpose of visualizing the result of the k -means clustering).

In a second step, we consider the second group, formed now by *novice* Chi meditators before and during the practice. We repeat the analysis in the panel (b) of Figure 4.13. Again the first two principal components capture more than 99% of the variability of the motifs considered. The scores related to the first principal component are very close to the ones found for the Yoga data subset (see appendix). For this meditation technique however it is not that easy to perfectly distinguish pre-meditative from meditative state clusters: the partition obtained with the k -means algorithm with input $k = 2$ (visualized by the black dotted line) contain ‘false meditators’ and ‘false non-meditators’. In order to quantify the performance of the clustering we use the so called *purity coefficient* [127] defined by:

$$\text{purity} = \frac{n_m^1 + n_n^0}{n_m^1 + n_m^0 + n_m^0 + n_n^1} \quad (4.32)$$

where n_m^1 is the number of meditators in the cluster 1, which is defined as the cluster where most of the meditators are found; n_m^0 is the number of meditators in the cluster 0, which is defined as the cluster where most of the non-meditators are found (‘false non-meditators’); n_n^1 is the number of non-meditators in the cluster 1 (‘false meditators’); n_n^0 is the number of non-meditators in the cluster 0. Purity takes value in $[0, 1]$ and was measured for the different partitions reported in Figure 4.13 (see Table 4-F); in this case we found $\text{purity} \simeq 0.83$. Now, as in this experiment the subjects were inexperienced Chi meditators, it is plausible that some of them were not able to concentrate or perform the task adequately, what would put their motif profile mixed amongst the pre-meditative state subjects. As we can see in the figure, there is some evidence of finding the ‘false non-meditators’ intertwined among non-meditators, but not the ‘false meditators’ intertwined among meditators.

We then perform the same analysis by considering data from the first two groups (Yoga group and the Chi group) altogether. Here we also aim at distinguishing meditative from pre-meditative states, however this is in principle much more delicate and problematic as we have different subjects performing different tasks. The results are reported in panel (c)

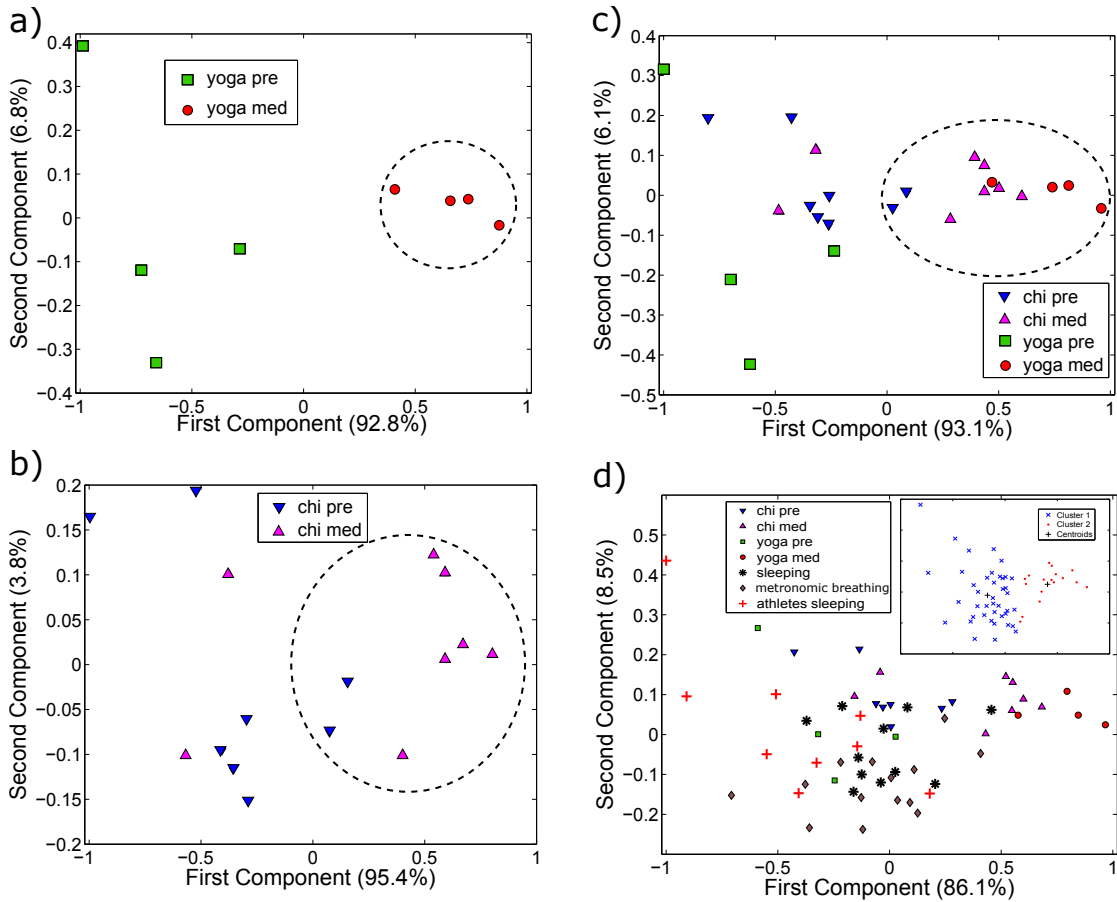


Figure 4.13: 2-dimensional Principal Component space of \mathbf{Z}^4 extracted from heart rate time series of subjects performing different tasks. (a) 4 Yoga meditators recorded during meditation (red dots) and during pre-meditation (green squares). The k -means algorithm (black dotted line) correctly assigns each of the 8 observations into the correct cluster. (b) 8 Chi meditators recorded during meditation (magenta triangles) and during pre-meditation (blue reverse triangles). The k -means algorithm correctly clusters 12 out of 16 observations, however in this case subjects were novice meditators, hence clusters are not that well defined (see the text). (c) The two clusters found by the k -means algorithm correctly clusters the points related to Yoga meditation and Yoga pre-meditation, and 12 out of 16 points related to different Chi meditators (black dotted line). (d): although the k -means clustering (small panel on the top right) fails to precisely distinguish a cluster related to meditation from a cluster related to non-meditation, all the meditation points are surprisingly well separated from the all remaining points, on the right side of the plane.

of Figure 4.13, and are consistent with the first two analysis conducted before. In PCA space, the first two principal components still capture more than 99% of the data variability (scores are reported in the appendix). k -means correctly clusters together most of the pre-meditative states and distinguishes them from the meditative states (Yoga and

Panel a	Panel b	Panel c	Panel d
1	0.83	0.75	0.81

Table 4-F: Purity measures [127] of the k-means clustering analysis depicted in the four panels of Figure 4.13: yoga meditators (panel a), chi meditators (panel b), yoga and chi meditators (panel c) and all states (panel d).

Chi-style), with purity= 0.75. There are two clear ‘false non-meditators’ which seem to correspond to two novice Chi meditators that falsely fall in the non-meditation state despite they were supposedly performing meditation. The two ‘false meditators’ are not mixed among the meditators but placed in the boundary of the cluster, meaning that a refined clustering algorithm would very likely do a better job. On the other hand, it is worth highlighting that meditators show lower scattering than non-meditators, and are placed at the right hand side of the plane. Among these, Chi meditators (the experienced subjects) appear even more towards the right hand side in the PCA plane. According to the motif scores (appendix), one can conclude that meditation promotes the onset of type-I motifs, that is to say, generates a relative decrease of high-frequency heart rate fluctuations.

Finally, in panel (d) of Figure 4.13 we show the results for the analysis of the whole data set (the projection in PCA space still gathers more than 94% of the data variability). Here we have highly heterogeneous subjects performing totally different tasks, which somehow can be classified into ‘meditative’ and ‘non-meditative’ states. In the inset panel of the same figure, each observation is labelled according to the result of k-means (crosses for non-meditative and dots for meditative states). Despite the heterogeneity of subjects, the purity of the partition obtained is high ($\simeq 0.81$), and most of the observations associated to the meditative state concentrate towards the right hand side of the PCA plane (which, again according to the scores, corresponds to an over-contribution of type-I motif). We conclude that meditative practices leave a unique physiological fingerprint in the heart rate time series of its practitioners, which can be distinguished from other relaxation techniques and states such as metronomic breathing or sleeping by using the HVG motif profile of each time series. This is a remarkable result, taking into account that this profile only consists of a vector of 6 numbers (actually 5 as $\mathbb{P}_2^4 = 0$) per observation.

4.7 Comparison of the performance in classification tasks between HVG and VG motifs

When dealing with empirical time series, the practitioner usually faces two different but complementary challenges, namely (i) the size of the series and (ii) the possible sources of measurement noise. The first challenge can be a problem when the statistics to be extracted from the series are strongly affected by finite-size effects, whereas for the second one needs to evaluate the robustness of those statistics against noise contamination. For a statistic or feature extracted from a time series to be not just informative but useful one usually requires that statistic or feature to be robust against both problems: it needs to have fast finite-size convergence speed and to be robust against reasonably large amounts of additive noise.

In [43] it has been already shown that the HVG motif profile has good convergence properties respect to the series size N and it is also robust respect to noise contamination. Here we explore these very same problems for the case of the VG motif profile and we make a detailed comparison of its performance with the HVG motif profile in a range of situations.

4.7.1 Convergence properties for finite size series

In general, due to finite size effects, the estimated value of any feature deviates from its asymptotic value. For classical features such as the mean or the variance of a distribution, these deviations are bounded and vanish with series size with a speed quantified by the central limit theorem. The estimation of the motif frequencies can be quantitatively affected by finite-size fluctuations and one can even observe missing motifs (motifs with estimated frequency $\Phi = 0$) which are not actually forbidden by the process but have not appeared by chance. This situation can be overemphasized in the presence of certain types of measurement noise.

Following an approach analogous to the one followed for the forbidden ordinal patterns in [128–130], we first perform a test to study the decay of missing motifs with the series size both in stochastic uncorrelated and correlated processes. In Figure 4.14 panel a) we plot $\langle R(N) \rangle$, the average number of missing motifs in a series of size N in the case of Gaussian white noise and coloured (red) Gaussian noise (for the red noise we consider the AR(1) process with $r = 0.5$ discussed in section III). For both types of noise $\langle R(N) \rangle$

decays exponentially to zero and already with a series of about 80-100 data points we can exclude the possibility of detecting missing motifs (for both HVG and VG) due to finite size fluctuations even in the case of correlated noise.

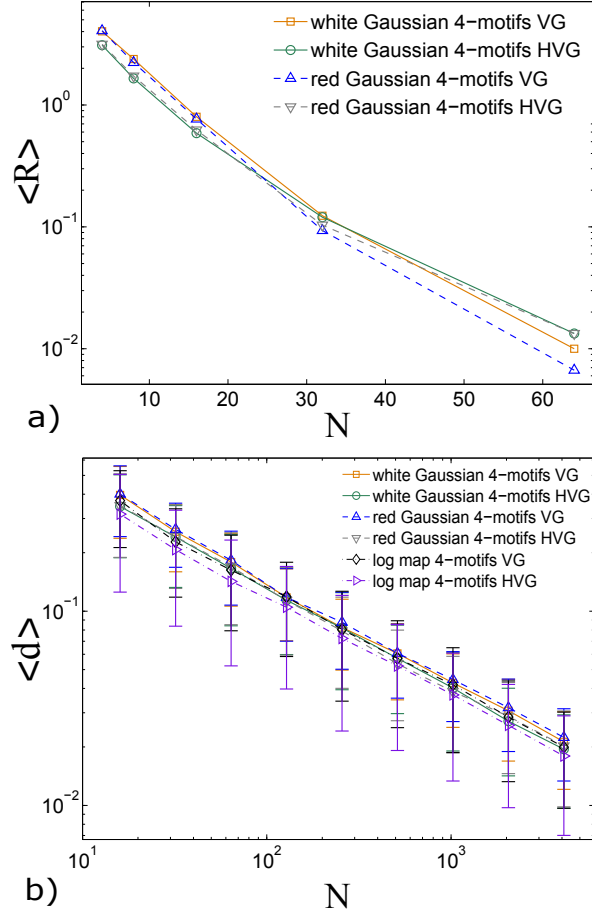


Figure 4.14: Robustness of VG and HVG motif statistic respect to finite series size effects, in the case of Gaussian white noise and coloured (red) Gaussian noise. Panel (a): Semi-log plot of the average number of missing motifs $\langle R \rangle$ vs series size N (each point is an average over 300 realisations of the corresponding process). $\langle R \rangle$ decays exponentially to zero, meaning that for $N \sim 100$ we can already exclude the possibility of detecting missing motifs due to finite size fluctuations for both types of noise. Panel (b): Log-log plot of the average distance $\langle d \rangle$ (see the text) between the observed motif profile and the theoretical profile as a function of the series size N (results are averaged over 300 realisations). $\langle d \rangle$ decreases as a power-law for all the processes considered.

As a second analysis, we explore the convergence speed of the estimated motif profile of uncorrelated and correlated stochastic series and of chaotic series (fully chaotic logistic map) of size N to the asymptotic profile solution given in section III. To do this we define the distance between the estimated 4-motif probabilities $\hat{\Phi}_m^4(N)$ and the asymptotic

value $\Phi_m^4 = \lim_{N \rightarrow \infty} \hat{\Phi}_m^4(N)$. We use ℓ_1 norm and accordingly define

$$d(N) = \sum_m |\hat{\Phi}_m^4(N) - \Phi_m^4| \quad (4.33)$$

In Figure 4.14 panel b) we show the trend of $d(N)$ in log-log scale (results are averaged over 300 realizations). The average distance appears to decrease like a power-law for all the processes considered, in agreement with a central-limit-theorem-like argument: since d quantifies the deviation between the estimated values (from a sample series of finite size N) and the exact (asymptotic) values of the motif frequencies, when the size of the sample series N increases, the deviation of the estimated mean values reduces, and d must converge to zero as $1/\sqrt{N}$ (deviations vanish as a power law of the sample size). For a series of $N = 10^3$ points $\langle d(N) \rangle$ is less than $5 \cdot 10^{-2}$ and the average distance $\langle d_m \rangle$ for each of the single components is less than 10^{-2} (not shown). These results suggest that VG and HVG motif profiles have very good convergence properties and are thus robust against finite size fluctuations.

4.7.2 Robustness against measurement noise

To test and compare the robustness of VG and HVG motif profiles when the effect of noise contamination combines with the finite size fluctuations we consider the fully chaotic logistic map dynamics x_t polluted with measurement (additive) noise η_t

$$\begin{cases} X_t = x_t + \eta_t \\ x_t = 4x_{t-1}(1 - x_{t-1}) \\ \eta_t = r\eta_{t-1} + \sqrt{\alpha}\xi_t, \quad \xi_t \in \mathcal{N}(0, 1) \end{cases} \quad (4.34)$$

in the two cases where η_t is respectively white Gaussian noise ($r = 0$) or colored Gaussian noise ($r = 0.5$). For both cases $\alpha \in [0, 1]$ is the parameter which tunes the noise-to-signal ratio (NSR) of the process defined as

$$\text{NSR}(\alpha) = \frac{\sigma^2[\sqrt{\alpha}\xi]}{\sigma^2[x]} \quad (4.35)$$

where $\sigma^2[\sqrt{\alpha}\xi]$ and $\sigma^2[x]$ are respectively the theoretical variance of the white Gaussian noise $\sqrt{\alpha}\xi$ and the theoretical variance of the dynamics (signal) x (note that with this definition we are underestimating the NSR in the case of correlated noise where $\sigma^2[\eta] = \sigma^2[\sqrt{\alpha}\xi]/(1 - r^2)$). The robustness of the observed motif profile $\hat{\Phi}_i^4[X(N, \alpha)]$ for a single realization of the process with given N and α can be defined as the distance between

this profile and the theoretical profile $\Phi_i^4[\eta(\alpha)]$ of the noise η for the given α .

$$\delta(N, \alpha) = \sum_m |\hat{\Phi}_m^4[X(N, \alpha)] - \Phi_m^4[\eta(\alpha)]|. \quad (4.36)$$

With such definition we expect $\delta(N, \alpha) \gg 0$ for low values of the NSR (dominant signal, $\hat{\Phi}_m^4[X(N, \alpha)] \simeq \Phi_m^4[x]$) and $\delta(N, \alpha) \simeq 0$ for high values of the NSR (dominant noise, $\hat{\Phi}_m^4[X(N, \alpha)] \simeq \Phi_m^4[\eta]$). Furthermore, $\delta(N, \alpha)$ is affected by finite size effects: if we assume to have few realizations N_r of the process X of small series size N , then we expect the variance $\sigma^2(\delta(N, \alpha))$ calculated over the realizations to be high. In particular we have to consider that a resolution limit δ^0 exists, such that when $\langle \delta(N, \alpha) \rangle_{N_r} = \delta^0$ we cannot say any more if the distance we measured is discriminating the signal x from the noise η or it is simply due to finite-size effects of the contamination noise η . We define this threshold δ^0 as the sum of the standard deviations of the estimated profile components $\hat{\Phi}_m^4[\eta]$ given N_r realizations of the noise process alone

$$\delta^0(N, \alpha) = \sum_m \sqrt{\langle (\hat{\Phi}_m^4[\eta(N, \alpha)] - \langle \hat{\Phi}_m^4[\eta(N, \alpha)] \rangle_{N_r})^2 \rangle_{N_r}} \quad (4.37)$$

It is thus convenient to work with the relative distance $\langle \delta \rangle_{N_r} / \delta^0$; when $\langle \delta \rangle_{N_r} / \delta^0 \leq 1$ we say that the resolution limit for the process $X(N, \alpha)$ -given its N_r realizations- is reached, and $\langle \delta \rangle_{N_r}$ is not any more a reliable indicator.

In Figure 4.15 panel a) we show the quantity $\langle \delta \rangle / \delta^0$ averaged over 300 realizations of the process X for different levels of contamination $\text{NSR} = [0, 0.2, 0.4, \dots, 8]$ at fixed size $N = 6400 = 100 \cdot 2^6$ (notice that for $N \geq 100$ the missing motifs are not found any more) respectively for white Gaussian noise and coloured Gaussian noise and for the HVG and the VG motif profile. The red solid line plotted in the figure represents the resolution limit threshold for the process. We can see that the HVG and the VG motif profiles are more robust respect to noise contamination when this noise is correlated. In this situation the HVG motif profile seems to perform better than the VG motif profile, while in the case of uncorrelated Gaussian noise the VG profile seems in turn slightly more robust than the HVG profile.

The last step of this robustness analysis is to consider the usual situation where only very few realizations (often a single one) of the same process are available. Our aim is to define a useful indicator θ which estimates for any given value of the size N the maximum amount of noise contamination level for which a measure δ computed with only one realization of the process X can be considered reliable. We define this to be

the value of the NSR such that $\langle \delta \rangle - \sigma(\delta) = \delta^0$, and thus

$$\theta(N) = \{\text{NSR}(\alpha) : \langle \delta(N, \alpha) \rangle - \sigma(\delta(N, \alpha)) = \delta^0\}. \quad (4.38)$$

$\theta(N)$ measures (in units of noise-to-signal ratio) the (statistical) reliability of the motif profile extracted from a single time series of size N of the signal x in the presence of measurement noise η .

In Figure 4.15 (panel a) we plot $\theta(N = 6400)$ for white Gaussian noise in the case of VG by considering the $\langle \delta \rangle / \delta^0$ curve marked by orange squares and by taking the smallest value of NSR for which the statistical error range allows to measure values of δ / δ^0 equal or smaller than the value of the resolution limit threshold (this value of NSR is highlighted in figure using a blue box). We find $\theta \simeq 2.2$, meaning that when working with a single time series of the process X with size $N = 6400$, the $\langle \delta \rangle$ distance measured by using the VG motif profile is reliable up to a level of white Gaussian noise contamination α such that $\text{NSR}(\alpha) \simeq 2.2$. In Figure 4.15 (panel b) we report the estimated value of θ for the VG and HVG motif profiles in the case of white Gaussian noise and correlated Gaussian noise in function of the series size $N = 100 \cdot 2, 100 \cdot 2^2, \dots, 100 \cdot 2^7$ (maximum noise contamination level considered was $\text{NSR}(\alpha) = 8$). We can see that the motif profile is in general a robust measure respect to the combined effect of measurement noise and finite size: working with a single time series of only 3000 points of the process X we can extract both the VG and the HVG motif profiles and expect those features to be informative respect to the underlying chaotic signal x up to a level of measurement noise for which $\text{NSR} = 1.5$ in the case of uncorrelated Gaussian noise and $\text{NSR} = 3$ in the case of correlated Gaussian noise.

Also and as observed before (Figure 4.11 panel b)), given the case of white Gaussian noise contamination the VG motif profile (orange squares) seems to perform slightly better than the HVG motif profile (green circles). For colored Gaussian noise the situation is the opposite and the HVG motif profile (reversed gray triangles) performs much better (almost a gap of one unit of NSR for $N > 1600$) than the VG motif profile (blue triangles). For both types of visibility graphs the motif profile is systematically more robust when polluted with colored noise than with white noise. This is probably due to the fact that white noise breaks up the correlation structure of the signal faster (respect to the size N) than correlated noise.

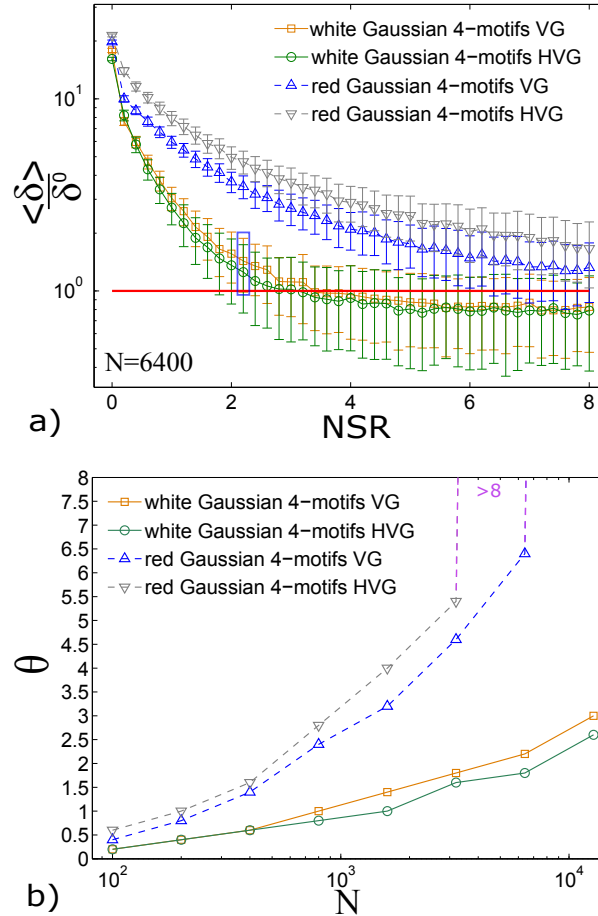


Figure 4.15: Robustness of HVG and VG motif profiles respect to measurement noise. Panel a) the average distance between the motif profile extracted from the polluted chaotic dynamics X (see Eq. 4.34) and the theoretical motif profile of the noise, normalized by the resolution limit threshold δ^0 for different level of contamination NSR (noise-to-signal ratio) at fixed size $N=6400$. When the curves reach the resolution limit (horizontal line) we cannot consider any more the motif profile informative about the underlying chaotic dynamics due to the noise effects. Panel b) the estimated values of the measure θ in function of the time series size N indicating the maximum amount of NSR for which the motif profile extracted from a single realization of the process X is reliably informative respect to the chaotic signal. The HVG and the VG motif profiles are more robust respect to noise contamination when the noise is correlated (red Gaussian). In this situation the HVG motif profile seems to perform better than the VG motif profile, while in the case of uncorrelated Gaussian noise the VG profile seems in turn slightly more robust than the HVG profile.

Chapter 5

Conclusion

Much of the information encoded in a network can be extracted from its mesoscopic structure by looking at the properties of motifs and modules or communities of nodes. Modularity is indeed a universal properties of the mesoscale structure of real world complex networks. To explain the emergence of the community or modular structure in real world evolving networks we have considered the simplest growing network model enforcing the mechanism of triadic closure, for which simple three nodes' motifs, the triangles, are formed by new nodes joining the network during its evolution. We have shown that triadic closure is alone capable to generate systems with all the characteristic properties of complex networks, from fat-tailed degree distributions to high clustering coefficients and strong community structure. Communities emerge naturally via triadic closure, which tends to generate cohesive subgraphs around portions of the system that happen to have higher density of links, due to stochastic fluctuations. When the modules become sufficiently large, their internal structure exhibits in turn link density inhomogeneities, leading to a progressive differentiation and eventual separation into smaller modules (separation in the sense that the density of links between the parts is appreciably lower than within them). This occurs both in the basic version of network growth model based on triadic closure, and in more complex variants. The strength of community structure is the higher, the sparser the network and the higher the probability of triadic closure. We have discussed a new variant of the model, in that link attractivity depends on some intrinsic appeal of the nodes, or fitness and we have seen that, when the distribution of fitness is not too heterogeneous, community structure still emerges, though it is weaker than in the absence of fitness. By increasing the heterogeneity of the fitness distribution, instead, we observe a major change in the structural organization of the network: communities disappear and are replaced by special subgraphs, whose nodes are connected only to superhubs of the network, i.e. nodes attracting most of the

links. Such structural phase transition is associated to very high values of the clustering coefficient.

The evolving network model enforcing triadic closure has been generalized to the case of multiplex network structures, where nodes can be connected on different layers (networks) describing different types or levels of interactions. This generalization is relevant if we consider that multiplex networks are ubiquitous and systems as different as social networks, transportation networks or cellular and brain networks require a multilayer description. In particular, an interesting aspect exhibited by multiplex social collaboration networks is that their mesoscopic structure usually displays cohesive communities spanning more than a single layer of interaction.

This multi-level communities' organization can be revealed by using an indicator function $\tilde{\Theta}^S$, based on the entropy of network ensembles, that is able to measure the mesoscopic similarities between the layers of a multiplex network. This indicator has been used to analyse the APS Collaboration Multiplex Network at the two levels of the PACS hierarchy, providing a bottom-up approach to identify how the organization of knowledge in physics is reflected in the structure of collaboration networks. The same APS Co-authorship Multiplex Network together with the IMDb Co-starring Multiplex Network have been analysed to test the generalised model. Interestingly the simplest version of the proposed model of growing multiplex network, based just on the interplay between intra- and inter-layer triadic closure, is actually able to explain much of the complexity observed in the micro- meso- and macroscopic structure of multidimensional collaboration networks in the different fields of science and movies, including not just the clustering but also intra- and inter-layer degree correlation patterns and the correspondence between the community structures at difference layers. Such levels of accuracy in reproducing the features of real-world systems have been obtained without the introduction of ad-hoc ingredients and the results suggest that, despite the apparent differences in the overall dynamics driving scientific cooperation and movie co-starring, triadic closure is a quite generic mechanism and might indeed be one of the fundamental processes shaping the mesoscale structure of multi-layer collaboration systems.

Finally we have seen that the motif formation also plays a important role in characterising the mesoscopic structure of particular types of network architectures, the visibility graphs, that result from mapping time series associated to a general dynamical processes into a graph. The theory of visibility graphs allows to describe and characterise time series and dynamics using the tools from graph theory and network analysis. The motif formation in the visibility graphs is intrinsically related with the characteristic trajectories of the dynamical process in its phase space. Thus if we look at the mesoscale structure of VGs, we can detect sequentially along the characteristic Hamiltonian path

small subgraphs, that we call sequential visibility graph motifs, which are highly informative about the time series structure and its underlying dynamics.

We have advanced a mathematically sound theory by which the HVG motif profile of large classes of stochastic and deterministic dynamics can be computed exactly. Interestingly, under the HVG framework, graph motifs are in direct correspondence with ordinal patterns [110]. This means, for instance, that the theory developed here can be exported to find rigorous results on the permutation entropy [131] and permutation spectra [112] of different dynamical systems. In the same vein, one could import concepts and ideas from ordinal patterns to the context of visibility graphs. For instance, one can define an HVG motif entropy $S_n = -\frac{1}{n} \sum \mathbf{Z}_i^n \log(\mathbf{Z}_i^n)$ and explore its similarities with the permutation entropy. More generally, the relation (and possible equivalences) between ordinal pattern analysis (so called permutation complexity [110]) and horizontal visibility graph analysis should be studied in more depth.

We have shown that the motif statistic is surprisingly robust, in the sense that it allows to distinguish amongst different dynamics even when the signals are polluted with large amounts of measurement noise, enabling its use in practical problems. As an application, we have tackled the problem of disentangling meditative from general relaxation states from the HVG motif profiles of heartbeat time series of different subjects performing different relaxation tasks. We have been able to provide a positive, unsupervised solution to this question by applying standard clustering algorithms on this simple feature.

We have extended the theory for exactly computing the motif profile to the realm of NVGs, for which the previous amount of known exact results was practically null. We have been able to give a closed form for the 4-node VG motif profile associated to general one dimensional deterministic and stochastic processes with a smooth invariant measure or continuous marginal distribution, for the cases where the variables belong to a bounded or unbounded interval. In the case where the time series is empirical and one does not have access to the underlying dynamics, the methodology still provides a linear time ($O(N)$, as for HVGs) algorithm to estimate numerically such profile. VG motifs have similar robustness properties as HVG, yet they depend on the marginal distribution of the process and as such yield different profiles for different marginals. This is at odds with the results found for HVGs, where the motif profiles did not depend on the marginals as they behave as an order statistic. Thus the deep similarity between HVG motifs and the so called ordinal pattern analysis -which holds mainly due to the fact that HVG is an order statistic- vanishes for VG motifs. This suggests that VG motifs provide different information than HVG ones and therefore stand as a complementary tool for time series description and classification, specially relevant when the marginals play a role in the analysis.

Future perspectives

Complex systems are always fascinating. From the cell, the smallest unit of life, to the brain, to societies, to the entire biosphere. Every scientist would like to have an answer to the question “How do they work?”. Network Science has not provided the answer yet, but thanks to its simple and powerful idea of describing those systems as networks has been the field who has contributed the most during the past two decades in shedding light on their principles and mechanisms, and has done that across all the disciplines of Natural Science.

I feel I have been very fortunate during the course of my Ph.D. in focusing my research on the mesoscopic structure of complex networks, on the emergence and evolution of communities and on network motifs. They all are important universal concepts in the theory of complex networks and their understanding helped me a lot in appreciating the interdisciplinarity, the beauty and the reasons of the success of this theory.

Working on multiplex networks, which is the most recent and promising advance in the field, has been very stimulating. Finding how communities can emerge in this network structures, that are a natural representation for the many complex systems in real world showing co-existence of interactions of different nature, has been very stimulating. However, what the ‘multiplex nature’ of communities is and how can be defined has still not been satisfactory elucidated and I think that future efforts to address this question should go in the direction of characterising the multiplex communities at the level of the links rather than at the level of the nodes, as ‘multi-link communities’.

On the other hand the idea of studying the mesoscale structure of visibility graphs via sequential motifs seems very promising for the analysis of time series and complex signals, because of the computational efficiency of the method, of its tractability and of its deep relation with important concepts in the theory dynamical systems such as ordinal patterns and permutation entropy. In this perspective I think that it would be interesting for future research to understand whether predictive algorithms could be defined using motif-based indicators for forecasting time series behaviour. In this direction it would be fundamental to study the higher order statistics of motif correlations along the Hamiltonian path of visibility graphs.

With these ideas in mind I will continue my investigation of complex networks and of their structures, that, to some extent, encodes the ‘mystery of complexity’ itself.

Appendix A

Methods for the analysis of multiplex social collaboration networks

Data sets. – We considered data from the APS and the IMDb collaboration networks. The APS collaboration data set is available from the APS website

<http://journals.aps.org/datasetsinthe>

in the form of XML files containing detailed information about all the papers published by all APS journals. The download is free of charge and restricted to research purposes, and APS does not grant to the recipients the permission to redistribute the data to third parties. We parsed the original XML files to retrieve, for each paper, the list of authors and the list of PACS codes. The PACS scheme provides a taxonomy of subjects in physics, and is widely used by several journals to identify the sub-field, category and subject of papers. We used the highest level of PACS codes to identify the ten main sub-fields of physics, and we considered only the papers published in Nuclear physics, Particle physics, Condensed Matter I and Interdisciplinary physics, respectively associated to high-level PACS codes starting with 1 (Particle physics), 2 (Nuclear physics), 6 (Condensed Matter I) and 8 (Interdisciplinary physics). We focused only on the authors who had at least one publication in each of the four sub-fields [20]. The co-authorship network of each of those four sub-fields constitutes one of the four layers of the APS multiplex. In particular, two authors are connected on a certain layer only if they have co-authored at least one paper in the corresponding sub-field. In the construction of the collaboration network of each sub-field we purposely left out papers with more than

ten authors, which represent big collaborations whose driving dynamics might be more complex than just triadic closure.

The IMDb data set is made available at the website

ftp : //ftp.fu – berlin.de/pub/misc/movies/database/

for personal use and research purposes. The data set comes in the form of several compressed text files, and we used those containing information about actors, actresses, movies and genres. We focused only on the co-starring networks of four movie genres, namely Action, Crime, Romance, and Thriller [20], obtained by merging information about participation of actors and actresses to each movie. In particular, two actors are connected by a link on a given layer (genre) only if they have co-starred in at least one movie of that genre. We considered only the actors who had acted in at least one movie of each of the four genres. We chose to restrict our analysis to just four layers for both the APS and the IMDb data set, which allowed us to consider the simplest formulation of our model, in which all the layers have the same clustering coefficient C . The use of the APS and the IMDb data sets does not require any ethical approval.

Synthetic multiplex networks. — We created synthetic networks according to our multi-layer network model by starting, on each layer, from a seed graph consisting of a triangle of nodes and simulating the intra- and inter-layer triadic closure mechanism for $N = 20000$ nodes, for different values of the parameters p and p^* . For each pair of values (p, p^*) we computed the mean clustering coefficient C on each single layer and the normalised mutual information NMI of the community partitions of the two layers over 30 different realisations. As observed from simulations, once the parameters (p, p^*) are fixed, the values of NMI and C do not vary substantially as the order N of the network increases. Notice that since the most simple formulation of the model we have set an identical value of p on both layers, the two layers will end up having the same clustering coefficient (up to small finite-size fluctuation).

Degree correlations. — We study the assortativity of real multiplex collaboration networks in terms of intra-layer, inter-layer and mixed degree correlations. The trend for intra-layer correlations is analysed by mean of the function $\langle K_{nn}^{[\alpha]}(k^{[\alpha]}) \rangle$, that is the average degree of the nearest neighbours on layer α of a node with given degree $k^{[\alpha]}$ on that layer. In particular, $\langle K_{nn}^{[\alpha]} \rangle$ is obtained as an average of $K_{nn,i}^{[\alpha]}$ over all nodes with the same degree $k^{[\alpha]}$. The node term can be computed as $K_{nn,i}^{[\alpha]} = \frac{\sum_{j \neq i} a_{ij}^{[\alpha]} k_j^{[\alpha]}}{k_i^{[\alpha]}}$, where $a_{ij}^{[\alpha]}$ are the entries of the adjacency matrix at layer α . Since such measure considers only a layer at a time, the layer index here is not strictly necessary but will be kept for symmetry with the other coefficients. It is interesting to notice that, in absence of intra-layer degree cor-

relations, $\langle K_{nn}^{[\alpha]}(k^{[\alpha]}) \rangle$ is a constant, while $\langle K_{nn}^{[\alpha]}(k^{[\alpha]}) \rangle$ is an increasing (resp., decreasing) function of $k^{[\alpha]}$ if assortative (resp., disassortative) degree correlations are present. To quantify inter-layer degree correlations we considered the quantity $\langle k^{[\beta]}(k^{[\alpha]}) \rangle$ [20, 132], that is the average degree on layer β of a node with degree $k^{[\alpha]}$ on layer α . Again, $\langle k^{[\beta]}(k^{[\alpha]}) \rangle$ will be an increasing function of $k^{[\alpha]}$ if nodes tend to have similar degrees on both layers (assortative inter-layer correlations), while $\langle k^{[\beta]}(k^{[\alpha]}) \rangle$ will decrease with $k^{[\alpha]}$ if a hub on one layer will preferentially have small degree on the other layer, and vice-versa. Finally, we measured the presence of mixed correlations through the function $\langle K_{nn}^{[\beta,\alpha]}(k^{[\alpha]}) \rangle$, that is the average degree on layer β of the nearest neighbours on layer β of a node with degree $k^{[\alpha]}$ on layer α [133]. In analogy with the case of intra-layer correlations, the node term is $K_{nn,i}^{[\beta,\alpha]} = \frac{\sum_{j \neq i} a_{ij}^{[\beta]} k_j^{[\beta]}}{k_i^{[\alpha]}}$. We remark here that there exists another possible definition of mixed correlations coefficient, which considers the nearest neighbours of a node on layer α rather than β (see Ref. [133] for details). The results for the alternative definition of mixed correlations are analogous to those observed for $\langle K_{nn}^{[\beta,\alpha]}(k^{[\alpha]}) \rangle$ and are not shown in the text.

In general, correlation functions might be affected by the degree sequence at each layer of the multiplex. In the simple scenario considered at first, however, we do not fit the parameter m from the data, to reduce as much as possible the complexity of the model. Instead, in order to still perform an accurate comparison between the synthetic multiplex networks constructed by our model and the real ones, in a second step we divided all the correlation functions by their (constant) value expected in the corresponding configuration model network. The correct normalisation for the intra-layer correlation function is $\frac{\langle (k^{[\alpha]})^2 \rangle}{\langle k^{[\alpha]} \rangle}$ [134], while for the inter-layer correlation function we have to divide $\langle k^{[\beta]}(k^{[\alpha]}) \rangle$ by $\langle k^{[\beta]} \rangle$. Finally, the mixed correlation function is correctly normalised by $\frac{\langle (k^{[\beta]})^2 \rangle}{\langle k^{[\alpha]} \rangle}$.

Appendix B

Explicit computation of the HVG motif profile for the fully chaotic logistic map

- \mathbb{P}_1^4

$$\mathbb{P}_1^4 = \int_0^1 f(x_0) dx_0 \int_0^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{x_1} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^{x_2} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3 + \int_0^1 f(x_0) dx_0 \int_{x_0}^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_1}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

the first integral on the right gives the following conditions:

$$\mathcal{H}^3(x_0) < \mathcal{H}^2(x_0)$$

$$\mathcal{H}^2(x_0) < \mathcal{H}(x_0)$$

which are never satisfied. The second integral gives:

$$\mathcal{H}^2(x_0) > \mathcal{H}(x_0)$$

$$\mathcal{H}(x_0) > x_0$$

which are satisfied for $x_0 \in [0, 1/4]$. Thus

$$\mathbb{P}_1^4 = \frac{1}{\pi} B_{[0, \frac{1}{4}]} \left(\frac{1}{2}, \frac{1}{2} \right) = \frac{1}{3} \quad (= 8/24).$$

- $\mathbb{P}_2^4 = 0$ since the probability of having $\mathcal{H}^2(x_0) = \mathcal{H}(x_0)$ is of zero measure.

- \mathbb{P}_3^4

$$\mathbb{P}_3^4 = \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_1}^{x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^{x_2} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3 + \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_0}^1 \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_0^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

In the first term:

$$\mathcal{H}(x_0) < x_0 \Rightarrow x_0 \in [3/4, 1]$$

$$\mathcal{H}^2(x_0) > \mathcal{H}(x_0) \cap \mathcal{H}^2(x_0) < x_0 \cap [3/4, 1] \Rightarrow x_0 \in [\frac{5+\sqrt{5}}{8}, 1]$$

$\mathcal{H}^3(x_0) < \mathcal{H}^2(x_0) \cap [\frac{5+\sqrt{5}}{8}, 1] \Rightarrow x_0 \in [\frac{5+\sqrt{5}}{8}, \frac{1}{2} + \frac{\sqrt{3}}{4}]$ Analogously for the second term,

$$\mathcal{H}(x_0) < x_0 \Rightarrow x_0 \in [3/4, 1]$$

$$\mathcal{H}^2(x_0) > x_0 \cap [3/4, 1] \Rightarrow x_0 \in [3/4, \frac{5+\sqrt{5}}{8}]$$

Altogether,

$$\mathbb{P}_3^4 = \frac{1}{\pi} B_{[3/4, \frac{1}{2} + \frac{\sqrt{3}}{4}]}(1/2, 1/2) = \frac{1}{6} (= 4/24)$$

- \mathbb{P}_4^4

$$\mathbb{P}_4^4 = \int_0^1 f(x_0) dx_0 \int_{x_0}^1 \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{x_1} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{x_2}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3 + \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_0}^{x_1} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{x_2}^{x_1} \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

the first integral on the right gives the following conditions:

$$\mathcal{H}^3(x_0) > \mathcal{H}^2(x_0)$$

$$\mathcal{H}^2(x_0) < \mathcal{H}(x_0)$$

$$\mathcal{H}(x_0) > x_0$$

which are satisfied for $x_0 \in [1/2, 3/4]$. The second integral gives:

$$\mathcal{H}^2(x_0) < \mathcal{H}^3(x_0) < \mathcal{H}(x_0)$$

$$\mathcal{H}^2(x_0) < \mathcal{H}(x_0)$$

$$\mathcal{H}(x_0) < x_0$$

which are satisfied for $x_0 \in [1/2 + \sqrt{3}/4, 1]$. Thus

$$\mathbb{P}_4^4 = \frac{1}{\pi} \left[B_{[1/2, 3/4]} \left(\frac{1}{2}, \frac{1}{2} \right) + B_{[1/2 + \sqrt{3}/4, 1]} \left(\frac{1}{2}, \frac{1}{2} \right) \right] = 8/24.$$

- \mathbb{P}_5^4

$$\mathbb{P}_5^4 = \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_{x_1}^{x_0} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{x_2}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

gives the following conditions:

$$\mathcal{H}^3(x_0) > \mathcal{H}^2(x_0)$$

$$\mathcal{H}(x_0) < \mathcal{H}^2(x_0) < x_0$$

which are satisfied for $x_0 \in [1/4 + \sqrt{3}/4, 1]$ and

$$\mathbb{P}_5^4 = \frac{1}{\pi} B_{\left[\frac{1}{4} + \frac{\sqrt{3}}{4}, 1\right]} \left(\frac{1}{2}, \frac{1}{2} \right) = \frac{1}{6} \quad (= 4/24).$$

• \mathbb{P}_6^4

$$\mathbb{P}_6^4 = \int_0^1 f(x_0) dx_0 \int_0^{x_0} \delta(x_1 - \mathcal{H}(x_0)) dx_1 \int_0^{x_1} \delta(x_2 - \mathcal{H}^2(x_0)) dx_2 \int_{x_1}^1 \delta(x_3 - \mathcal{H}^3(x_0)) dx_3$$

gives the following conditions:

$$\mathcal{H}^3(x_0) > \mathcal{H}(x_0) > \mathcal{H}^2(x_0)$$

$$\mathcal{H}(x_0) < x_0$$

which are never satisfied for the $\mathcal{H}(x)$ map (this is indeed based on the fact that the pattern $x_i > x_{i+1} < x_{i+2}$ is indeed a forbidden pattern in the orbit of $\mathcal{H}(x)$). Hence

$$\mathbb{P}_6^4 = 0.$$

$$\begin{aligned} \mathbb{P}_5^4 &= \int_{1/2 + \sqrt{3}/4}^1 \frac{1}{\pi \sqrt{x_0(1-x_0)}} dx_0 = \\ &= \frac{1}{\pi} B_{\left[\frac{1}{2} + \frac{\sqrt{3}}{4}, 1\right]} \left(\frac{1}{2}, \frac{1}{2} \right) = \frac{1}{6} (= 4/24) \end{aligned}$$

Appendix C

Motif profiles for all subjects in the empirical study

In Figure C.1 we give an overview of the HVG 4-node motif profiles, measured for the different subjects in the different states. Interestingly, the motif that shows more variability in each of the given states is the one related to the type-1 motif, which we have seen to play a minor role in the case of the chaotic dynamics polluted with noise.

The scores of the two components in terms of motifs are reported in Tables ??, and as expected the highest contribution to the first component (0.874) is given by motif of type 1.







	Yoga		Chi	
	First Component	Second Component	First Component	Second Component
	0.874	0.0346	0.871	0.171
	-0.029	-0.203	-0.023	-0.239
	-0.379	0.074	-0.376	0.207
	-0.204	0.731	-0.272	0.666
	-0.043	0.01	-0.048	-0.173
	-0.219	-0.647	-0.152	-0.631

Table C.1: Principal component scores obtained from PCA considering the Yoga meditators data subset (left) and the Chi meditators data subset (right).

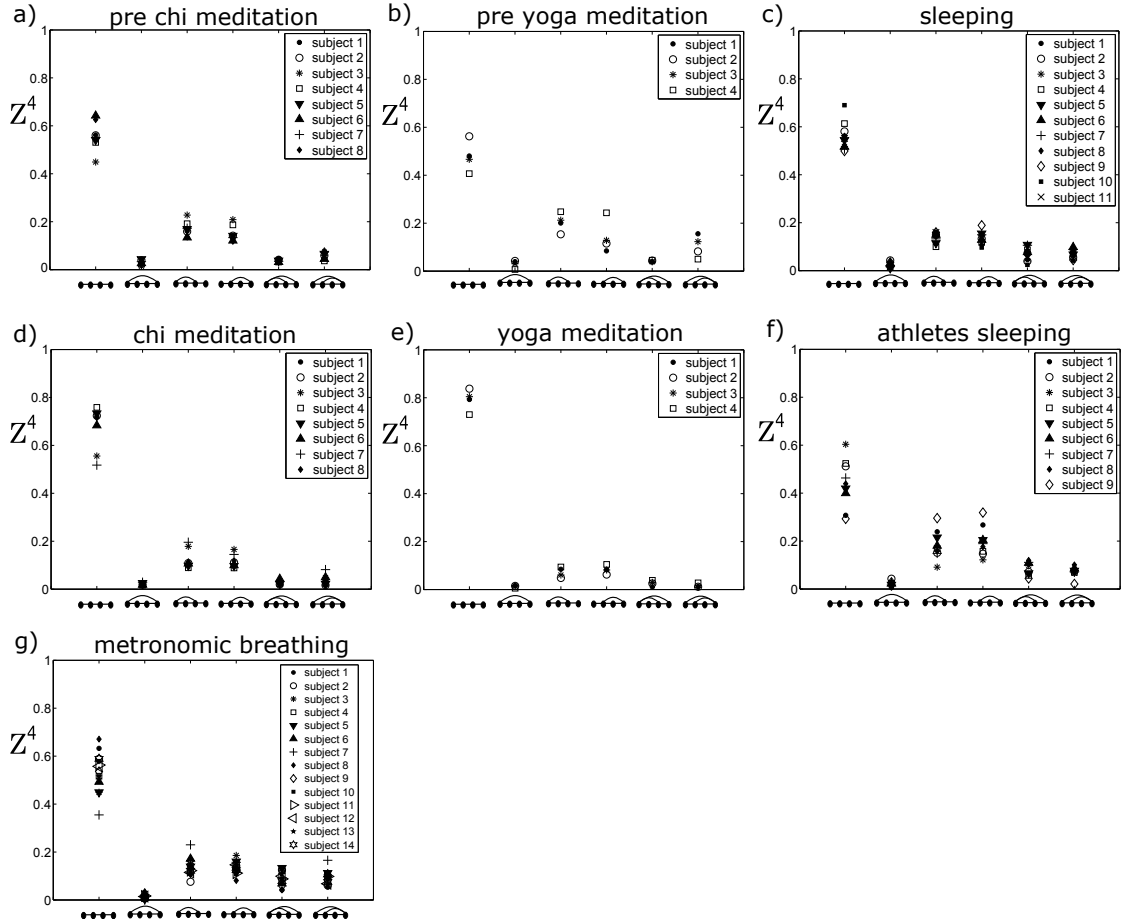


Figure C.1: HVG motif significance profile Z^4 obtained by analysing heart rate time series from different groups of subjects in different states: a) 8 Chi meditators before the meditation practice; b) 4 Yoga meditators before the meditation practice; c) 11 subjects during sleeping; d) same 8 Chi meditators of a) during the meditation practice; e) same 4 Yoga meditators of b) during the meditation practice; f) 9 elite athletes during sleeping; g) 14 subjects during metronomic breathing at 0.25 Hz.

	Chi&Yoga		All States	
	First Component	Second Component	First Component	Second Component
	0.874	0.08	0.881	0.133
	-0.027	-0.181	0.007	0.073
	-0.378	0.089	-0.315	0.477
	-0.235	0.714	-0.294	0.4
	-0.045	-0.039	-0.13	-0.58
	-0.188	-0.664	-0.15	-0.503

Table C.2: Principal component scores obtained from PCA considering the subset data of Yoga and Chi meditators together (left) and considering all data set (right).

References

- [1] Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics* 74:47–97.
- [2] Newman MEJ (2003) The Structure and Function of Complex Networks. *SIAM Review* 45:167–256.
- [3] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Physics reports* 424:175–308.
- [4] Milo R, et al. (2002) Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298:824–827.
- [5] Alon U (2007) Network motifs: theory and experimental approaches. *Nature Reviews Genetics* 8:450–461.
- [6] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297:1551–1555.
- [7] Kreimer A, Borenstein E, Gophna U, Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences* 105:6976–6981.
- [8] Rives AW, Galitski T (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences* 100:1128–1133.
- [9] Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. *Nature Reviews Genetics* 8:921–931.
- [10] Meunier D, Lambiotte R, Bullmore ET (2010) Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience* 4:200.
- [11] Olesen JM, Bascompte J, Dupont YL, Jordano P (2007) The modularity of pollination networks. *Proceedings of the National Academy of Sciences* 104:19891–19896.
- [12] Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. *Physical review E* 68:065103.
- [13] Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Science* 99:7821–7826.
- [14] Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. *Current opinion in microbiology* 6:125–134.
- [15] Sekara V, Stopczynski A, Lehmann S (2016) Fundamental structures of dynamic

- social networks. *Proceedings of the national academy of sciences* 113:9977–9982.
- [16] Boccaletti S, et al. (2014) The structure and dynamics of multilayer networks. *Physics reports* 544:1–122.
- [17] Kivela M, et al. (2014) Multilayer networks. *Journal of Complex Networks* 2:203.
- [18] Fienberg SE, Meyer MM, Wasserman SS (1985) Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association* 80:51–67.
- [19] Cardillo A, et al. (2013) Emergence of network features from multiplexity. *Scientific Reports* 3:1344.
- [20] Nicosia V, Latora V (2015) Measuring and modeling correlations in multiplex networks. *Physical Review E* 92:032805.
- [21] Menichetti G, Remondini D, Panzarasa P, Mondragón RJ, Bianconi G (2014) Weighted multiplex networks. *PloS one* 9:e97857.
- [22] Cantini L, Medico E, Fortunato S, Caselle M (2015) Detection of gene communities in multi-networks reveals cancer drivers. *Scientific Reports* 5:17386.
- [23] Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10:186–198.
- [24] Reis SD, et al. (2014) Avoiding catastrophic failure in correlated networks of networks. *Nature Physics* 10:762–767.
- [25] Battiston F, Nicosia V, Latora V (2014) Structural measures for multiplex networks. *Physical Review E* 89:032804.
- [26] Bianconi G (2013) Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E* 87:062806.
- [27] Szell M, Lambiotte R, Thurner S (2010) Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Science* 107:13636–13641.
- [28] Cardillo A, et al. (2013) 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* 215:23–33.
- [29] Iacovacci J, Wu Z, Bianconi G (2015) Mesoscopic structures reveal the network between the layers of multiplex data sets. *Physical Review E* 92:042806.
- [30] Battiston F, Iacovacci J, Nicosia V, Bianconi G, Latora V (2016) Emergence of Multiplex Communities in Collaboration Networks. *PLoS ONE* 11:e0147451.
- [31] Zhang J, Small M (2006) Complex Network from Pseudoperiodic Time Series: Topology versus Dynamics. *Physical Review Letters* 96:238701.
- [32] Kyriakopoulos F, Thurner S (2007) *Directed network representation of discrete dynamical maps* (Springer), pp 625–632.
- [33] Xu X, Zhang J, Small M (2008) Superfamily phenomena and motifs of networks induced from time series. *Proceedings of the National Academy of Science* 105:19601–19605.
- [34] Donner RV, Zou Y, Donges JF, Marwan N, Kurths J (2010) Recurrence networks: a novel paradigm for nonlinear time series analysis. *New Journal of Physics*

- 12:033025.
- [35] Donner R, et al. (2011) The geometry of chaotic dynamics—a complex network perspective. *The European Physical Journal B* 84:653–672.
 - [36] Bollobás B (2013) *Modern graph theory* (Springer Science & Business Media) Vol. 184.
 - [37] Pollock DSG, Green RC, Nguyen T (1999) *Handbook of time series analysis, signal processing, and dynamics* (Academic Press).
 - [38] Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) *Time series analysis: forecasting and control* (John Wiley & Sons).
 - [39] Lacasa L, Luque B, Ballesteros F, Luque J, Nuño JC (2008) From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Science* 105:4972–4975.
 - [40] Luque B, Lacasa L, Ballesteros F, Luque J (2009) Horizontal visibility graphs: Exact results for random time series. *Physical Review E* 80:046103.
 - [41] Gutin G, Mansour T, Severini S (2011) A characterization of horizontal visibility graphs and combinatorics on words. *Physica A: Statistical Mechanics and its Applications* 390:2421–2428.
 - [42] Lacasa L, Toral R (2010) Description of stochastic and chaotic series using visibility graphs. *Physical Review E* 82:036120.
 - [43] Iacovacci J, Lacasa L (2016) Sequential visibility-graph motifs. *Physical Review E* 93:042309.
 - [44] Fortunato S (2010) Community detection in graphs. *Physics reports* 486:75–174.
 - [45] Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* 9:09008.
 - [46] Capocci A, Servedio VD, Caldarelli G, Colaioni F (2005) Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications* 352:669–676.
 - [47] Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerà R (2004) Community analysis in social networks. *European Physical Journal B* 38:373–380.
 - [48] Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. *Physical review E* 69:026113.
 - [49] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10:10008.
 - [50] Guimera R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. *Physical Review E* 70:025101.
 - [51] Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Science* 105:1118–1123.
 - [52] Leck R, von Martin Reiterer R (2000) *Georg Simmel and Avant-Garde Sociology. The Birth of Modernity* (Prometheus Books, New York).

- [53] Caplow T (1956) A theory of coalitions in the triad. *American sociological review* 21:489–493.
- [54] Rapoport A (1953) Spread of information through a population with socio-structural bias: I. assumption of transitivity. *The bulletin of mathematical biophysics* 15:523–533.
- [55] Rapoport A (1953) Spread of information through a population with socio-structural bias: II. various models with partial transitivity. *The bulletin of mathematical biophysics* 15:535–546.
- [56] Barrat A, Barthelemy M, Vespignani A (2008) *Dynamical processes on complex networks* (Cambridge university press).
- [57] Newman M (2010) *Networks: an introduction*. United States: Oxford University Press Inc., New York pp 1–2.
- [58] Albert R, Jeong H, Barabási AL (1999) Internet: Diameter of the World-Wide Web. *Nature* 401:130–131.
- [59] Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382.
- [60] Pastor-Satorras R, Vespignani A (2001) Epidemic Spreading in Scale-Free Networks. *Physical Review Letters* 86:3200–3203.
- [61] Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. *Science* 286:509–512.
- [62] Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393:440–442.
- [63] Ramasco JJ, Dorogovtsev SN, Pastor-Satorras R (2004) Self-organization of collaboration networks. *Physical review E* 70:036106.
- [64] Newman ME (2001) The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98:404–409.
- [65] Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. *Physical Review E* 65:026107.
- [66] Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks. *Physical Review Letters* 88:128701.
- [67] Dorogovtsev S, Mendes J, Samukhin A (2001) Size-dependent degree distribution of a scale-free growing network. *Physical Review E* 63:062101.
- [68] Bianconi G, Rahmede C, Wu Z (2015) Complex quantum network geometries: Evolution and phase transitions. *Physical Review E* 92:022815.
- [69] Vázquez A (2003) Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E* 67:056104.
- [70] Jackson MO, Rogers BW (2007) Meeting strangers and friends of friends: How random are social networks? *The American economic review* 97:890–915.
- [71] Jin EM, Girvan M, Newman ME (2001) Structure of growing social networks. *Physical review E* 64:046132.

- [72] Toivonen R, Onnela JP, Saramäki J, Hyvönen J, Kaski K (2006) A model for social networks. *Physica A Statistical Mechanics and its Applications* 371:851–860.
- [73] Kumpula JM, Onnela JP, Saramäki J, Kaski K, Kertész J (2007) Emergence of communities in weighted networks. *Physical review letters* 99:228701.
- [74] Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) *The web as a graph: measurements, models, and methods* (Springer), pp 1–17.
- [75] Krapivsky PL, Redner S (2005) Network growth by copying. *Physical Review E* 71:036118.
- [76] Vázquez A, Flammini A, Maritan A, Vespignani A (2002) Modeling of protein interaction networks. *Complexus* 1:38–44.
- [77] Solé RV, Pastor-Satorras R, Smith E, Kepler TB (2002) A model of large-scale proteome evolution. *Advances in Complex Systems* 5:43–54.
- [78] Ispolatov I, Krapivsky PL, Yuryev A (2005) Duplication-divergence model of protein interaction network. *Physical Review E* 71:061911.
- [79] Pastor-Satorras R, Smith E, Solé RV (2003) Evolving protein interaction networks through gene duplication. *Journal of Theoretical biology* 222:199–210.
- [80] Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America* 102:13773–13778.
- [81] Dorogovtsev SN, Mendes JF (2002) Evolution of networks. *Advances in physics* 51:1079–1187.
- [82] Fortunato S, Barthelemy M (2007) Resolution limit in community detection. *Proceedings of the National Academy of Science* 104:36–41.
- [83] Marsili M, Vega-Redondo F, Slanina F (2004) The rise and fall of a networked society: A formal model. *Proceedings of the National Academy of Science* 101:1439–1442.
- [84] Holme P, Saramäki J (2012) Temporal networks. *Physics reports* 519:97–125.
- [85] Bianconi G, Barabási AL (2001) Bose-Einstein Condensation in Complex Networks. *Physical Review Letters* 86:5632–5635.
- [86] Caldarelli G, Capocci A, De Los Rios P, Munoz MA (2002) Scale-free networks from varying vertex intrinsic fitness. *Physical review letters* 89:258702.
- [87] Lancichinetti A, Kivela M, Saramäki J, Fortunato S (2010) Characterizing the Community Structure of Complex Networks. *PLoS ONE* 5:e11976.
- [88] Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP (2010) Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* 328:876.
- [89] De Domenico M, Lancichinetti A, Arenas A, Rosvall M (2015) Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems. *Physical Review X* 5:011027.
- [90] (2013) Aps data sets for research.
- [91] (2013) Pacs 2010 regular edition - american institute of physics.

- [92] Bianconi G (2008) The entropy of randomized network ensembles. *EPL (Europhysics Letters)* 81:28005.
- [93] Bianconi G (2009) Entropy of network ensembles. *Physical Review E* 79:036114.
- [94] Peixoto TP (2012) Entropy of stochastic blockmodel ensembles. *Physical Review E* 85:056122.
- [95] Bianconi G, Pin P, Marsili M (2009) Assessing the relevance of node features for network structure. *Proceedings of the National Academy of Science* 106:11433–11438.
- [96] Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. *Physical review E* 78:046110.
- [97] Rosvall M, Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Science* 104:7327–7331.
- [98] Sokal R, Michener C (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38:1409–1438.
- [99] Sokal RR, Rohlf FJ (1962) The comparison of dendrograms by objective methods. *Taxon* pp 33–40.
- [100] Zhao Y, Karypis G, Fayyad U (2005) Hierarchical clustering algorithms for document datasets. *Data mining and knowledge discovery* 10:141–168.
- [101] Newman MEJ (2006) From the Cover: Modularity and community structure in networks. *Proceedings of the National Academy of Science* 103:8577–8582.
- [102] Jaccard P (1912) The distribution of the flora in the alpine zone. *New phytologist* 11:37–50.
- [103] Rand WM (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66:846–850.
- [104] Kuncheva LI, Hadjitodorov ST (2004) *Using diversity in cluster ensembles* (IEEE), Vol. 2, pp 1214–1219.
- [105] Bianconi G, Darst RK, Iacovacci J, Fortunato S (2014) Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E* 90:042806.
- [106] Milo R, et al. (2004) Superfamilies of evolved and designed networks. *Science* 303:1538–1542.
- [107] Masoudi-Nejad A, Schreiber F, Kashani Z (2012) Building blocks of biological networks: a review on major network motif discovery algorithms. *IET systems biology* 6:164–174.
- [108] Lonardi J, Patel P (2002) *Finding motifs in time series* pp 53–68.
- [109] Mueen A, Keogh E, Zhu Q, Cash S, Westover B (2009) *Exact discovery of time series motifs* (SIAM), pp 473–484.
- [110] Bandt C, Pompe B (2002) Permutation Entropy: A Natural Complexity Measure for Time Series. *Physical Review Letters* 88:174102.
- [111] Amigó J (2010) *Permutation Complexity in Dynamical Systems*.

- [112] Kulp CW, Zunino L (2014) Discriminating chaotic and stochastic dynamics through the permutation spectrum test. *Chaos* 24:033116.
- [113] Gutin G, Mansour T, Severini S (2011) A characterization of horizontal visibility graphs and combinatorics on words. *Physica A Statistical Mechanics and its Applications* 390:2421–2428.
- [114] Iacovacci J, Lacasa L (2016) Sequential motif profile of natural visibility graphs. *Physical Review E* 94:052309.
- [115] Lacasa L (2014) On the degree distribution of horizontal visibility graphs associated with Markov processes and dynamical systems: diagrammatic and variational approaches. *Nonlinearity* 27:2063.
- [116] Lacasa L, Flanagan R (2015) Time reversibility from visibility graphs of nonstationary processes. *Physical Review E* 92:022817.
- [117] (2013) Wolfram mathematica.
- [118] van Kampen NG (1995) Stochastic processes in physics and chemistry.
- [119] Bishop CM (2006) Pattern recognition. *Machine Learning* 128:1–58.
- [120] Lutz A, Greischar LL, Rawlings NB, Ricard M, Davidson RJ (2004) Long-term meditators self-induce high-amplitude gamma synchrony during mental practice. *Proceedings of the National academy of Sciences of the United States of America* 101:16369–16373.
- [121] Lutz A, et al. (2009) Mental training enhances attentional stability: neural and behavioral evidence. *Journal of Neuroscience* 29:13418–13427.
- [122] Tang YY, Hölzel BK, Posner MI (2015) The neuroscience of mindfulness meditation. *Nature Reviews Neuroscience* 16:213–225.
- [123] Jiang S, Bian C, Ning X, Ma QD (2013) Visibility graph analysis on heartbeat dynamics of meditation training. *Applied Physics Letters* 102:253702.
- [124] Sarkar A, Barat P (2008) Effect of meditation on scaling behavior and complexity of human heart rate variability. *Fractals* 16:199–208.
- [125] Peng CK, et al. (1999) Exaggerated heart rate oscillations during two meditation techniques. *International journal of cardiology* 70:101–107.
- [126] Goldberger A, et al. (2000) Physiobank, physiokit, and physionet: Components of a new research resource for complex physiologic signals. *circulation* [online]. 101 (23), pp. e215–e220.
- [127] Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23:1495–1502.
- [128] Amigó JM, Zambrano S, Sanjuán MAF (2007) True and false forbidden patterns in deterministic and random dynamics. *EPL (Europhysics Letters)* 79:50001.
- [129] Carpi LC, Saco PM, Rosso OA (2010) Missing ordinal patterns in correlated noises. *Physica A Statistical Mechanics and its Applications* 389:2020–2029.
- [130] Rosso OA, et al. (2012) Causality and the entropy-complexity plane: Robustness and missing ordinal patterns. *Physica A Statistical Mechanics and its Applications*

391:42–55.

- [131] Zanin M, Zunino L, Rosso OA, Papo D (2012) Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review. *Entropy* 14:1553–1577.
- [132] Nicosia V, Bianconi G, Latora V, Barthelemy M (2013) Growing Multiplex Networks. *Physical Review Letters* 111:058701.
- [133] Nicosia V, Bianconi G, Latora V, Barthelemy M (2014) Nonlinear growth and condensation in multiplex networks. *Physical Review E* 90:042807.
- [134] Catanzaro M, Boguñá M, Pastor-Satorras R (2005) Generation of uncorrelated random scale-free networks. *Physical Review E* 71:027103.