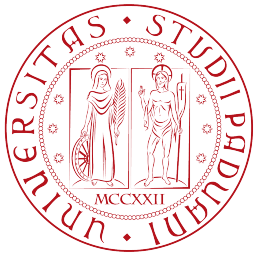


UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics

Final Dissertation

Quantifying the robustness of interconnected systems from their collective dynamics

Thesis supervisor

Prof. Manlio De Domenico

Thesis co-supervisor

Dr. Oriol Artime

Candidate

Chiara Veronese

Academic Year 2021/2022

Contents

Preface

Complex networks are powerful tools to investigate real-world systems in many areas of science, especially those of interest in Physics such as biophysical and ecological systems. Networks can model complex systems through a convenient representation in terms of nodes and their connections. However, we cannot always describe interactions among elements through physically present links, such as electric power transmission links in a power grid: in some cases, the structural features of a complex system have to be statistically inferred by observing the time course of relevant physical quantities. That is a well-known inverse problem which, to date, has not been yet satisfactorily solved, although it is a very common scenario when modelling interconnected systems. Solving this problem would allow us to characterize the macroscopic features of a networked system, from mesoscale organization to critical behaviour.
To this aim, the goal of my thesis is to bridge the observation of collective dynamics, in terms of multivariate time series, with the structural and functional robustness of the underlying complex network.

Specifically, we will explore the limits of the fuzzy network approach, recently introduced, where uncertainty about the existence of edges results in an ensemble network reconstruction. Besides being a new approach for network inference, it has not yet been used to investigate the resilience behaviour of a complex network, both from a theoretical and computational viewpoint.
In network science, we use percolation theory to address robustness, a key aspect in understanding systems' macroscopic response to failures and perturbations. Our theoretical analysis aims at extending this formalism to the case where edges are defined by an existence probability since existing techniques are only valid for physical links.
In parallel, the computational analysis considers synthetic networks and Kuramoto dynamics on top of them to characterize robustness through the time course of oscillator phases far from equilibrium. By assuming no knowledge about the network behind the Kuramoto dynamics, we first test the standard procedure where interactions are inferred by thresholding the correlation matrix obtained from time series on each network node. We study the limit of this method by analyzing various structural properties. We then elude the reconstruction phase and opt for the fuzzy network approach, where we deal with a network ensemble sampled by the probability adjacency matrix reconstructed from the correlation one. The analysis focuses on finding the right conditions under which the

network ensemble better represents the robustness properties of the synthetic network underneath.

In the first chapter, after briefly introducing the original Kuramoto model, we focus on the toy model from which we extract the multivariate time series analyzed in this work. Indeed, the first chapter provides the reader with the main properties of the stochastic Kuramoto model on networks (SKMN). Moreover, we discuss the importance of considering all the relevant dynamical stages characteristic of the synthetic data we create. We highlight the importance of time integration to capture the entire synchronization process (from the transient time until stationarity) since oscillators' phase time series are the unique observables from which to recover network topology.
The second chapter is devoted to network topology reconstruction by thresholding procedure. That is widely used as a standard procedure to infer the network interaction structure between dynamical units from multivariate time series. Though in the third chapter we explore the limits of this standard reconstruction procedure, this discussion is of central importance. Indeed, we look for the best model parameter set (coupling, noise and time region) for which the thresholding reconstruction procedure shows the best reconstruction.
The third chapter opens by discussing the reliability of the thresholding criterion choice to assess the connection among units in a system. Therefore, we introduce an alternative network inference method: the fuzzy network model. That is the main reconstruction procedure of this thesis. We discuss the random permutation (RP) and iterative amplitude fast Fourier transform (IAFFT) methods to create the null hypothesis testing to recover the fuzzy matrix Π from the SKMN dynamics. Π resembles a weighted adjacency matrix, instead, each entry π_{ij} defines the existence probability for the ij link in the network. After computing Π, in the last chapter, we focus on the probabilistic perspective of this approach. All the network descriptors must be re-defined as random variables since they reflect the uncertainty about edge existence. Therefore, we can define the fuzzy counterpart of some basic structural descriptors as the node excess degree distribution. Finally, we focus on the robustness problem. After extending the theory of percolation to fuzzy networks, we want to assess whether the reconstructed network resilience properties are representative of the synthetic network underneath the SKMN model.

Chapter 1

Modelling dynamical processes on complex networks

Networks are powerful tools for studying real-world systems in many areas. Networks of citations between papers, power grids, social networks, human transportation, brain neural structure or earth climate are some examples of a vast range of systems we can model through network tool.
Networks can model real systems through a convenient representation in terms of nodes and edges. Nodes individuate the many units that make up the analyzed system whereas edges individuate relationships or interactions among them.
In the following, networks and graphs are synonyms: we use the word graph to refer to the abstract mathematical concept, while network mainly concerns real system modelling. Both terms can be used without distinctions, however, we state a clear preference over the term network.
Network theory has its roots in Graph theory which dates back to Euler's famous Königsberg bridge problem [1]; since then it has become a relevant interdisciplinary topic, pervading many fields from sociology to communication from physics and biology. In the past few years, the availability of large amounts of data and a considerable increase in computing power has given a new empirically driven momentum to such an attractive and powerful theoretical framework.

1.1 Basic introduction to networks

Most of the basic theoretical tools introduced to describe and analyze real-world networks come from graph theory. An undirected (directed) *graph* $G = (N, E)$ is a collection of N *nodes* linked by E *edges*. We represent each edge in our network as an unordered (ordered) pairs of distinct vertices, so that $e_{ij} \in E \subseteq N \times N$ (where i and j both belong to N set). A network that has neither *self-edges* ($e_{ii} = 0,\ \forall i \in V$) nor *multiple edges* (multiple definitions for e_{ij}) is called a *simple network* or *simple graph*. A network with multiple

edges is called a *multi-graph*.
A simple graph is completely determined by the (N, E) collection and this information can be written in the following matrix form:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{NN} \end{bmatrix} \tag{1.1}$$

A, the *adjacency matrix*, is the uniquely related to each graph $G = (N, E)$. $A \in \mathcal{M}(N \times N, \mathbb{Z})$ records each link between node i to node j with 1 in the i-th row and the j-th column; 0 if the link is not present.

$$a_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

Networks with symmetric A are undirected otherwise they are directed. Self-edges, also called loops, contribute to a one to the main diagonal entries. Moreover, we can take into account multiple edges by substituting the 1 value in (1.2) with the opportune multiplicity m_{ij}
This basic mathematical structure can be complicated in many ways: e.g. real numbers can be attached to each edge. Then, we can write $a_{ij} = w_{ij}$: that is the *weighted adjacency matrix*.

With the help of the adjacency matrix we can define many important properties of a network. Newman's *Networks: An Introduction* [2] offers a clear exposition of common classes of graphs and their properties.Let us introduce some of them.
The number of edges $|E|$ can be recovered as $|E| = \frac{1}{2}\sum_i \sum_j a_{ij}$.
The degree k_i of i-th node is the number of nodes to which i is connected to.

$$k_i = \sum_j a_{ij} = \sum_j a_{ji}, \tag{1.3}$$

for an undirected graph. If G is directed the *in-degree* k_i^{in} is the number of i's in-coming edges, while the *out-degree* k_i^{out} counts its out-going edges.

$$k_i^{in} = \sum_j a_{ji} \qquad \text{and} \qquad k_i^{out} = \sum_j a_{ij}. \tag{1.4}$$

An important property of networks is the *mean vertex degree*, which stands for the average number of neighbours of a generic node in G.

$$\langle k \rangle = \frac{\sum_i k_i}{N} = \frac{2|E|}{|N|}. \tag{1.5}$$

Then, we define the network connectance (*sparsity* or *density*) as the fraction of edges to all possible links between N vertices:

$$\rho = \frac{|E|}{\binom{N}{2}} \tag{1.6}$$

To conclude the list of most important properties, we introduce the concept of *path*, a sequence of edges connecting two nodes and its *length* is the number of edges traversed in the path. A *component* (connected component C_G is a subgraph of G, where each vertex $i \in C_G \subseteq N$ can be reached bu all other vertices $j \in C_G$ following paths running along edges. We summarize the main properties in Figure 1.1.

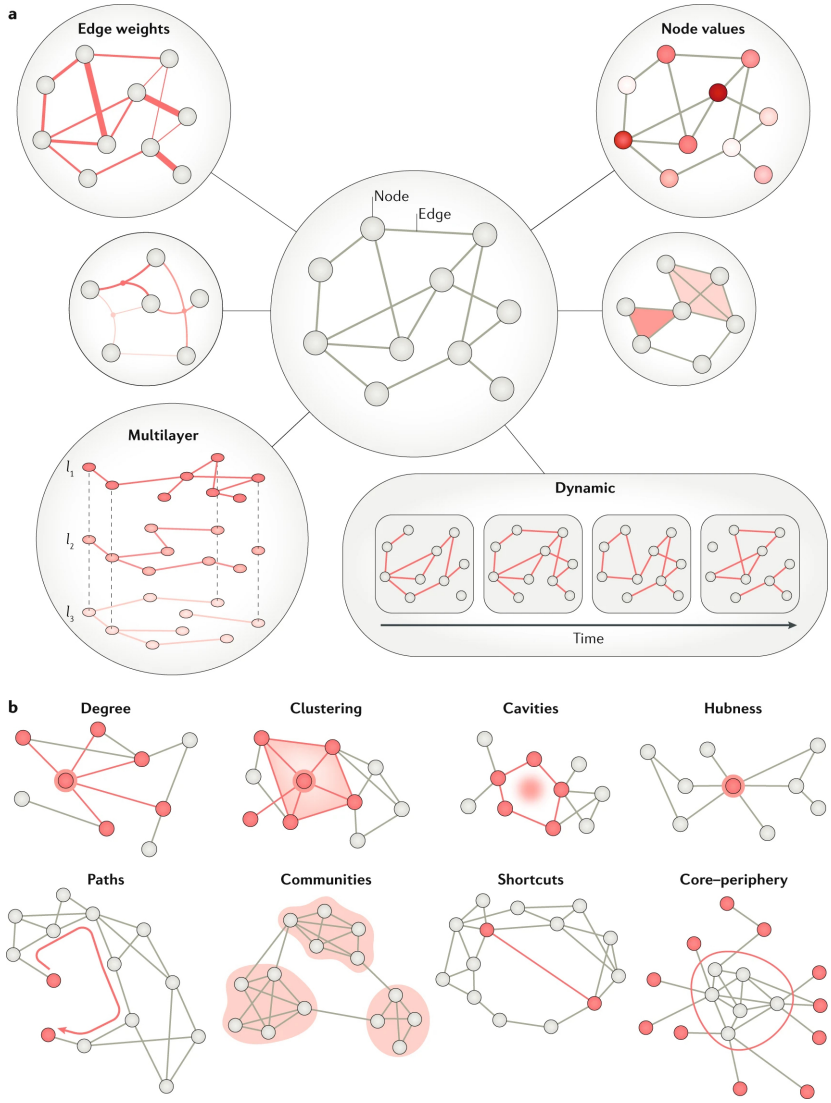


Figure 1.1: (a) The simple network model that represents the pattern of connections (edges) between neural units (nodes). In figure are also shown more sophisticated network models such as weighted graphs, graphs with explicit functional forms for their dynamics or multi-layer networks. (b) Common measures of interest in a network G. Image from [3].

1.2 Introduction to dynamics on networks

Network theory has recently become of interest to investigate empirical systems in many areas of science. Indeed, it provides a modern and powerful tool in complex systems analysis, to better understand their behaviour with the help of network modelization. One key question is how to turn the modelling results into conclusions or predictions about how the overall system behaves.
The statistical analysis revealing the underlying principles of highly complex topological structures is only one step towards this kind of understanding. On top of that, another relevant issue is to highlight the interplay between structure and function (Figure 1.2).
Let us consider, for instance, the extreme importance of the combination of those two characteristics in the study of the human brain [4,5] or in the new insights for the stability analysis of the climate system [6,7].

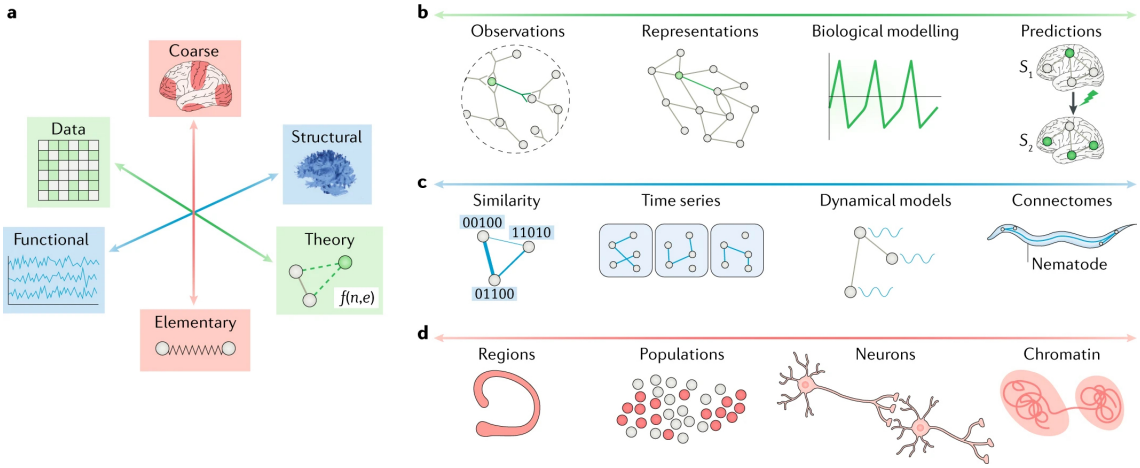


Figure 1.2: (a) Three dimensions of network model types. The human brain network. The image highlights the first dimension in green: (b) the network theory representation of the brain from data to graph modelling. In blue (c) the before-mentioned interplay between structure and function which allows us to acknowledge how connectomes work by looking for example at EEG time series. In red (d) the third dimension, we do not discuss, extends network model from elementary descriptions to coarse-grained approximations. Image from [3]

Nowadays, studying the emergence of collective dynamics in complex networks has increased the attention to relating the dynamics of a network to its topology and local properties. Recent studies show that dynamical processes, like network synchronization, one focus of this work, are strongly influenced by the structure of the underlying network [8–10]. Furthermore, in many realistic systems, dynamics feedback can reshape the network topology, but this is far beyond the goal of this work, although absolutely fascinating.

In this thesis work, we will focus on the interplay between network structure and dynamics. To be more specific, in this first chapter, we will mainly deal with the stochastic Kuramoto model, the dynamics we choose to study on top of various synthetic networks.

A generic complex system is composed of non linearly interacting, many N components. Each i-th system unit can be generally described by one or more physical observables which define its own state. Let $\boldsymbol{x}_i$ denote this internal D-dimensional state, then $\boldsymbol{x}_i(t) = [x_i^{(1)}(t), x_i^{(2)}(t), ..., x_i^{(D)}(t)]^T \in \mathbb{R}^D$ is its complete notation at time t.
The general evolution of the state is governed by a system of N differential equations:

$$\dot{\mathbf{x}}_i(t) = \boldsymbol{f}_i(\boldsymbol{x}_i(t)) + \sum_{j=1}^{N} A_{ij}\boldsymbol{g}_{ij}(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t)) + \boldsymbol{I}_i(t) + \boldsymbol{\xi}_i(t) \tag{1.7}$$

where $i, j \in \{1, 2, ..., N\}$; $\boldsymbol{f}_i : \mathbb{R}^D \to \mathbb{R}^D$ and $\boldsymbol{g}_{ij} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$, respectively, define the intrinsic and interaction dynamics of the D-dimensional units.
The function $\boldsymbol{I}_i(t)$ represents external drivers, whereas $\boldsymbol{\xi}_i(t)$ stands for the stochastic term due to noise, be it additive or multiplicative [11].
Indeed, eq. (1.7) is a system of N stochastic differential equations (SDE). Two different components contribute to the system's evolution: the generally non-linear and time-dependent drift term and the volatility component. The drift term is identified by the nonlinear function $\boldsymbol{f}$, $\boldsymbol{g}$ and $\boldsymbol{I}$. For each $d \in [0, 1, \ldots, D]$, the second term $\xi_i^{(d)}(t) = \Sigma(\boldsymbol{x}^{(d)}(t))dW_i(t)/dt$ introduces a stochastic contribution, where $\boldsymbol{W}^{(d)}(t)$ is a N-dimensional Wiener process with the following constraints:

$$\left\langle \boldsymbol{\xi}^{(d)}(t)(\boldsymbol{\xi}^{(d)}(t'))^\tau \right\rangle = 2\epsilon \mathbb{1}_N \delta(t - t') \text{ and } \Sigma(\boldsymbol{x}^{(d)}(t))\Sigma(\boldsymbol{x}^{(d)}(t))^\tau = D(\boldsymbol{x}^{(d)}(t)). \tag{1.8}$$

Here ϵ represents a non-negative numerical constant taking into account noise strength. Referring to Σ matrix, the random noise can be additive or multiplicative considering or not the $\boldsymbol{x}^{(d)}$ dependence.
Last but not least for our analysis, the term A_{ij} defines the interaction topology in terms of the adjacency matrix A, such that $A_{ij} = 1$ if there is a direct physical interaction from the unit j to i and $A_{ij} = 0$ otherwise.
Adjacency matrix univocally defines a network, which is an abstraction used to model a system that contains discrete, interconnected elements. In network modelling, nodes (also called vertices) individuate interacting units in the system, whereas the interconnections are represented by edges (links).

Let us now consider a specific realization of eq. (1.7), where the i-th unit is one of N oscillators in a system and its unique internal state is its phase. The drift term contains both an intrinsic but trivial contribution given by i-th oscillator's natural frequency and an interaction term considering a complex topology. At last, our dynamics model includes an additive noise term. This is the stochastic Kuramoto model on network, but first let us introduce the original Kuramoto case.

1.3 Introduction to Kuramoto model

Kuramoto model [12] takes part in the mathematical approaches to tackle the problem of collective synchronization. It models a system made up of oscillators, which exert a phase-dependent influence on the others. Such a system shows an equilibrium phase transition.

If the coupling is too weak compared to the natural frequencies of the oscillators, the system behaves incoherently as oscillators cannot synchronize. However, all oscillators constantly evolve in the phase space into synchrony when the coupling is strong enough. Kuramoto proposed the following governing equations for each oscillator phase θ_i in the system:

$$\dot{\theta}_i(t) = \omega_i + \frac{\sigma}{N} \sum_{j=1}^{n} \sin(\theta_i(t) - \theta_j(t)) \tag{1.9}$$

where σ is the coupling strength among oscillators and ω_i is the natural frequency of the i-th oscillator. The frequencies ω_i are distributed according to some function $g(\omega)$, which is usually assumed to be unimodal and symmetric about its mean frequency. The factor $1/N$ is included to ensure the good behaviour of the model in the thermodynamic limit.

The macroscopic phase coherence $r(t)$ describes the collective dynamics of the whole population of our problem, moreover, it is the order parameter linked to the phase transition Kuramoto system shows at varying coupling (Figure 1.3). The concept of the order parameter is very useful for a quantitative theory of phase transitions, in general, and critical phenomena, in particular [14, 15]. It consists of a characteristic (macroscopic) quantity (r_∞) of the system assuming values different from zero above a threshold (critic) parameter (σ_c) and being zero otherwise (Figure 1.4). When symmetry is broken by lowering the coupling strength below σ_c, this special quantity will be linked to that symmetry. In other words, this quantity measures the degree of order/synchronization in the system.

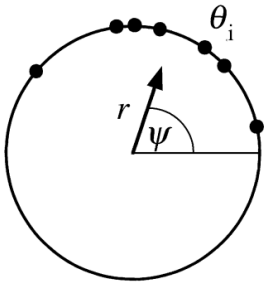


Figure 1.3: Geometric interpretation of the order parameter (1.10). The phases θ_i are plotted on the unit circle. Their centroid is given by the complex number $re^{i\psi}$, shown as an arrow. Image from [13].

$$r(t)e^{i\psi(t)} = \frac{1}{N} \sum_{j=1}^{n} e^{i\theta_j(t)} \tag{1.10}$$

where $\psi(t)$ is the average phase. The module of $r(t)$ ranges in $0 \leq r(t) \leq 1$, the two limits $r(t) \simeq 0$ and $r(t) \simeq 1$ respectively describe the condition in which all the oscillators are either phase locked or move incoherently.
Manipulation of eq. (1.9) and eq. (1.10) allows for an analytical treatment in terms of the mean field approach.

$$r(t)\sin(\psi(t) - \theta_i(t)) = \frac{1}{N} \sum_{j=1}^{n} e^{i\theta_j(t)} \tag{1.11}$$

$$\dot{\theta}_i(t) = \omega_i + \sigma r(t)\sin(\psi(t) - \theta_i(t)) \tag{1.12}$$

Mean field approach analytical treatment operates in equation (1.12) as each oscillator interacts with all the others only through the global quantities $r(t)$ and $\psi(t)$. The oscillator phase is indeed the result of its intrinsic natural frequency ω_i, besides second term in eq. (1.12) provides the most relavant and interesting contribution. The factor $r(t)$ provides a positive feedback loop: as $r(t)$ increases the collective dynamics becomes more coherent, resulting in strengthening the interconnection between the oscillators and including more and more of them in the "coherent pack".

Moreover, we calculate the critical coupling σ_c by looking for steady solutions of eq. (1.12). Stationarity is then recovered at the $\lim_{t\to\infty}$ and allows for the assumption $r(t)$ and $\psi(t)$ being constant. Without loss of generality, we can set $\psi = 0$, therefore going in the so-called co-moving reference frame. This transformation leads to the equations of motion [12, 16]:

$$\dot{\theta}_i = \omega_i - \sigma r \sin(\theta_i) \quad (i = 1, ..., N) \; . \tag{1.13}$$

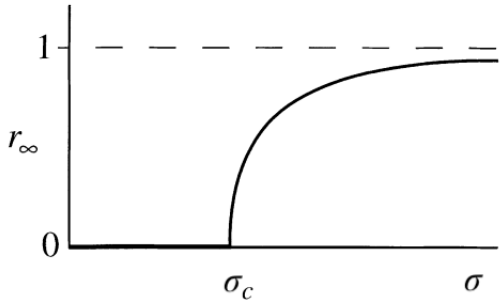


Figure 1.4: Dependence of the steady-state coherence r_∞ on the coupling strength σ. Image from [13].

When the coupling is larger than the critical value, σ_c, the solution of eq. (1.13) contemplates two different types of long-term behaviour for oscillators in the system, depending on the size of $|\omega_i|$ relative to σr. Numerical simulations of the model verified the following results. Oscillators for which $|\omega_i| \leq \sigma r$, approach a stable fixed point defined implicitly by:

$$\omega_i = \sigma r \sin\theta_i, \tag{1.14}$$

These oscillators belong to a phase-locked bulk at frequency Ω in the original frame. The remaining oscillators, subject to $|\omega_i| > \sigma r$, drift around the circle, sometimes accelerating and sometimes rotating at lower frequencies.
The following self-consistent equation can be derived for r, by imposing that drifting oscillators follow some stationary distribution [13],

$$r = \sigma r \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\cos^2\theta\right) g(\omega) d\theta, \tag{1.15}$$

where $\omega = \sigma r \sin(\theta)$. This equation admits a non-trivial solution, beyond which $r > 0$.

$$\sigma_c = \frac{2}{\pi g(0)}. \tag{1.16}$$

Expression in eq. (1.16) is the value of critical coupling at the onset of synchronization. Moreover, near the onset of synchronization, the order parameter, r, obeys the following law,

$$r \sim (\sigma - \sigma_c)^{\beta} \tag{1.17}$$

with $\beta = 1/2$. Indeed, it turns out that β critical exponent gives much deeper information than the critical coupling σ_c itself [17]. The reason is that the latter significantly relies on the microscopic details of the system, while the exponents are, in a certain sense, universal as they depend only on a few fundamental parameters. This introduces us to the beautiful concept of universality [18].

The Kuramoto model (KM, from now on) approach to synchronization was a breakthrough for understanding synchronization in large populations of oscillators.
However, when dealing with more realistic cases, e.g. for finite populations of oscillators or including some noise term, analytical attempts show unsolved problems and rise questions regarding global stability results [19]. In what follows, we introduce noise term and a non-trivial connection topology [2].

1.4 Stochastic Kuramoto model on network

In the following we reformulate the governing equations in the Kuramoto model to include both noise and some connectivity different from the complete-graph topology involved in the original KM in eq. (1.9).
The dynamics of the stochastic Kuramoto model on network (SKMN) is then described by:

$$\dot{\theta}_i(t) = \omega_i + \frac{\sigma}{k_i} \sum_{j=1}^{N} A_{ij} \sin(\theta_j(t) - \theta_i(t)) + \xi_i(t) \tag{1.18}$$

where A_{ij} takes into account the new complex topology , the coupling is normalized on the node degree k_i as prescripted in section 3.1.2. in [20]: the inclusion of weights in the interaction impose a dynamic homogeneity that masks the real topological hetereogeneity of the network. The stochastic part of eq. (1.18) involves some additive Gaussian noise, as

$$\langle \xi_i(t) \rangle = 0 \quad \text{and} \quad \langle \xi_i(t) \xi_j(t') \rangle = 2\epsilon \delta_{ij} \delta(t - t'). \tag{1.19}$$

We introduce the noise term in our model in order to make it more realistic. Moreover, the stochastic term dampens the phase oscillator synchronization, as it could be problematic in order to reconstructing the topology.

1.4.1 Comments on SKMN and computational implications

After having introduced the SKMN, we remind that, in the next two chapters, we want to recover the system's topology from Kuramoto collective dynamics.
As phase time series from N oscillators are the unique observables from which to recover such relevant information, we must be very attentive to the different parameters involved in the SKMN.
First, we notice that coupling σ and noise intensity ϵ are two competing parameters.
An increase in σ implies that the oscillator's phases in the systems synchronize quickly. Viceversa, a noise intensity increase makes the synchronization less stable and slower.
Besides influencing the time at which the collective dynamic synchronizes, both features affect the collective amount of synchronization in the system. Indeed, the global order parameter r shows a critical transition in varying the coupling, but it is also partially conditioned to the stochastic term of SDE eq. (1.18).

$$r(t)e^{i\psi(t)} = \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j(t)}, \tag{1.20}$$

From previous considerations on quicker and slower dynamic responses to varying parameters, let us consider T, the total integration time of the SDE.
In fact, T is a crucial feature of time series analysis. As we are dealing with the Kuramoto model, we want to set T long enough to capture the entire synchronization process, from incoherence to coherence (see both panels in Figure 1.5). Then, the integration time setting must consider both the time transient, in which oscillators are still decoupled and random (e.g. low $r(t)$ in Figure 1.5B, and the stationary state, where synchronization of the system reaches its maximum strength.

Using different T times in the Kuramoto processes has two relevant computational implications, involving the algorithm to integrate the SDE.
In this regard, we here present the computational integration method we opt for in this work, which is the Euler-Marujama (E-J) scheme [21].
We know that the E-J algorithm is more accurate and stable in integrating a general SDE, more the infinitesimal integration interval dt is small.
This implies that:

- Whenever we are dealing with low coupling, we choose dt slightly higher to avoid long computational times, but meantime such that not to nullify Euler-Marujama precision.
- We pay attention to having sufficiently "dense" time series.
  If $\theta_i(t) = [\theta_i(0), \theta_i(dt), ..., \theta_i(T - dt), \theta_i(T)]$ is the considered time series we want the components to be numerous enough to compute a meaningful correlation matrix. Correlation matrix is one of the main component for the network inference we will deal with in the next chapters.

The following table specifies the time integration T and dt interval which must be chosen

to obtain a meaningful collective dynamics from SKMN eq. (1.18) on the chosen network.

σ	1.5	1.75	2	2.5	3	3.5	7	11	15
E-R T	60	50	40	25	22.5	20	5	4	3
B-A T	62.5	52.5	42.5	26	23.5	22.5	7	4	3
W-S T	65	55	45	30	27.5	25	10	5	4
dt	0.001	0.001	0.001	0.001	0.001	0.001	0.00025	0.00025	0.0001

Table 1.1: dt and T values for integrating eq. (1.18) depending on the coupling and network chosen. E-R stands for Erdos-Renyi, B-A is the Barabasi-Albert model and W-S is Watts-Strogatz. All networks have similar average degree equal to ~ 12.

We notice that the time integration T depends mainly on the chosen coupling strength σ. Moreover, we highlight the fact that T is chosen such that the dynamics reaches its steady state (r_∞) more or less when $t \to T/2$ (obviously for $\sigma > \sigma_c$).
We then report some graphs validating the fact that the entire synchronization process is captured.

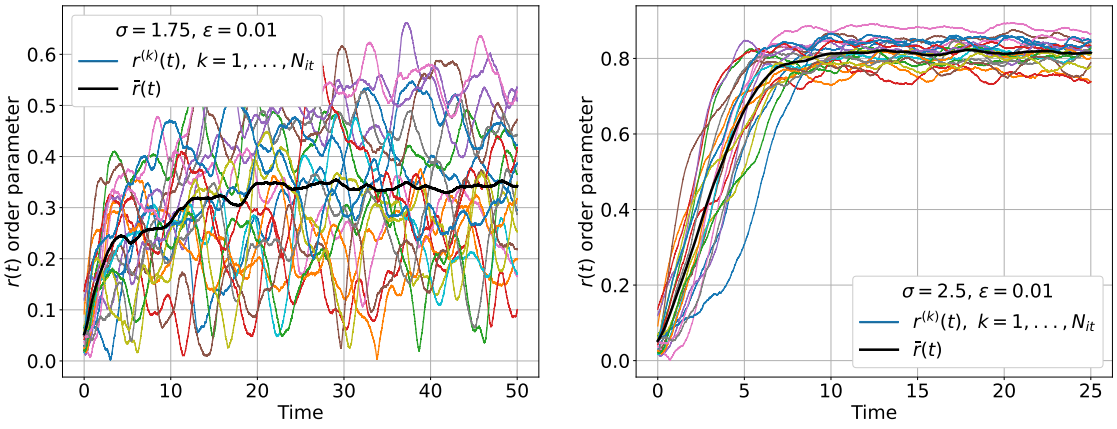


Figure 1.5: The soft coloured lines represent the global order parameter $r^{(k)}(t)$, where superscript $^{(k)}$ individuates the k-th process in the noise ensemble $\{(r^{(k)}(t))\}_{k=[1,2,\ldots,N_{it}]}$, $N_{it} = 20$. Instead, the solid black line is $\bar{r}(t)$ the average synchronization parameter. Both panels show the order parameter derived from SKMN on Erdos-Renyi network ($\langle k \rangle \sim 12$) and $\epsilon = 0.01$, however (A) has $\sigma = 1.75$, whereas (B) $\sigma = 2.5$, integration time T and time interval dt are choosen such that ehe entire synchronization process is captured (see Table 1.1).

From Figure 1.5, we highlight the presence of two panels. The introduction of noise in SKMN dynamics forces us to consider some statistical ensemble of synchronization dynamics processes with equal parameters; we then change the noise seed and the initial conditions both on $\theta_i(0)$, the initial phases of the oscillators, and ω_i, their natural frequencies.
The soft coloured lines represent the global order parameter $r^{(k)}(t)$, where superscript $^{(k)}$

individuates the k-th process in the noise ensemble $\{(r^{(k)}(t))\}_{k=[1,2,\ldots,N_{it}]}$. Instead, the solid black line is $\bar{r}(t)$, the average synchronization parameter.

$$\bar{r}(t) = \frac{1}{N_{it}} \sum_{k=1}^{N_{it}} r^{(k)}(t) \tag{1.21}$$

To have a significant ensemble but small enough not to have a computationally expensive algorithm, we consider an ensemble of $N_{it} = 20$ processes for each noise intensity and coupling.
That is an arbitrary choice but allows us to determine whether or not the average synchronization process reaches stationarity.

1.5 Computational results

In the previous section, we describe the main characteristics of the SKMN, introduce the computational scheme to integrate its SDE (1.18) and highlight the importance of capturing the entire synchronization process to implement a fair analysis of our model.
In the following discussion, we focus on how the two main parameters coupling σ and noise intensity ϵ affect the dynamics.
Specifically, we aim to recover the landscape, which describes the stationary $\lim_{t\to\infty} r(t)$ macroscopic complex order parameter in eq.(1.22), as a function of the two free parameters of model (1.18): ϵ and σ.

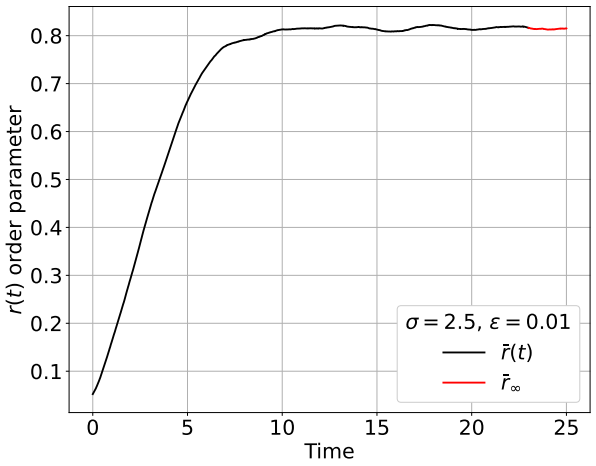


Figure 1.6: Synchronization process on an Erdos-Renyi network of 256 nodes ($\langle k \rangle \sim 12$), with $\epsilon = 0.01$ and $\sigma = 2.5$, T=25 and dt=0.001.

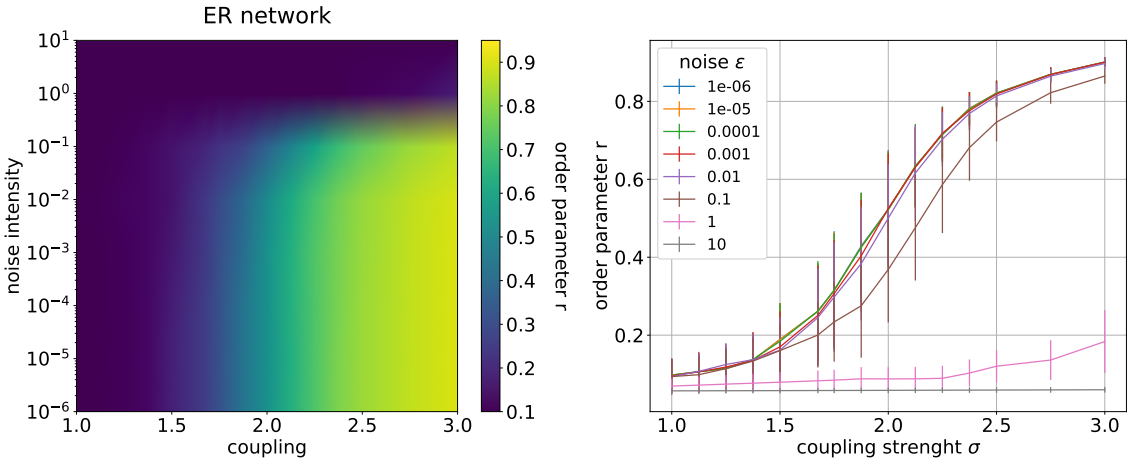
For a fixed network topology, we calculate the stationary r_∞ global order parameter by

looking at the synchronization process in Figure 1.6.
Following what we learned in section 1.4.1, we fix the coupling and the noise and set T the total time of the process long enough to capture the synchronization process, on average. The fact we are looking to the average synchronization process is due to the fact that we are working with a stochastic model. Indeed, we must consider the average order parameter $\bar{r}(t)$ in eq. (1.21) due to the noise ensemble N_{it} we introduced above.

Fixing coupling and noise, we iterate the process for $N_{it} = 20$ times by changing the noise seed and the initial conditions both on $\theta_i(0)$, the initial phases of the oscillators and ω_i, their natural frequencies.
N.B.: $\theta_i(0)$ are extracted from a uniform distribution $\mathcal{U}_{[a,b]}$ where $[a, b] = [-2\pi, 2\pi]$, whereas ω_i come from a normal distribution $\mathcal{N}(0, 1)$. Both distributions do not affect the Kuramoto model; specifically, the ω_i distribution only affects how the critical threshold coupling is approached and not the value itself.

Therefore to calculate r_∞ we work with $\bar{r}(t)$ quantity defined in eq. (1.21). In the example in Figure 1.6, the image represents the averaged order parameter $\bar{r}(t)$ at various instants in time. We are left with a process of $N_{steps} = \frac{T}{dt}$; we average over the final 2000 steps (red part of the plot in Figure 1.6) to find the stationary r_∞ parameter.

In the following we present the results for various network typologies.

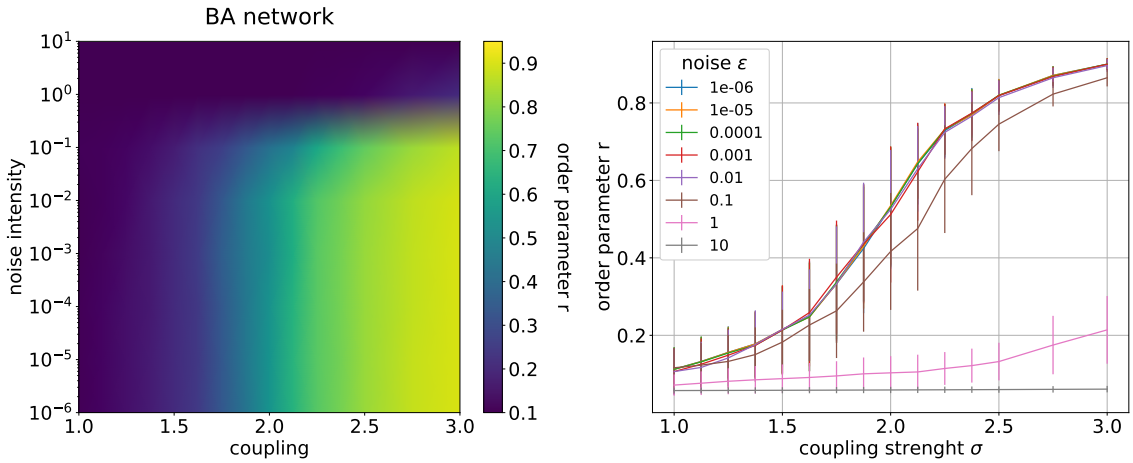


(A) (B)

Figure 1.7: (A) Landscape representing the order parameter r_∞ at stationary conditions, by varying the two main parameters of SKMN dynamics, that are noise intensity ϵ and coupling strength σ. The network underneath the model is the Erdos-Renyi with $\langle k \rangle \simeq 12$. **Pay attention to the log10 scale in noise intensity axis**, (B) 2-dimensional reproduction of results in panel A.

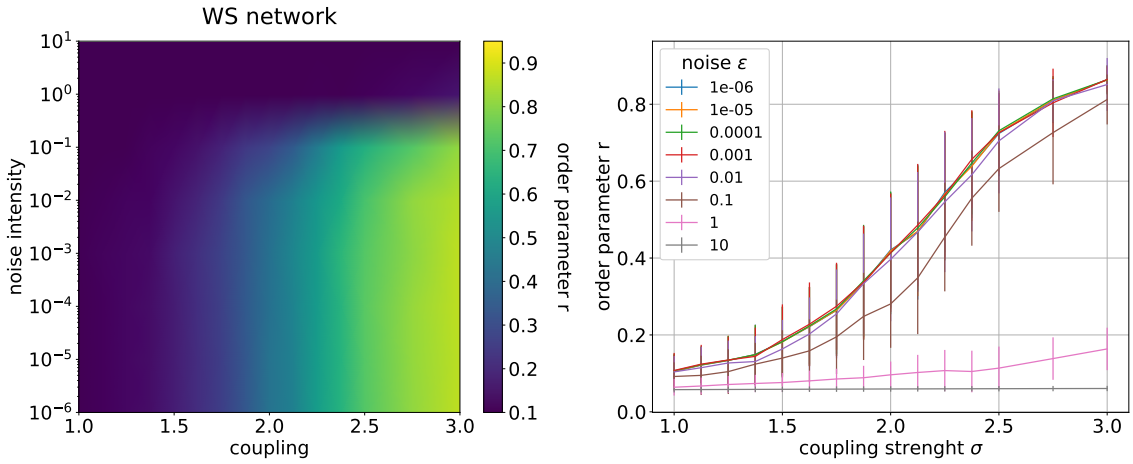
We want to reproduce the results of the following landscape as they show a non-trivial behaviour. We hope that some formula can analytically fit this landscape and, specifically,

we would like to reproduce the behaviour of the critical coupling σ_c as a noise function. The fact the transition is smooth (see Figure 1.7B) rather than having the theoretical form in Figure 1.4 is due to finite-size effect and to the role of noise in giving a realistic taste to the model [19].



(A) (B)

Figure 1.8: (A) Landscape representing the order parameter r_∞ at stationary conditions, by varying ϵ and σ in the SKMN dynamics. The network underneath the model is the Barabasi-Albert with $\langle k \rangle \simeq 12$. (B) 2-dimensional reproduction of results in panel A.



(A) (B)

Figure 1.9: (A) Landscape representing the order parameter r_∞ at stationary conditions, by varying ϵ and σ in the SKMN dynamics. The network underneath the model is the Watts-Strogatz with $\langle k \rangle \simeq 12$. (B) 2-dimensional reproduction of results in panel A.

1.5.1 Future research perspectives

At the end of previous paragraph we obtain three different landscapes (see Figure 1.7, 1.8 and 1.9), each for a different typology of synthetic network. Each image show a very interesting result about the dependence of r_∞ on both σ and ϵ. Indeed, the $r_\infty(\sigma, \epsilon)$ is not trivial and it seems to show a metacritical point.
Moreover the landscape reflects what said before, paragraph 1.4.1, about the competition between the coupling and the noise intensity. For each noise intensity, we explore different types of coupling: from values which not cause synchronization, to coupling at the critical threshold and above.

To recover an analytical behaviour of landscapes in Figures 1.7, the idea is to follow an approach similar to Arenas [20] (from eq.(25) to eq.(29)) to have some analytical insights on the critical coupling dependence both on network topology and noise (the last parameter is not included in the approach by Arenas).

Instead of defining a global order parameter, the one used in the fully connected original KM

$$r(t)e^{i\phi(t)} = \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j(t)}, \tag{1.22}$$

we define a local order parameter:

$$r_i(t+dt)e^{i\theta_i(t+dt)} = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}e^{i\theta_j(t)}; \tag{1.23}$$

As we will look for the order parameter in stationary conditions the previous equation reduces to:

$$r_i e^{i\theta_i} = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}e^{i\theta_j}. \tag{1.24}$$

A new global order parameter to measure the macroscopic coherence is readily introduced as:

$$re^{i\phi} = r_i e^{i\theta_i} \tag{1.25}$$

Moreover from eq. (1.25) we find that $r \simeq r_i$ as in the fully synchronized state we have a uniformly rotating one cluster state where $\phi = \theta_1 = \theta_2 = ... = \theta_N$.
If we take into account ansatz in eq. (1.25), we have that at steady state each node is coupled to a local field (r_i) that is equal to the global one (r). Here ϕ is the global phase and at the steady state $\dot{\theta}_1 = \dot{\theta}_2 = ... = \dot{\theta}_N = 0$. Then:

$$re^{i\phi} = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}e^{i\theta_j}. \tag{1.26}$$

By observing that:

$$re^{i(\phi-\theta_i)} = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}e^{i(\theta_j-\theta_i)} = \overbrace{\frac{1}{k_i}\sum_{j=1}^{N} A_{ij}\cos\left(\theta_j-\theta_i\right)}^{r\cos(\phi-\theta_i)} + i\overbrace{\frac{1}{k_i}\sum_{j=1}^{N} A_{ij}\sin\left(\theta_j-\theta_i\right)}^{r\sin(\phi-\theta_i)}. \quad (1.27)$$

Substituting in eq.(1.18), we obtain

$$\dot{\theta}_i = \omega_i - \sigma r\sin\left(\theta_i-\phi\right) + \xi_i(t) \quad (1.28)$$

Close to the (meta)stable state: $\theta_i \simeq \phi \rightarrow \sin(\phi-\theta_i) \simeq \phi-\theta_i$, then:

$$\dot{\theta}_i = \omega_i - \sigma r(\theta_i-\phi) + \xi_i(t) \quad (1.29)$$

By introducing $\psi_i = \theta - \theta_i$ we obtain:

$$\dot{\psi}_i = -\omega_i - \sigma r\psi_i - \xi_i(t) \quad (1.30)$$

As a future step, we can solve the analytical quest of σ_c as function of the network topology and noise by noticing that this is a set of N decoupled Langevin equations.

In the following chapter, we aim to review the thresholding procedure for ad-hoc time series. Indeed, we analyze data from the SKMN dynamics to infer the topology underneath, which we assume to be "unknown".

Chapter 2

Network reconstruction from thresholding

In the previous chapter, we discussed the main features of the stochastic Kuramoto model on a network (SKMN). In the following discussion, we aim to review the thresholding procedure for ad-hoc time series. Indeed, we analyze data from the SKMN dynamics to infer the topology underneath, which we assume to be "unknown".
We explore the results for the Erdos-Renyi model [2]. We try to address the topological reconstruction from the most general viewpoint by considering several values of the parameters involved in the SKMN model. Our goal is to find the best parameters set for which the thresholding reconstruction procedure shows the best reconstruction.

Network science is concerned with understanding and modelling the behaviour of interconnected systems. The analysis of these systems in terms of networks has recently become a relevant interdisciplinary topic, which provides a modern tool to study spatial and temporal data. Synthetic or observational data are a starting point for many developments in the field, especially when dealing with network inference, whose aim is to solve the following inverse problem [24, 25]: from information about the dynamics, reconstruct the network of interactions.
In Figure 2.1 [26], we illustrate the procedure to inference network topology from indirect measurements (incomplete and erroneous networks, time series dynamics and proximity events) coming from some unknown underlying network structure.

Usually, the dynamics is individuated by the vector $\boldsymbol{s}(\boldsymbol{x}(t))$, a N-dimensional multivariate time series of measured observables from an empirical system. Since we cannot always take into account all the details of such system, $\boldsymbol{s}(\boldsymbol{x}(t))$ is a function of $\boldsymbol{x}(t)$, described by eq. (1.7). In many cases, the reconstruction problem relies solely on the $\boldsymbol{s}(\boldsymbol{x}(t))$. However, in the following chapter, we will deal with synthetic $\boldsymbol{x}(t)$ derived by the SKMN to cope with the topology reconstruction problem.

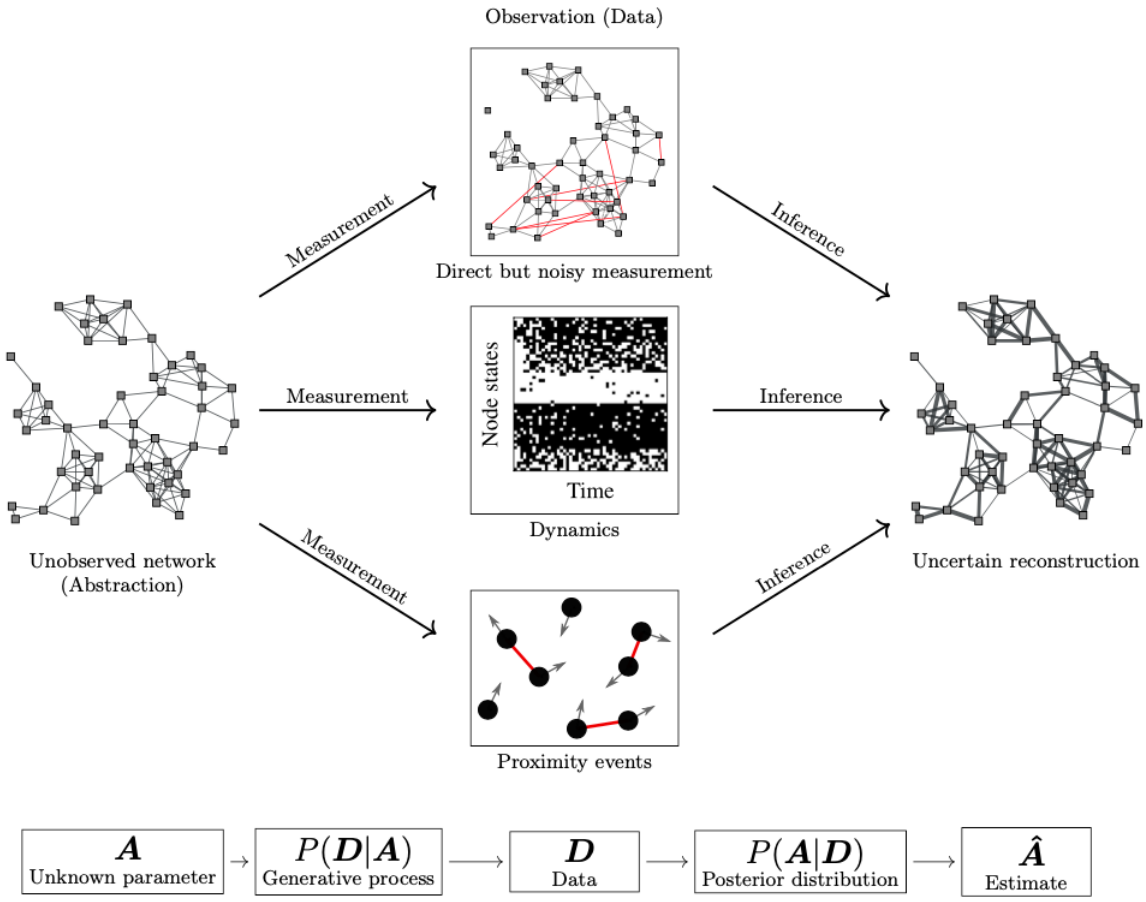


Figure 2.1: The figure from [26] shows how to infer the network topology from the observational data knowledge. An unknown interconnected system, an network of interactions A, gives us a result some kind of observational data D. In this work we deal with the collective noisy dynamics (central picture). Inference reconstruction reminds us to recognize that the data D is the result of measurement process $P(D|A)$ that is conditioned on the unseen network, but is to some extent unavoidably decoupled from it. To estimate the underlying network, we need to perform an inferential step $P(A|D)$, which needs to include our modelling assumptions about how the network and the data are generated. This results in $\hat{A}$, which will have an uncertainty that reflects the experimental design, accuracy of the measurements and overall feasibility of the particular reconstruction problem.

State of the art in literature provides many different methods to infer the network interaction structure between dynamical units from multivariate time series. Besides, the most widely used procedure is the standard thresholding one. This method consists of the computation of an appropriate statistical similarity measure (SSM) to quantify interdependencies between nodes. We typically measure pairwise correlations or statistical causality between network vertices time series and then apply a criterion to decide whether the measured interaction is significant or not. Therefore, we discard values below a chosen threshold such that an edge is assigned only between units whose interaction is sufficiently strong. In this thesis, we choose as SSM the Pearson correlation coefficient, CC.

In statistics, Pearson's r, also known as the correlation coefficient, is a measure of linear correlation between two sets of data, let us say X and Y. The general definition provides the following formula. We denote Pearson correlation coefficient as $CC_{X,Y}$:

$$CC_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

$CC_{X,Y}$ is essentially a normalized measurement of the covariance, the ratio between the covariance of two variables and the product of their standard deviations

As we will deal with time series $s_i(t)$ data, the previous definition eq.(2.1) must include some temporal parameter. From now on, the correlation coefficient is specified as $CC_{ij}(\tau)$:

$$\begin{aligned} CC_{ij}(\tau) = CC(s_i(t), s_j(t+\tau)) = \frac{\mathrm{cov}[s_i(t)s_j(t+\tau)]}{\sigma_{s_i(t)}\sigma_{s_j(t+\tau)}} = \\ = \frac{\sum_{t=0}^{L-\tau}(s_i(t)-\bar{s_i})(s_j(t+\tau)-\bar{s_j})}{\sqrt{\sum_{t=0}^{L-\tau}(s_i(t)-\bar{s_i}))^2}\sqrt{\sum_{t=0}^{L-\tau}(s_j(t+\tau)-\bar{s_j}))^2}} \end{aligned} \tag{2.2}$$

What differentiates the time series formulation of Pearson's coefficient eq (2.2), from the general one in eq (2.1) is, as anticipated, the temporal parameter.
Indeed, the correlation measure depends on τ: this relevant quantity reflects the time lag between two time series.
For the sake of simplicity, we take $\tau = 0$, because it is computationally easier to calculate. Moreover, our analysis will deal with synchronization dynamics whose time lag τ would be, anyway, near to the zero value. We drop out the τ notation in $CC_{ij}(\tau)$ considering the time-lag as $\tau = 0$, if not differently specified.

2.1 Thresholding recontruction methods

Now, let us introduce the thresholding reconstruction procedure. We denote with $\theta_i(t)$ notation the time series linked to the i-th vertex of a generic N nodes network, G.
Since in the following we will deal with synthetic time series, we consider some statistical ensemble to add information to gain a better reconstruction. Therefore, we analyze the ensemble $\{\{\theta_i(t)\}^{(1)}_{i\in G}, \{\theta_i(t)\}^{(2)}_{i\in G}, ..., \{\theta_i(t)\}^{(N_{it})}_{i\in G}\}$ where the N_{it} ensemble is due to a noise

term we introduce in our toy model dynamics. Here, we propose two methods to reconstruct the network of interactions from information about the dynamics, $\{\theta_i(t)\}_{i=1,...,N}$: let us call them **Before Reconstruction (BR)** and **After Reconstruction (AR)** methods.
Both procedures involve the noise statistical ensemble, where we take into account N_{it} processes, $\{\{\theta_i(t)\}^{(1)}_{i\in G}, \{\theta_i(t)\}^{(2)}_{i\in G}, ..., \{\theta_i(t)\}^{(N_{it})}_{i\in G}\}$.
However, the actual reconstruction, where we discard values below a chosen threshold to recover the interaction matrix[1] A^{*R}_{ij}, takes place at different steps in the two methods.
Once we have the $\{\theta_i(t)\}^{(k)}_{i\in G}$ time series, where $^{(k)}$ individuates the k-th element in the noise ensemble, we first choose a statistical similarity measure, the correlation coefficient matrix (CC) to quantify interdependencies between pairs of time series.

In the BR-Method, for each k in the noise ensemble, we consider all the $\{\theta_i(t)\}^{(k)}$ time series in the network. We compute the correlation matrix $CC^{(k)}_{ij}$ for each pair (i, j) time series. Hence, we obtain N_{it} correlation matrices, we averaged them and use the mean correlation matrix $\overline{CC}_{ij} = \sum_{k=1}^{N_{it}} CC^{(k)}_{ij}/N_{it}$ to obtain the inferred network from thresholding. The BR-Method allows for the reconstruction of a unique network A^{BR}_{ij} starting with the N_{it} processes in ensemble.

Whereas in AR-Method for each $\{\theta_i(t)\}^{(k)}$ time series we compute the correlation matrix $CC^{(k)}_{ij}$. Therefore we reconstruct a graph $A^{(k)}_{ij}$ by thresholding each $CC^{(k)}_{ij}$. From the size N_{it} ensemble $\{A^{\mathrm{AR}}_{ij}\} = \{A^{(k)}_{ij}\}$, we recover the different topological properties that are then averaged.
In the following scheme we summarize the thresholding procedure for both methods:

BR-Method	**AR-Method**
$\overline{CC}_{ij} = \sum_{k=1}^{N_{it}} CC^{(k)}_{ij}/N_{it}$	$\{CC_{ij}\}^{\mathrm{AR}} = \{CC^{(k)}_{ij}\}_{k=1,...,N_{it}}$
$A^{\mathrm{BR}}_{ij} = \Theta(\overline{CC}_{ij} - W)$	$\{A_{ij}\}^{\mathrm{AR}} = \{\Theta(CC^{(k)}_{ij} - W)\}_{k=1,...,N_{it}}$

where Θ is the Heaviside function $\Theta(x) := \mathbb{1}_{x>0}$ and W is the chosen thresholding criterion.

From the adjacency matrix A^{BR}_{ij}, we can recover all the topological properties to compare the unique reconstructed network with the original one. Whereas from $\{A_{ij}\}^{\mathrm{AR}}$ ensemble we derive an ensemble of topological properties to average and compare with the ones of the original network G.

[1] The apex ^{*R} stands for the fact that A^{*R}_{ij} is the reconstructed adjacency matrix instead of the original one linked to the G network.

2.1.1 On the sparsity thresholding condition

Now, let us specify the chosen thresholding criterion W to decide whether the measured interaction CC_{ij} is significant or not.

$$A_{ij}^{*R} = \Theta(CC_{ij} - W) \tag{2.3}$$

The thresholding procedure is chosen such that the network reconstructed has the same number of links of the original network G.

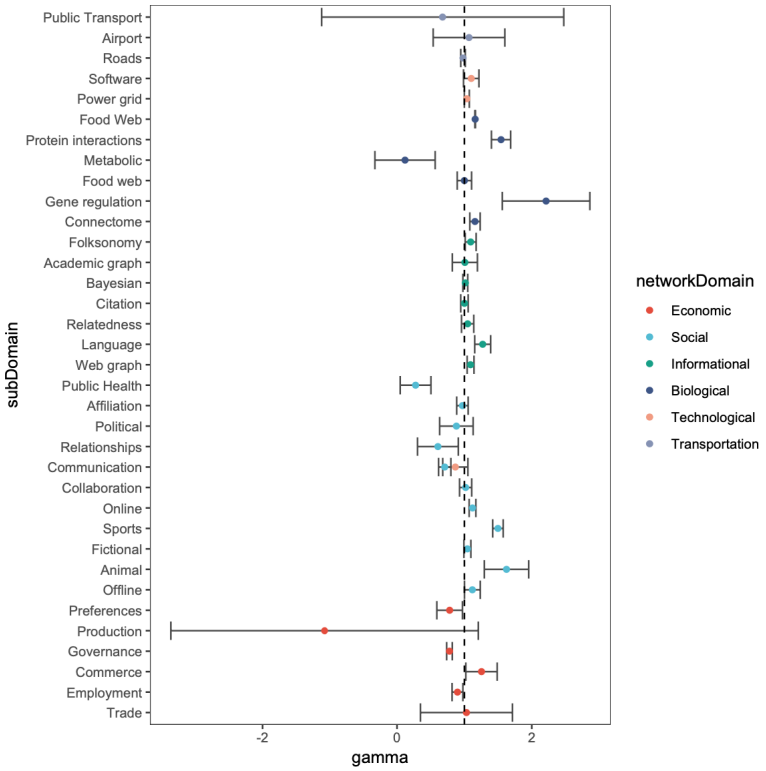


Figure 2.2: Results from [27]. Allometric scaling γ parameter for edges size $|E|$ calculated for different types of networks. The legend on the right shows the area of interest of various networks, which involve several systems from Economics, to Transportation or Biology. It is interesting how γs from different systems are concentrated around a unique value. $\langle\gamma\rangle = 1.04 \pm 0.36$.

Recent studies [28,29] have revealed that many real networks are sparse, i.e the percentage of active interactions is inversely proportional to the system size. Sparsity condition is individuated by the fraction of edges in a network $|E|$ respect to the all possible links, which are $\binom{N}{2}$, if N is the number of nodes. Then the sparsity (density) ρ is:

$$\rho = \frac{2|E|}{N(N-1)} \tag{2.4}$$

The analysis of about 500 various types of networks (Figure 2.2) confirms literature's results on sparsity and computes an average behaviour for the edge's number.

$$|E| \sim N^{\gamma} \qquad \langle\gamma\rangle = 1.04 \pm 0.36 \tag{2.5}$$

The result of a unique γ is very interesting since it implies that many real-systems follow some optimization law to assess the connection among units, which obviously reflects in systems' stability, explorability, and efficiency.
Though the outcome is worthy of further analysis, we use it to threshold the network to impose the realistic sparsity ansatz. Therefore as the original synthetic network is already sparse, we want the reconstructed network to have the same number of links.

2.1.2 Time interval subdivision

Moreover, whether the thresholding procedure analysis deals with the BR-Method or with the AR-Method, we both consider the whole time series $\{\theta_i(t)\}_{i=1,\dots,N}$, or we divide them into subsequential intervals to further investigate the dynamics at different times [30].
The following Figure 2.3 makes our procedure clearer.

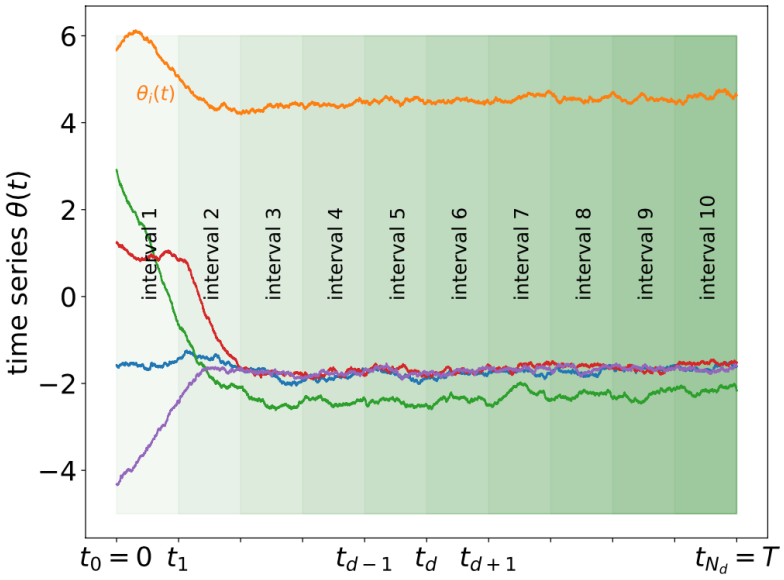


Figure 2.3: Time interval subdivision.

As shown in Figure 2.3 our time series are divided into subsequential time intervals, *interval d*, to infer the topology network by use of both BR or AR thresholding method. If $\theta_i(t)$ goes from 0 to T, we divide $[0, T]$ interval into N_d sub-intervals: then each *interval d* is individuated by $[t_{d-1}, t_d]$, where $d = 1, \dots, N_d$ ($t_0 = 0$ and $t_{N_d} = T$).
Indeed, for each *interval d* we use the "cut" time series $\theta_i^{\text{cut}}(t)$ with $t \in [t_{d-1}, t_d]$. Therefore, we find the correlation matrix CC_{ij} from the time series (i, j) pairs restricted to the $t \in [t_{d-1}, t_d]$ interval and proceed to network inference by thresholding as usual. The time

interval subdivision procedure allows to individuate different time region of the collective dynamics. If we refer to the SKMN synchronization problem we will deal with, we find that for *interval* d with $d < 3$ there is no synchronization since phases are random; when $d \in [3-5]$ dynamics starts to synchronize due to coupling σ; finally for $d > 5$ dynamics reaches equilibrium. That is a consequence of the time integration T we set in Table 1.1. Then, the analysis of network inference restricted to *interval* d allows us to find an optimal dynamics time region for reconstruction.

2.2 Assesing the quality of reconstructed networks

In the previous pages we discuss the thresholding reconstruction methods (BR and AR), the statistical similarity measure (CC) and the thresholding criterion (sparsity ansatz). Now, we can proceed analyzing data from the stochastic Kuramoto model on a network (SKMN) to infer the topology underneath, which we assume to be "unknown". We explore the results for the Erdos-Renyi model [2].

The synthetic Erdos-Renyi case

We here report the equation of the SKMN dynamics:

$$\dot{\theta}_i = \omega_i + \frac{\sigma}{k_i} \sum_{j=1}^{N} A_{ij} \sin(\theta_j - \theta_i) + \xi_i(t) \tag{2.6}$$

In the following section, we deal with the SKMN on the synthetic Erdos-Renyi network, which parameters are specified by the following table

Network	degree $< k >$	spars. ρ	avg path length l	assort. r_{am}	clust. C
Erdos-Renyi	~ 12.65	0.05	2.46	0.034	0.048

Now we want to assess the quality of network reconstruction by thresholding procedure by looking at specific parameters' for the SKMN. To do that, we use the results obtained in previous chapter about the order parameter r_∞ at stationary conditions (Figure 2.4). We notice the σ and ϵ parameters individuate three relevant regions depending on whether or not the synchronization is present and on its strength:

- Before synchronization, the bluish area;
- At the onset of synchronization, the green area;
- When synchronization in the system is strong, yellow areas.

Since every colour in Figure 2.4 is linked to a different dynamical behaviour of our system of N oscillators, we fix several pairs of (σ,ϵ) parameters to set the collective dynamics in such meaningful regions.
Then, for each (σ,ϵ) pair, we analyze the $\{\{\theta_i(t)\}^{(1)}_{i\in G}, \{\theta_i(t)\}^{(2)}_{i\in G}, ..., \{\theta_i(t)\}^{(N_{it})}_{i\in G}\}$ time series

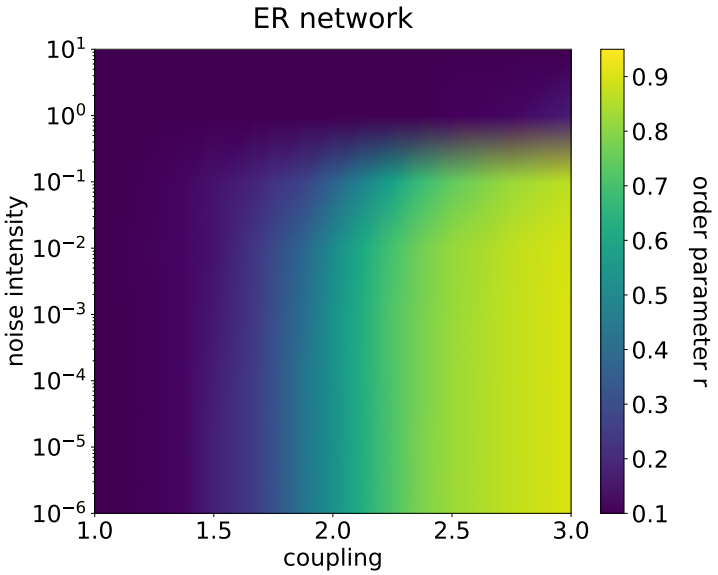


Figure 2.4: Landscape representing the order parameter r_∞ at stationary conditions, by varying the two main parameters of SKMN dynamics, that are noise intensity ϵ and coupling strength σ. The network underneath the model is the Erdos-Renyi specified by the parameters in the above table.

ensemble to infer the adjacency matrix A^{*R} [2].
In the following the thresholding procedure analysis will use both the full time series, or the 'cut' time series by dividing $\theta_i(t)$ in 10 time intervals as discussed in section 2.1.2.

Before assessing the quality of topology reconstruction, we must decide whether we prefer the BR or AR method described in section 2.2.
It is important to remind we use the sparsity ansatz we discussed in section 2.1.1 to include only CC_{ij} highlighting a sufficiently strong interaction. Therefore, the W thresholding criterion would change depending on (σ,ϵ) pair choice.
At first, we discuss the change in the threshold by varying (σ,ϵ). Secondly, we discuss differences in the Frobenius norm, then the fraction of correctly inferred links (true positive ratio, TPR) and we conclude with a comparison of the so-found degree distribution by use of the Jensen-Shannon distance. All these properties should give us an idea of the network reconstruction quality.

In each of the following figures, the * points represent the inferred network property, when the reconstruction method, being it BR or being it AR, involves the analysis of the full time series coming from the SKMN in eq. (2.6); whereas the solid line with ° points shows the same measures when dealing with the 'cut' time series in one of the ten intervals in which we divide our dynamics. To consider growing time regions for Kuramoto dynamics

[2] $*R$ stands for BR or AR methods

has the purpose not only to evidence the quality of reconstruction when the oscillators collective dynamics reaches the steady state, but aims at evaluating the inference method behaviour for dynamics at transient time.

2.2.1 Threshold

Here we present the plots (Figures 2.5, 2.6 and 2.7) for the threshold value W for different coupling strength σ and noise intensity ϵ in SKMN in eq. (2.6).

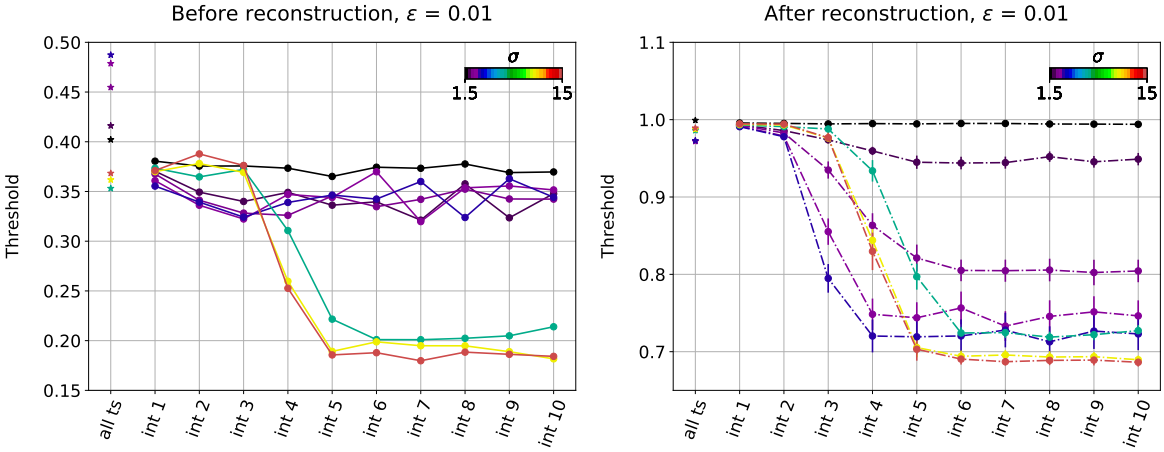


Figure 2.5: On the left: BR; on the right AR. * points represent the W criterion value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

The two methods are different: it is evident that the threshold W is higher in the AF-Method than in BF-Method. Both methods show interesting patterns if we look at *interval d* behaviour. In AR method we observe that when coupling is low there is no difference in analyzing the full time series or the time series restricted to *interval d*. This is because σ strength is not enough for the collective dynamics to become coherent. However, once we reach the critical coupling σ_c we observe that the W value decrease as the dynamics starts synchronizing and the threshold settles to lower values when stationary conditions are reached.
Moreover, in both methods, we observe that an increase of noise intensity implies the threshold to decrease (Figure 2.6 and 2.7). That is due to the before-mentioned action of noise to make the time series less synchronized and, therefore, less pairwise correlated, resulting in setting a lower W.

The trend is quite different in BR method: for high noise intensities the threshold is lower (it makes sense) and the behaviour of W decreases when synchronization stationarity is

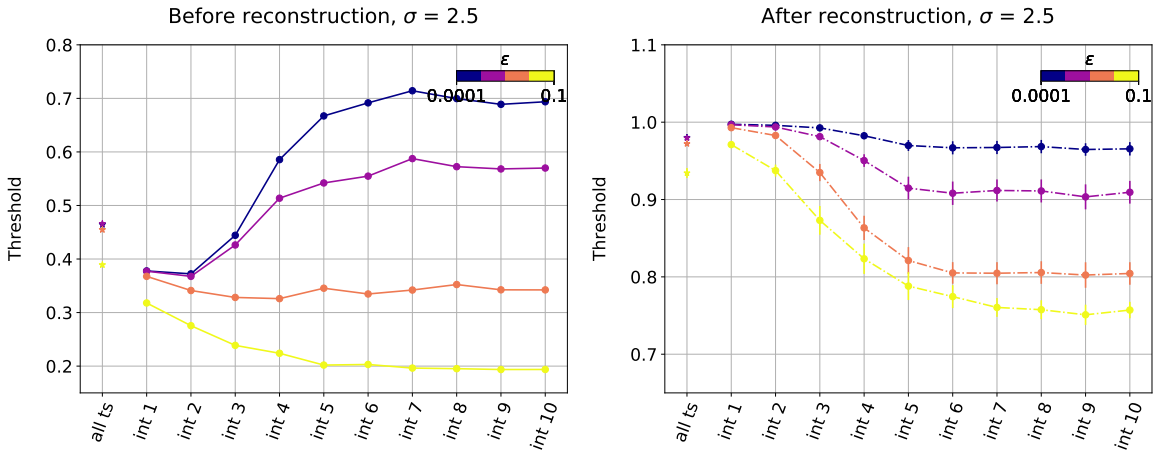


Figure 2.6: Same description of Figure 2.5. However, here we fix coupling $\sigma = 2.5$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

reached. However, for lower ϵ and considering intervals increasing in time the chosen threshold increases.

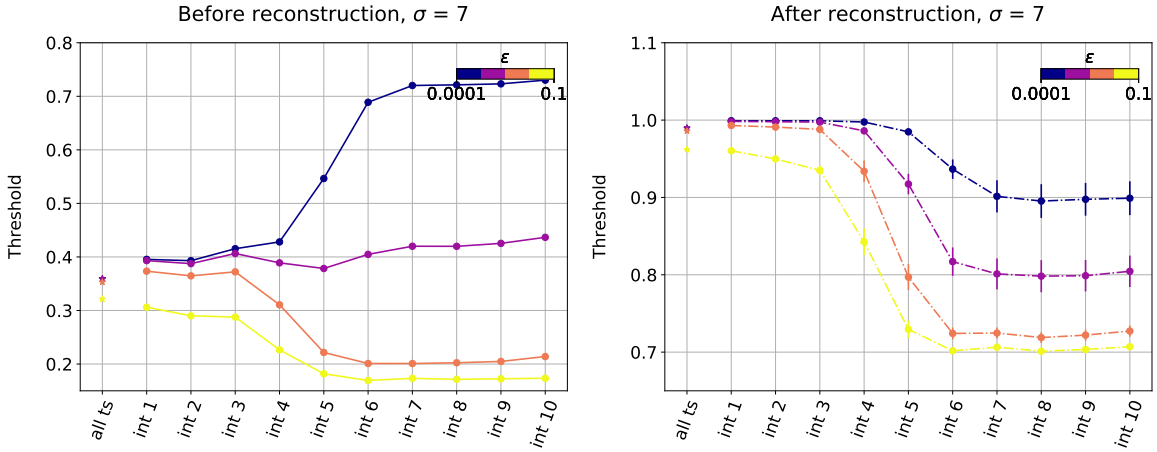


Figure 2.7: Same description of Figure 2.5. However, here we fix coupling $\sigma = 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

2.2.2 Frobenius norm

After discussing the choice of the threshold to cut-off the edges in the reconstructed network we finally deal with some measures to compare the quality of the topological inference. Let us introduce a first measure to compare the reconstructed network with the original

synthetic one. We choose to introduce the Frobenius norm F:

$$F = \frac{||A^{\text{recontructed}} - A^{\text{original}}||}{Normalization} = \frac{\sqrt{\sum_{i>j} |A_{ij}^{\text{recontructed}} - A_{ij}^{\text{original}}|}}{Normalization} \tag{2.7}$$

When considering a general system of N vertices, we can describe two opposite networks: the fully connected A^{fc} and the completely disconnected A^{fd} graph. The adjacency matrices, individuating the two cases, have all ones and all zeros entries, respectively.
Therefore, the normalization of the Frobenius norm is its value in comparing the fully connected and the completely disconnected graph. That is to say:

$$Normalization = ||A^{\text{fc}} - A^{\text{fd}}|| = \sqrt{\sum_{i>j}^{N} |A_{ij}^{\text{fc}} - A_{ij}^{\text{fd}}|} = \sqrt{\sum_{i>j}^{N} |1 - 0|} = \sqrt{\frac{N(N-1)}{2}} \tag{2.8}$$

Frobenius norm from reconstruction is shown in the following Figures 2.8, 2.9 and 2.10. Best reconstruction is achieved lower is the F value.

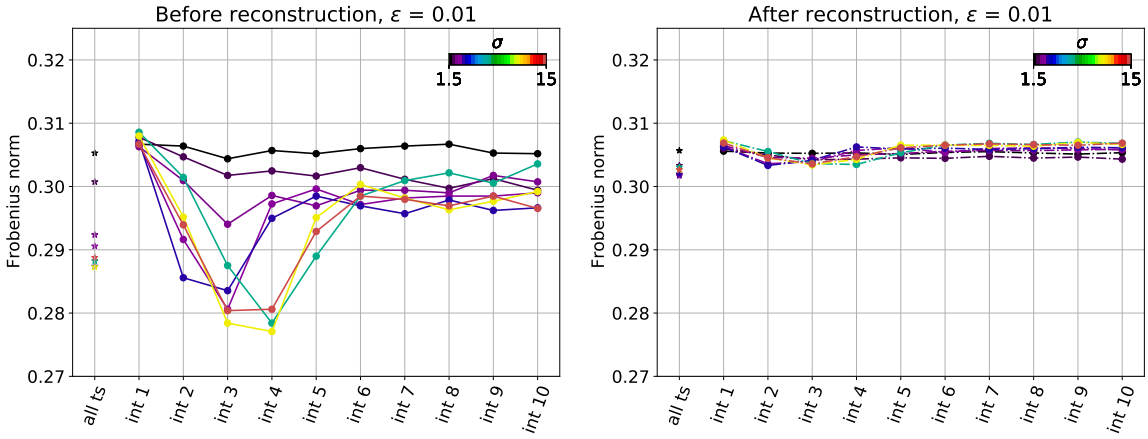


Figure 2.8: On the left: BR; on the right AR. * points represent the Frobenius norm value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

What we can say from the previous pictures is that the two reconstruction method are similar, but BR-Method seems to better reconstruct expecially on the time region at the onset of synchronization (which is around the 4^{th} and the 5^{th} interval). This is even more evident when considering higher coupling and lower noise.
However by looking at Figure 2.8 we notice that after a certain coupling $\sigma = 11$ the trend is inverted and the Frobenius norm increases again.
The best reconstruction provides a Frobenius norm assessing around 0.3%. Even though

the reconstruction is not optimal we can identify some interesting patterns depending on the considered time interval, the coupling and the noise.

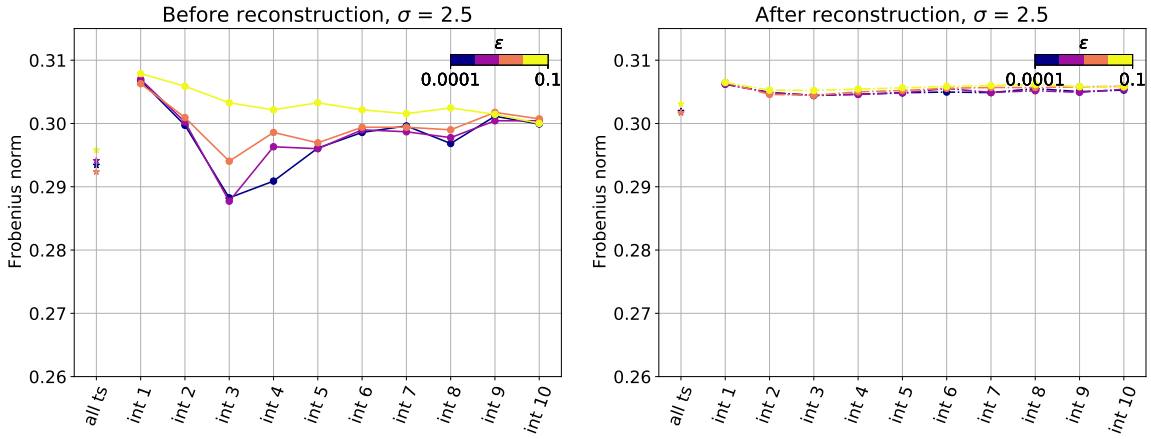


Figure 2.9: Same description of Figure 2.8. However, here we fix coupling $\sigma = 2.5$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

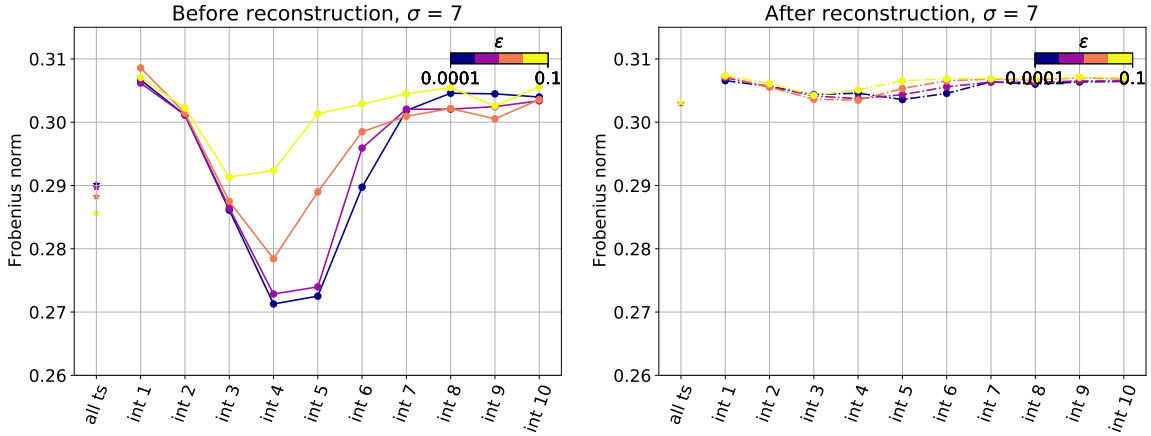


Figure 2.10: Same description of Figure 2.8. However, here we fix coupling $\sigma = 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

2.2.3 True and False positive ratio

Then, we evaluated the performance of the inference method by plotting the fraction of correctly inferred links (true positive ratio, TPR) and the fraction of wrongly inferred links that are not present in the structural network (false positive ratio, FPR).

In Figure 2.11, 2.12 and 2.13, we plot only the true positive ratio. That is because of

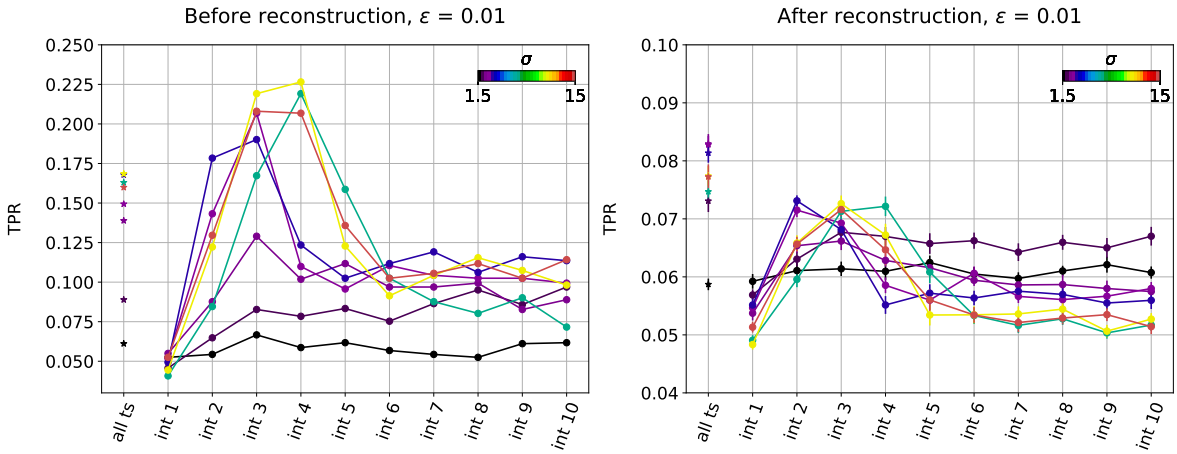


Figure 2.11: On the left: BR; on the right AR. * points represent the TPR value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

the constraint of the inferred network to have the same links as the original synthetic one. Therefore, if we correctly estimate a fraction of edges equal to TPR, the remaining inferred links belong to the FPR fraction, which is exactly 1-TPR.
We notice an evident difference in the BR and AR methods: the Before Reconstruction method systematically provides more TFR (less FPR) than the After inference procedure.

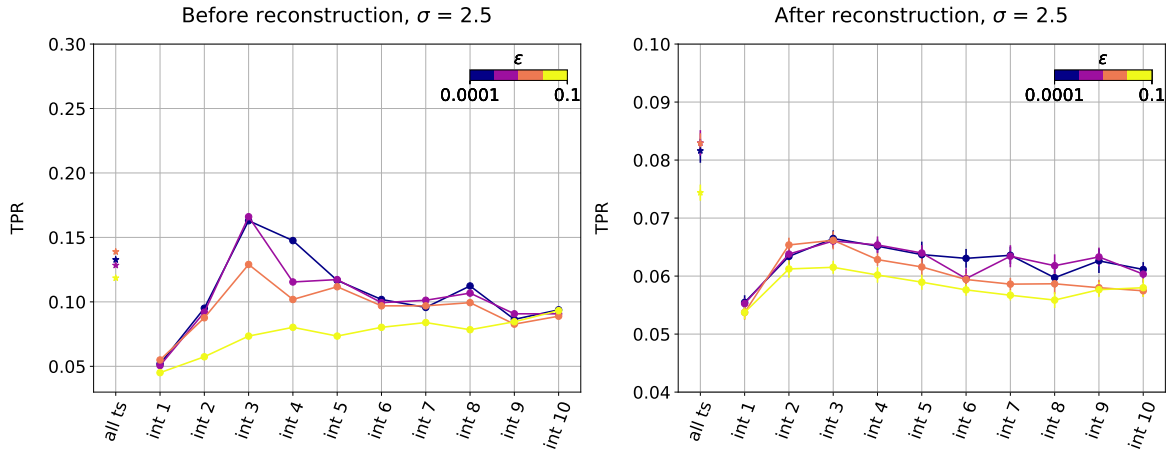


Figure 2.12: Same description of Figure 2.11. However, here we fix coupling $\sigma = 2.5$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

Concerning the BR method, what TPR plots (Figure 2.11A, 2.12A and 2.13A) suggest

is similar to the result found for the Frobenius norm. Indeed the correctly reconstructed edges do not exceed the 0.3% of the correctly inferred edges in the original network.
Instead, the AR procedure (Figure 2.11B, 2.12B and 2.13B) shows worse results than the ones in Frobenius plots since the true positive ratio does not exceed 0.1% of the edges correctly inferred, which suggests a bad network topology reconstruction. That may suggest that the BR procedure better works in reconstruction, however, this improvement is significant only in the middle *interval d*, that is the transient region before synchronization and stationarity condition takes place in the dynamics.

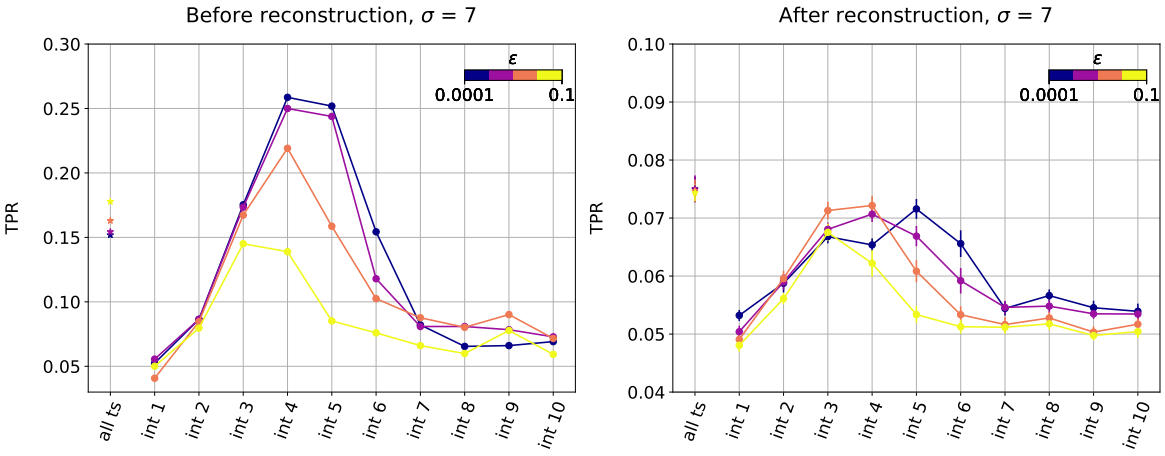


Figure 2.13: Same description of Figure 2.11. However, here we fix coupling $\sigma = 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

If we focus on Figures 2.12 and 2.13, the pattern is even more interesting. In fact, the two figures show the TPR result for σ coupling near the critical σ_c and above this threshold value. By increasing the coupling above σ_c, the true positive ratio tends to be higher, especially in the region at the onset of synchronization. However, from Figure 2.11, we notice that after $\sigma = 11$, the TPR values start to decrease again.
The results for different noise intensity are very similar to the ones for Frobenius norm. Focusing on BR method and the region at the onset of synchronization, lower noise intensities $\epsilon = 0.0001$ and $\epsilon = 0.001$ show both in Figure 2.12 and 2.13 a better score than for $\epsilon = 0.01$ and $\epsilon = 0.1$.

2.2.4 Jensen-Shannon distance

We conclude the comparison between BR and AR methods in thresholding network inference by looking at the Jensen-Shannon distance between the degree distribution for the reconstructed network and the original one.
Let us call the inferred network degree distribution as $P^{*\mathrm{R}}(k)$, since *R=BR,AR are the chosen reconstruction method and k denoted the degree. Instead, $P^{\mathrm{original}}(k)$ individuates the original network degree probability distribution.

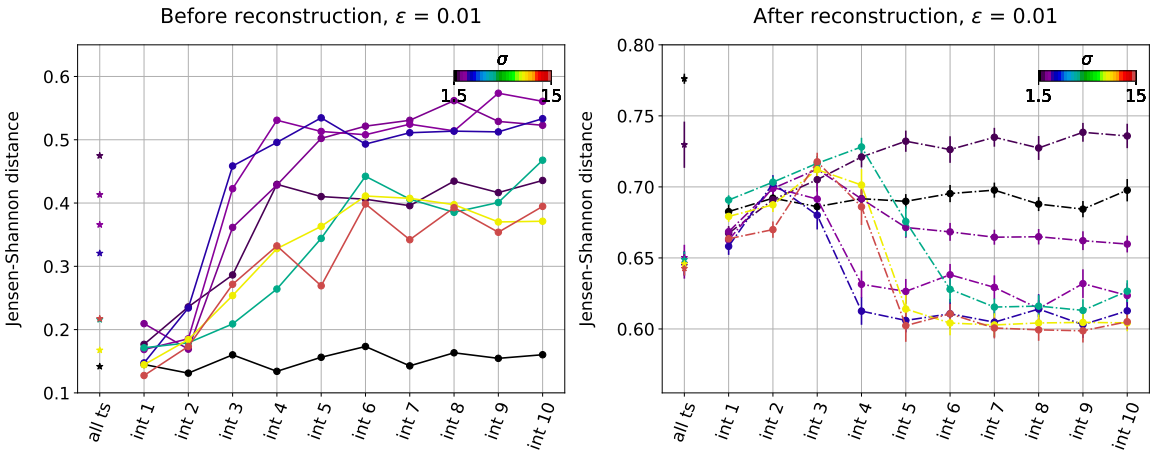


Figure 2.14: On the left: BR; on the right AR. * points represent the degree distribution JSD value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

The Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the Kullback–Leibler divergence where $P^{*\mathrm{R}}(k)$ and $P^{\mathrm{original}}(k)$ are the two distributions to be compared.

$$JSD = \sqrt{\frac{D(P^{*\mathrm{R}}||M) + D(P^{\mathrm{original}}||M)}{2}} \tag{2.9}$$

where

$$M = \frac{P^{*\mathrm{R}} + P^{\mathrm{original}}}{2} \tag{2.10}$$

$D(P^{*\mathrm{R}}||M)$ quantifies the information loss of a probability distribution $M(k)$ from another probability distribution $P(k)$ by means of the Kullback-Leibler divergence:

$$D(P(\cdot)||M(\cdot)) := \sum_{\mathbf{k}} P(\mathbf{k}) \log \frac{P(\mathbf{k})}{M(\mathbf{k})} \tag{2.11}$$

The KL divergence is clearly non-negative and vanishes if and only if the two distributions are equal, but it is not symmetric since it properly quantifies the loss of information when Q is used to approximate P.

We plot the JSD results in the Figure 2.14, 2.15 and 2.16. Comparison between BR and AR method still shows a preference for BR procedure since JSD is sistematically lower. Moreover by looking at Figure 2.15 and 2.16, we notice that higher coupling show a smaller JSD value for the full time series analysis. When considering the time intervals *interval d* before the oscillators reaches stationary state, we notice that for $\sigma = 7$ a lower noise implies an inferred degree distribution more similar to the original one.

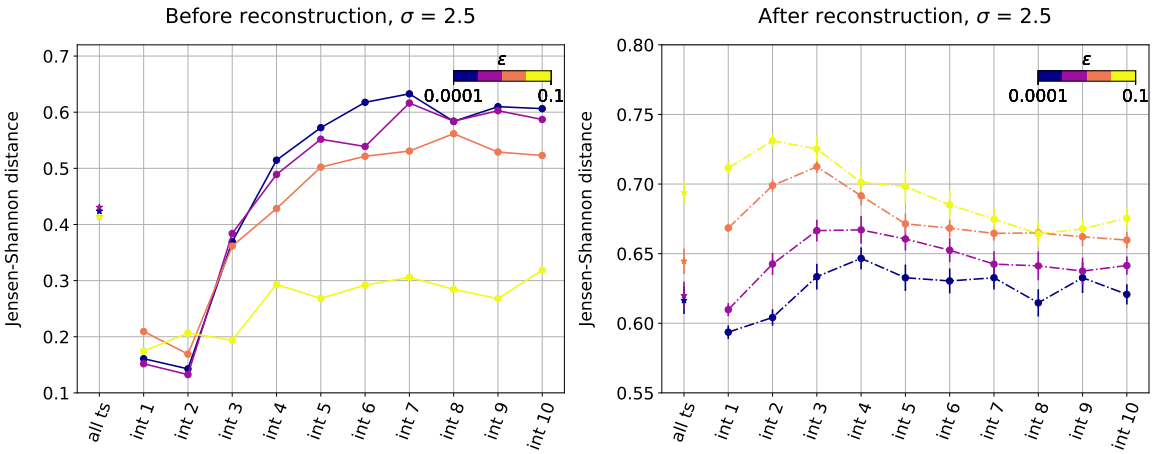


Figure 2.15: Same description of Figure 2.14. However, here we fix coupling $\sigma = 2.5$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

The trend is the opposite for *interval* d with $d > 4$, since the JSD is lower when considering higher noise intensities.

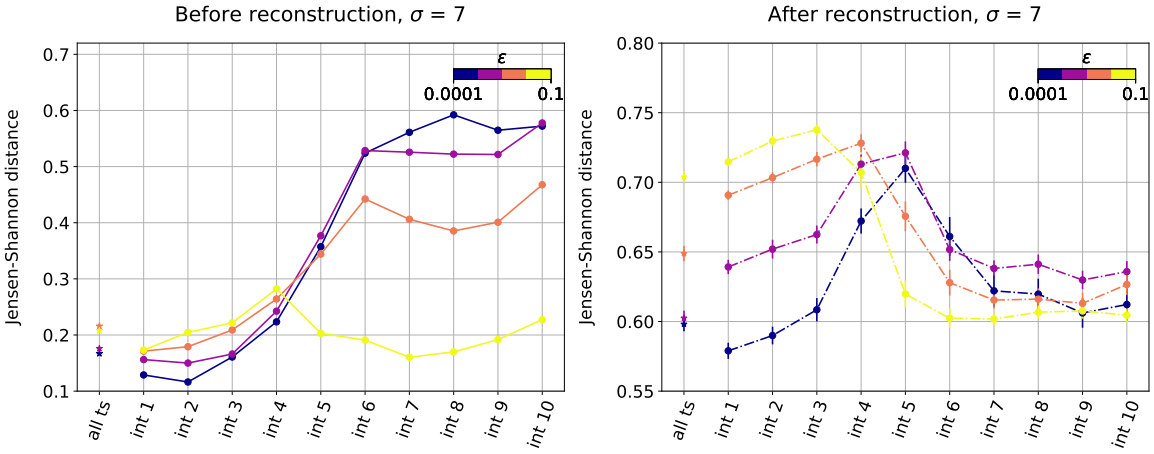


Figure 2.16: Same description of Figure 2.14. However, here we fix coupling $\sigma = 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

We conclude this section by stating that the BR method of reconstruction is the preferable over the AR procedure.

2.3 Inferring macroscopic topological indicators

By knowing the structure of a network, the adjacency matrix A, we can calculate from it some measures that capture particular features of the graph topology.
In the following, we focus on the Before method reconstruction since it appears to be the best one. Though we cannot be exhaustive, we look at some of the network's measures to see if the inference results are still promising for such relevant topological indicators. In particular, we discuss the average path length, the assortativity and clustering.

2.3.1 Average path length

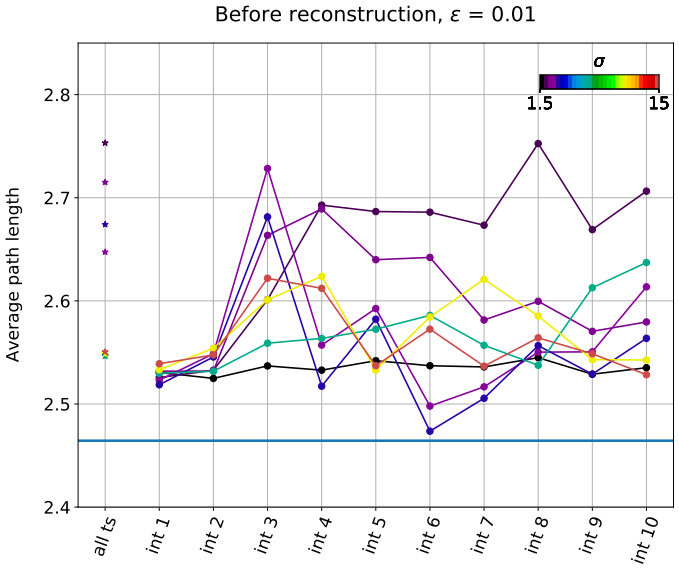


Figure 2.17: BR plot where * points represent the l value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

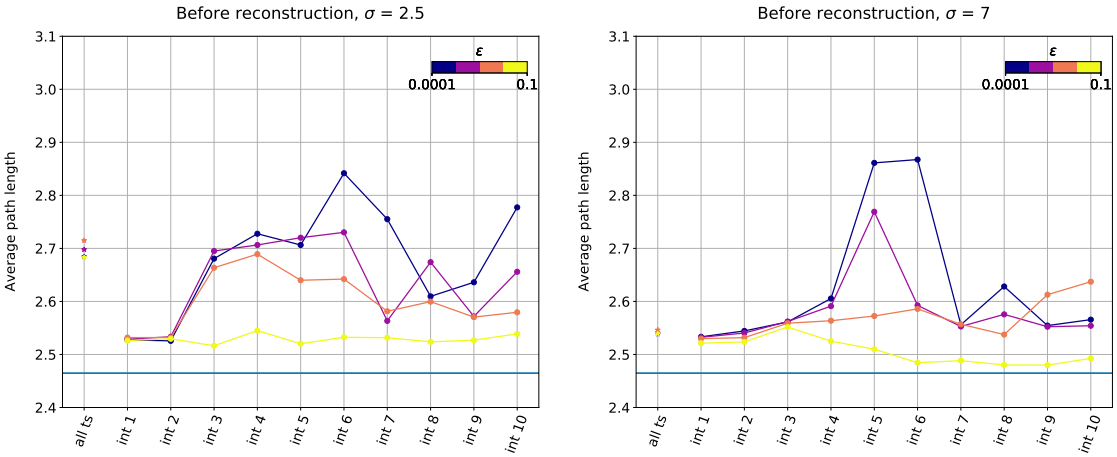
As suggested by its name, average path length gives information on the distance between two nodes in a network, where distance is measured as the minimum number of edges linking the two. Then, to put the definition into a useful equation, we first define the mean distance l_i between node i and all the other nodes in the network.

$$l_i = \frac{1}{n} \sum_j d_{ij} \tag{2.12}$$

where d_{ij} is the number of edges linking i and j node in the shortest path.
Then we average l_i over all vertices in the graph to obtain the mean distance l, which is

the average path length.

$$l = \frac{1}{n^2} \sum_{ij} d_{ij} \tag{2.13}$$



(A) (B)

Figure 2.18: Same description of Figure 2.17. However, here we fix coupling $\sigma = 2.5, 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

The results concerning the average path length are shown in Figure 2.17 and 2.18. We are happy with the results since the l seems to be well recovered both by the full time series analysis and by the network reconstructed at intervals of the SKMN time series. Comparing Figure 2.18A and 2.18B, we notice that the l measure is more similar to the average path length of the original Erdos-Renyi network (light blue orizontal line) when coupling is higher. If higher noise seems to better recover l property, by paying more attention we notice the result is quite similar even for lower noise intensities and especially for time series analyzed in the time regions before stationarity is reached.

2.3.2 Assortative mixing

Assortative mixing or homophily is the tendency of nodes to connect to other vertices with similar characteristics. In the following we choose to measure the assortative mixing by degree. This measure the tendency of nodes to connect other vertices with degrees similar to their own.
Assortative mixing by degree can be quantified in several different ways. We choose the one in eq. (21) in [31] which gives us:

$$r_{am} = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)k_i k_j}{\sum_{ij}(k_i \delta_{ij} - k_i k_j/2m)k_i k_j} \tag{2.14}$$

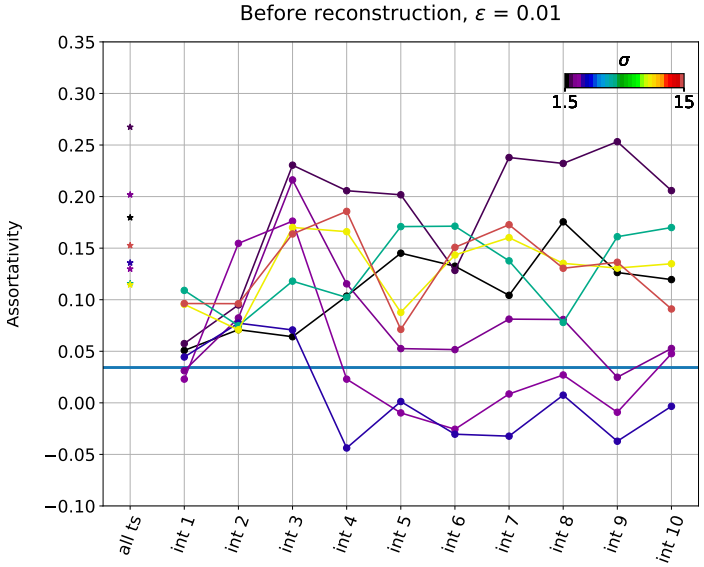
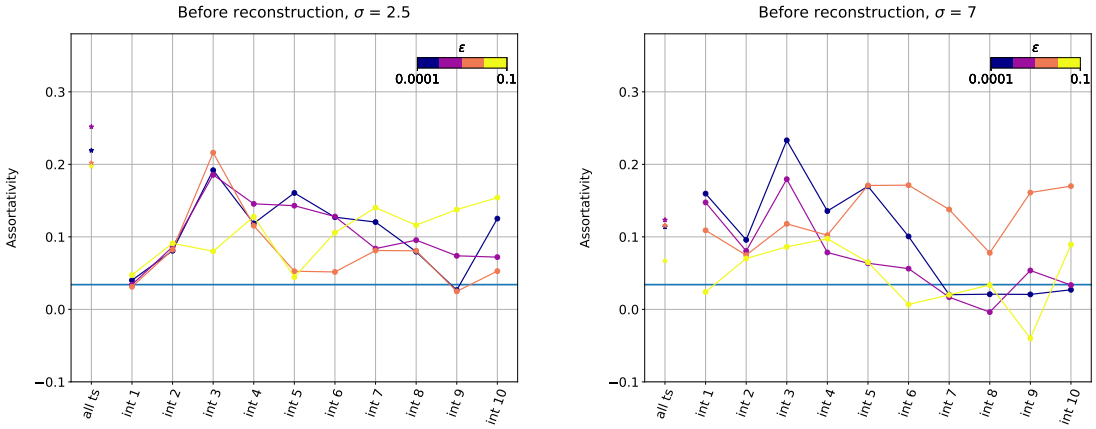


Figure 2.19: BR plot where * points represent the assortativity r_{am} value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

Figure 2.19 shows no significant pattern, since r_{am} measure seems not having a specific behaviour. However, by comparing Figure 2.20A and 2.20B it is a little be easier to notice that for higher coupling the measure is better inferred especially for those time intervals for which stationary state is still not reached.



(A) (B)

Figure 2.20: Same description of Figure 2.19. However, here we fix coupling $\sigma = 2.5, 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

2.3.3 Clustering coefficient

The clustering coefficient C is the average probability that two neighbours of the same node are, in turn, connected. Since the clustering coefficient involves a transitive property, we can define it as the density of triangles in a network. The following plots show the results for the reconstruction compared with the value in the original synthetic network.

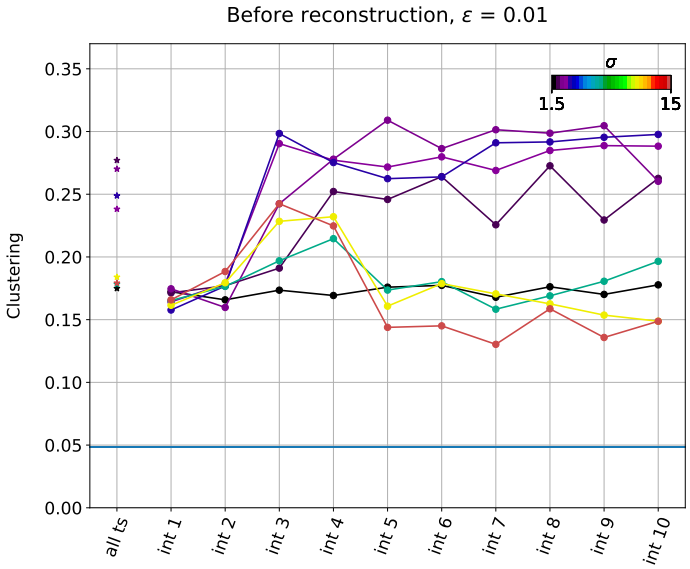


Figure 2.21: BR plot where * points represent the C value when the reconstruction method involves the analysis of the full time series coming from the SKMN in eq. (2.6) (*all ts*); whereas the solid line with ° points shows the same measures when dealing the ten intervals in which we divide our dynamics (*int d*). We fix noise intensity to $\epsilon = 0.01$, while the lines are coloured depending on the coupling strength $\sigma \in [1.5, 2, 2.5, 3, 3.5, 7, 11, 15]$.

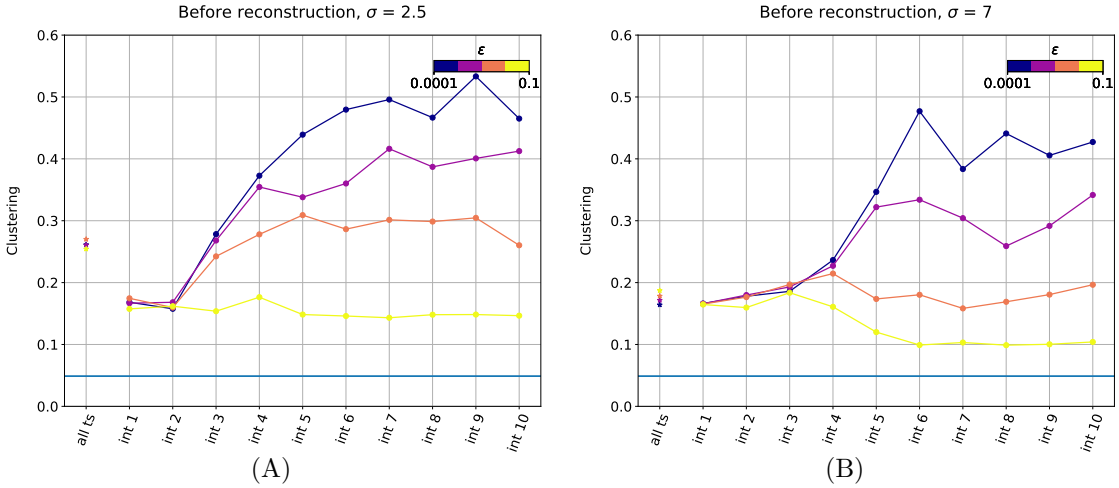


Figure 2.22: Same description of Figure 2.21. However, here we fix coupling $\sigma = 2.5, 7$, while the lines are coloured depending on the noise intensity $\epsilon \in [0.0001, 0.001, 0.01, 0.1]$.

By comparing result in Figure 2.22A and 2.22B, we notice that a higher coupling $\sigma = 7$ better recover the clustering property than for $\sigma = 2.5$.
Moreover, we notice an interesting behaviour depending on the time region we conduct our analysis on. Before and at onset of synchronization the C measure remains near the original network value, however when time series stationarity is reached ($int\ d$, with $d > 4$ in the case with higher coupling), the clustering coefficient increases (when considering lower noise intensities).

2.4 Synthetic networks results from the optimal region

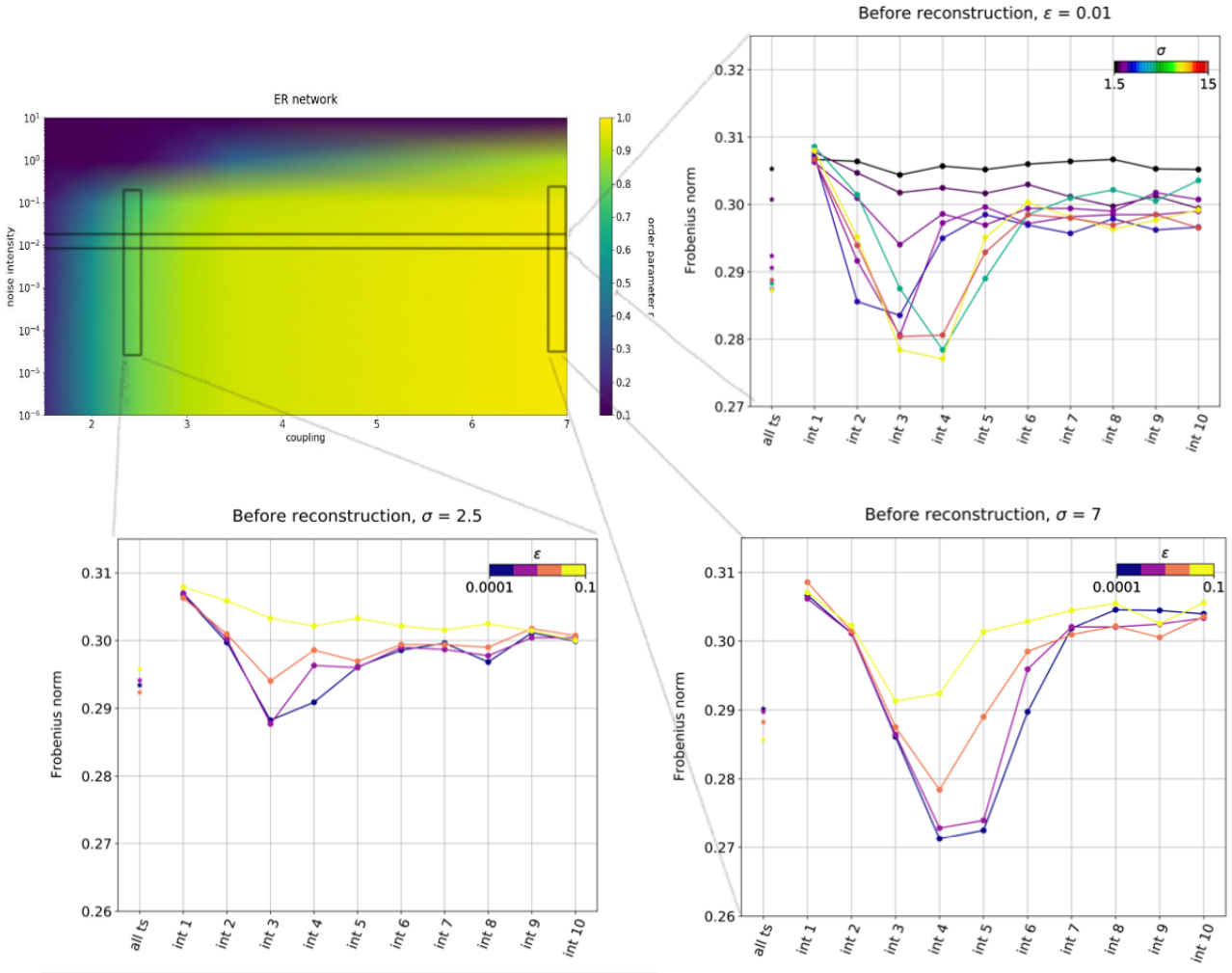


Figure 2.23: This figure summarizes the thresholding reconstruction process for different σ and ϵ parameters on Erdos-Renyi network with $\langle k \rangle \simeq 12$. For each of the three relevant regions individuated we plot the Frobenius norm value.

In this section, we summarise the results learned in the previous paragraphs. We try to address the previous questions on best network inference from the most general viewpoint, considering all parameters involved in the SKMN model.
Indeed, we are not interested in finding an optimal reconstruction; besides, we focus on finding the best parameters in the thresholding reconstruction procedure, which would be relevant to exploring the fuzzy-network framework. From comments on the previous plots, we can derive that for the Erdos-Renyi synthetic network under the SKNM, the optimal parameters are low noise intensity and coupling above the critical σ_c. This result can be summarized by reporting the Frobenius norm (Figure 2.23), which better shows this pattern of optimal reconstruction. Moreover, we notice that the time interval subdivision is useful to better recover the network topology. By including the transient time, the onset of synchronization and the stationarity part in the time series analysis, we can find that a better reconstruction is achieved just before the collective dynamics have reached synchronization.

As we will see at the beginning of the next chapter, thresholding introduces some arbitrarity in the reconstruction process since this method relies on the choice of a thresholding criterion. Consequently, we witness the inference of complex features even when no complex structure is present. Therefore, in next chapter, we introduce and define a new approach to network inference, the fuzzy network model to overcome the problem and gain a meaningful reconstruction method.

Chapter 3

Network reconstruction from fuzzy network modelling

In the previous chapter, we explore the limits of the thresholding reconstruction of network topology. Moreover, we assess the optimal parameters and time region to infer the network underneath the SKMN dynamics. In this chapter, we discuss why the thresholding procedure is not always reliable by looking at a simple but significant example. Therefore, we introduce and define a new approach to network inference, the fuzzy network model. In the last section, we calculate the main information of this method, the Π matrix: an adjacency matrix analogous.
But first, let us introduce some general knowledge about null hypothesis and surrogate testing that will be of use in the present chapter.

3.1 Surrogate data testing

Since we use the analysis of pair of nodes (i, j) time series as a proxy for the structural connectivity of the system, the very first question is whether the dynamics of the N dimensional system results in purely random noise, or exhibits some deterministic features. Surrogate methods can provide some relevant answers to this question. However, it is important to estimate surrogate data by correctly assuming the underlying null hypothesis, not leading to false rejection.

How many surrogates do we create?

Typically, to test the deviation of original data from the distribution of the surrogates, we generate several realizations of the chosen model. But how many surrogates do we need for a minimal significance requirement? For a residual probability α of false rejection, corresponding to a level of significance $(1 - \alpha) \times 100\%$, we generate $M = 2K/\alpha - 1$ surrogates sequences, where K is a positive integer for a two-sided test.

Thus, including the data itself, we have $2K/\alpha$ sets. It is quite an obvious result that larger values of K give a more sensitive test than $K = 1$. For a minimal significance requirement of 95% and $K = 1$, we thus need at least 39 surrogate time series for a two-sided test.

In the following, we discuss various methods to estimate the null hypothesis model.

3.1.1 Random permutation surrogates

Random permutation (RP) surrogates are central to checking whether there is any temporal structure in the data or whether analyzed time series are just uncorrelated noise. The algorithm consists of randomly re-shuffling the original time series to obtain M surrogates. The synthetic RPs possess the same mean and variance as the original signal, but any temporal structure is destroyed.
Any correlations present in the data will lead to the null hypothesis rejection as data possess some temporal structure. Besides this statement, we can not conclude anything further about the nature of this structure without further tests. At the same time, when testing against noise in empirical systems, we can regularly face a low signal-to-noise ratio: a failure to reject does not necessarily imply that the data are just noise. In the Kuramoto model and the majority of cases, data can be verified as not being uncorrelated noise by visual inspection.

3.1.2 Nonlinear testing: FT, AAFT and IAAFT surrogates

Very few real-world time series, suspected of showing non-linearity, follow a Gaussian single-time distribution. It follows that non-linearity testing has its simplest signature from non-Gaussianity [1]

Thus, a possible null hypothesis is the one where surrogate data are generated by a stationary Gaussian linear stochastic process. The most general univariate linear process is given by:

$$s_t = \sum_{i=1}^{M} a_i s_{t-1} + \sum_{i=0}^{T} b_i \eta_{t-1} \tag{3.1}$$

where $\{\eta_t\}$ are Gaussian uncorrelated random increments.
Since we want to test against a whole class of processes, without specifing one particular linear process only (a_i and b_i in eq. (3.1)), we are dealing with the so-called *composite null hypothesis*.

Among various approaches to testing for a composite null hypothesis, the most attractive seem to be the ones creating constrained realisations [34]. Instead of thinking to s_t underlying model equations eq. (3.1), this means we look at surrogate datas by creating sequences with the same first and second order properties of the original data (mean, variance and auto-covariance function), but which are otherwise random.

[1] We pay attention to the fact that data may be distorted in the measurement process, leading to non-Gaussian distribution even if the underneath dynamic is linear.

Indeed, the *Fourier transform* (FT) surrogates procedure is a phase randomization process where we preserve linear behaviour, i.e the power spectrum/autocorrelation, but destroy any non-linearity. Constrained realizations of this null hypothesis would require the generation of random sequences with the same power spectrum (fully specifying the linear process) and the same single-time distribution (specifying the effect of the measurement function) as the observed original data. The algorithm requires the calculation of $F_k(s_t)$ the discrete Fourier transform of the original signal, let call it $s(t) = s_t$ where time t is discrete with T time-steps.

$$|F_k(s_t)|^2 = \left| \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} s_t \exp\left(\frac{2\pi i k t}{T}\right) \right|^2, \tag{3.2}$$

i.e. the periodogram estimator of the power spectrum. Then, we generate a vector of random phases α_k ($0 \leq \alpha_k < 2\pi$ are independent uniform random numbers).
We then create the new phase randomized vector $F_k(\bar{s}_t)$, by multiplying the Fourier transform of the data by random phases at previous step, $|F_k(\bar{s}_t)| = e^{i\alpha_k}|F_k(s_t)|$.
Finally, we recover surrogate time series $\{\bar{s}_t\}_{t=0,\dots,T}$ by transforming back to the time domain:

$$\bar{s}_t = \frac{1}{\sqrt{T}} \sum_{k=0}^{T-1} |F_k(r_t)| \exp\Big(-\frac{2\pi i k t}{T}\Big), \tag{3.3}$$

The FT null hypothesis discussed so far (Gaussian linear processes) is not what we want to test against in most realistic situations. In particular, the most obvious deviation from the Gaussian linear process is usually that the data do not follow a Gaussian single-time probability distribution. It is frequent that empirical data, obtained by measuring intervals between events, can lead to false rejections of the null hypothesis. Attention must be paid to ensure that a conclusion of non linearity does not arise in the system during measurement rather than resulting from the underlying dynamics.
However, there is a simple generalisation of the FT null hypothesis that explains deviations from the normal distribution by the action of an an invertible time-independent instantaneous (i.e. no time delays) measurement function h:

$$s_t = h(x_t), \qquad x_t = \sum_{i=1}^{M} a_i x_{t-1} + \sum_{i=0}^{T} b_i \eta_{t-1} \tag{3.4}$$

where the restriction that h is invertible(i.e. $x_t = h^{-1}(s_t)$), time-independent (stationary/autonomous) and instantaneous is very severe and essential.
The *Amplitude Adjusted Fourier Transform* (AAFT) method proposed in [35] attempts to invert the measurement function h by rescaling the data time series to a Gaussian distribution. Then the Fourier phases are randomized and the rescaling is inverted.
As discussed in [36], this procedure is biased toward a flatter spectrum. In the same reference, a scheme is introduced that removes this bias by iteratively adjusting the distribution and the spectrum of the surrogates, the iterative amplitude adjusted Fourier transform

(IAAFT). Alternatingly, the surrogates are rescaled to the exact values taken by the data and then the Fourier transform is brought to the exact amplitudes obtained from the data. The discrepancy between both steps either converges to zero with the number of iterations or to a finite inaccuracy, which decreases with the length of the time series, to achieve a closer match between both the distribution and the power spectrum in the original data and the surrogates.

3.2 Random network model

Let us start with a simple model where we consider an N-dimensional vector, $\mathbf{s} = [s_1, s_2, \ldots, s_i, \ldots, s_N]$. For each vector component i, we build an artificial L-dimensional discrete time series, whose components are random numbers extracted from a generic distribution $\mathcal{D}$.
Therefore for each vector entry i, we have: $s_i(t) = [s_i(0), ..., s_i(L)]$ where $s_i(m) \in \mathcal{D}$ with $m \in (0, 1, ..., L)$. We obtain the following (N, L)-matrix:

$$\overbrace{\left.\begin{bmatrix} s_1(0) & s_1(1) & \ldots & s_1(L-1) \\ s_2(0) & s_2(1) & \ldots & s_2(L-1) \\ \vdots & \vdots & \ddots & \vdots \\ s_N(0) & s_N(1) & \ldots & s_N(L-1) \end{bmatrix}\right\} N \text{ rows}}^{L \text{ columns}} \tag{3.5}$$

By just using the information about time series, we want to learn the topology between every i-th component, namely the nodes of a hypothetic and unknown N-dimensional network.
Even the least attentive reader would notice the evident anomaly of the task, as there should not be any proper link among the N components, because we are looking at random time series.
Starting with N time series, we want to end up with a unique adjacency matrix individuating the associated network.
In order to make the procedure clear, we describe each step of the Simple Random Network model (SRNM) in the following paragraphs.

- The first one 3.2.1 deals with the calculation of a similarity measure to compute the correlation level between any pair of time series s_i;
- The second 3.2.2 provides Z-scores calculation from the previously calculated similarity measures by comparing them to some null hypothesis;
- At last 3.2.3, we decide on some threshold to cut off lower Z-scores, in order to point out the paired nodes from the most highly correlated time series.

3.2.1 Measuring statistical similarity between time series

We first compute some statistical similarity measure, SSM, between each pair of nodes (i, j) in the N-dimensional system as a proxy for the structural connectivity of the system. In general we can apply any statistical descriptor, here we list some of them: Pearson correlation coefficient (CC), Spearman's rank correlation (SC), or the spectral coherence (SpeCoh) [37]; an information-theoretic tool as mutual information (MI) [38]; or a state-space reconstruction tool, convergent cross mapping (CCM) [39].
In this thesis, we choose as SSM the Pearson correlation coefficient. In the following section 3.4, our choice will be soon clear.

As we are dealing with time series $s_i(t)$ data, the correlation coefficient is specified as $CC_{ij}(\tau)$:

$$\begin{aligned} CC_{ij}(\tau) = CC(s_i(t), s_j(t+\tau)) = \frac{\mathrm{cov}[s_i(t)s_j(t+\tau)]}{\sigma_{s_i(t)}\sigma_{s_j(t+\tau)}} = \\ = \frac{\sum_{t=0}^{L-\tau}(s_i(t) - \bar{s_i})(s_j(t+\tau) - \bar{s_j})}{\sqrt{\sum_{t=0}^{L-\tau}(s_i(t) - \bar{s_i}))^2}\sqrt{\sum_{t=0}^{L-\tau}(s_j(t+\tau) - \bar{s_j}))^2}} \end{aligned} \tag{3.6}$$

The correlation measure depends on τ: this relevant quantity reflects the time lag between two time series.
For the sake of simplicity, we take $\tau = 0$, but the results of SRNM (eq.3.5) would be equal, independent of the τ choice. We drop out the notation considering the time-lag as $\tau = 0$ if not differently specified.
In the following, the apex label $^{(obs)}$ stands for the fact that the matrix $CC_{ij}^{(obs)}(\tau)$ is a unique realization that univocally individuates the system 3.5 in analysis; the observed system indeed.

3.2.2 Null hypothesis testing

The second step is to build a null model to compare with our specific observable, the present realization of random numbers for each component of the N vector.
Such a model should consider an adequate null hypothesis $H_{ij}^{(null)}$: that is the lack of relationships between the nodes i and j.
The simplest way to calculate such model distribution is to re-shuffle the observed time course at each site, to destroy any temporal correlation [40].

Algorithm: null hypothesis distribution

1: for k in $[1, 2, \ldots, M]$:
2: for i in range(N):
3: $s_i^{(null)}(t) =$ re-shuffle$(s_i(t))$
4: $\forall k \quad CC_{ij}^{(null)} = CC(s_i^{(null)}(t), s_j^{(null)}(t))$

The previous table summarizes the algorithm to create RP surrogates from the original time series. We consider a significant statistic M to computationally fit the Gaussian distribution whose mean and standard deviation $(\overline{CC}_{ij}^{(null)}, \sigma_{ij}^{(null)})$ are involved in the Z-score calculation eq. (3.7).

$$Z_{ij} = \frac{CC_{ij}^{(obs)} - \overline{CC}_{ij}^{(null)}}{\sigma_{ij}^{(null)}} \tag{3.7}$$

$$p_{ij} = 1 - \mathrm{erf}\left(\frac{Z_{ij}}{\sqrt{2}}\right) \tag{3.8}$$

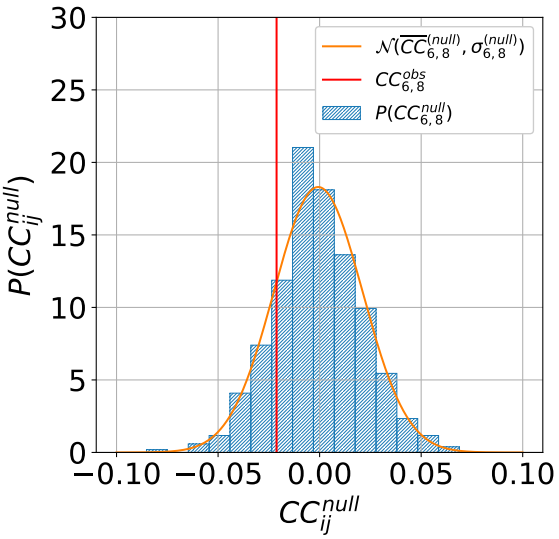
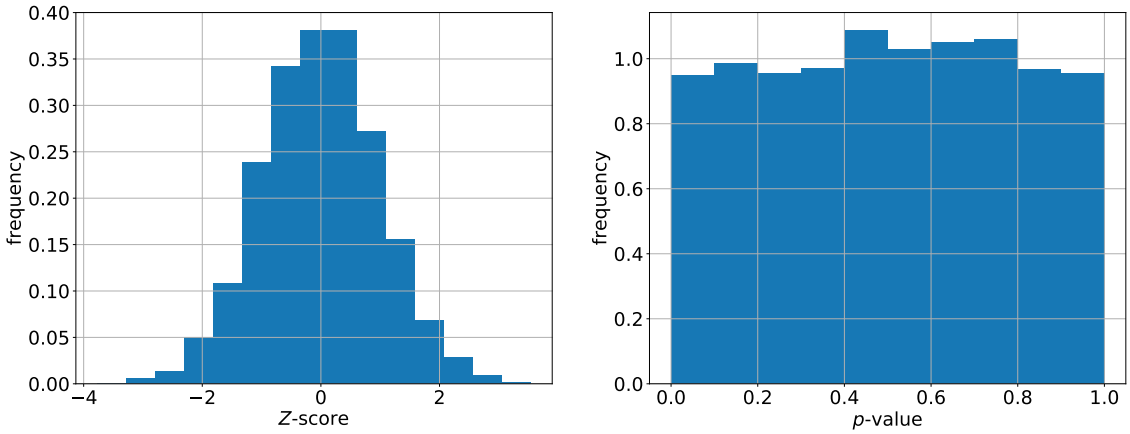


Figure 3.1: Null hypothesis distribution for link $(i, j) = (6, 8)$ The orange line is the gaussian distribution with $\mu = \overline{CC}_{ij}^{(null)} = -0.0007$ and $\sigma = \sigma_{ij}^{(null)} = 0.0218$. The red vertical line $CC_{ij}^{(obs)} = -0.0211$ individuates our observed case for the specific edge.

We then compare our observable towards the null model distribution and recover the linked Z-score.

We highlight at this point the parameters involved in the SRNM model: the number of nodes, N=100, the time series length, L=2000, the chosen distribution is a Bernoulli $\mathcal{D} = \mathcal{B}(p) = \begin{cases} 0 & p \\ 1 & 1-p \end{cases}$, with $p = 0.5$, and the size of the null ensemble is M=500.
Figure 3.1 highlights the null model comparison for edge $(i, j) = (6, 8)$: the plot show we can not reject the lack of connectivity between the two chosen vertices.

The following plot shows the distribution of every p-value related to each pair of nodes in our artificial network.
From both panels of Figures 3.2, we recover an important piece of information. The p-value distribution is uniform (Figure 3.2B). This is the result of the Gaussian distribution of correlation values between any two time series in the SRNM. Precisely Z-scores are Gaussian distributed indeed (Figure 3.2A).



(A) (B)

Figure 3.2: (A) Z-score distribution for every possible edge $\binom{100}{2}$ in the SRNM, where $N = 100$, $L = 2000$ and the distribution D from which to extract the time series is a Bernoulli $\mathcal{B}(p)$, with $p = 0.5$, (B) p-value distribution for the SRNM.

3.2.3 Network obtained from thresholding

Once we obtained in paragraph 3.2.2 the $p - values$ corresponding to each possible link in the $N = 100$ network, we impose a significance level of $p - value < \alpha$ in rejecting the null hypothesis.

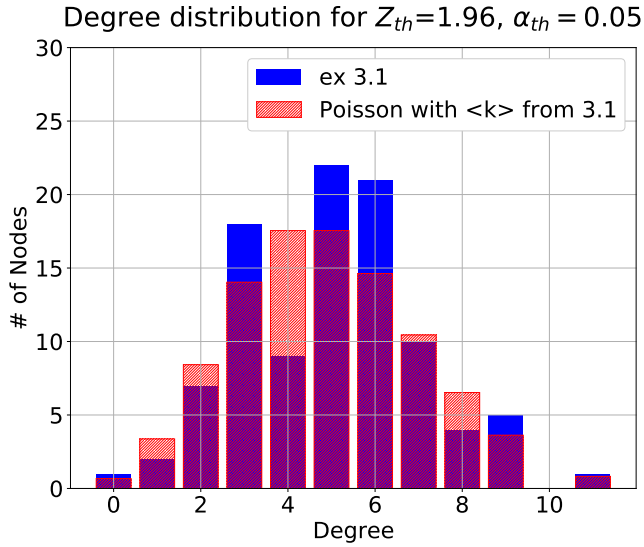


Figure 3.3: In blue we plot the network degree distribution derived from thresholding our problem with significance level $\alpha = 0.05$. Red histogram shows a comparison of a random network Poisson distribution with link existence probability p related to the significance level cut-off.

The analyzed case 3.5 shows that, when choosing such a threshold, the cut-off provides an Erdos-Renyi network whose link probability p is related to the chosen significance level α, but is actually the result of spurious correlation. Therefore, the so-found network itself is a spurious network correlation, the meaning of this result will be better explained in next section 3.2.3.
In Figure 3.3, we show the result of SRNM for a significance level $\alpha_{th} = 0.05$.

Why is multiple testing a problem?

This simple model is a reminder to pay attention to the thresholding procedure in inferring a network from any time series. Indeed, the previous SRNM model is an example of multiple testing. In fact, we simultaneously test a set of hypotheses: the existence of a link between any pair of vertices in a network.
As discussed in paragraph 3.2.3, statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypothesis is low.
It is important to observe that the probability of observing a rare event increases when testing multiple hypotheses. This increased probability enlarges the likelihood of incorrectly rejecting a null hypothesis[2].
This is exactly our case, where we wish to define the link's presence for $\binom{N}{2}$ possibilities.
Let us now ask the question about the probability of observing at least one significant result just due to chance. Consider the case where we have 20 hypotheses to test (e.g. possible edges in a network), and we impose a significance level of 0.05. If S describes the event of having at least one significant result, S^c is the event of no significant results.

$$\begin{aligned} P(S) &= 1 - P(S^c) = \\ &= 1 - (1 - 0.05)^{20} \simeq 0.64 \end{aligned} \tag{3.9}$$

This simple calculation shows that we have a 64% chance of observing at least one significant result, even if all of the 20 tests are actually not significant.
We do not leave to reader's imagination how many significant results we can find for our case with $\binom{N}{2}$ possibilities, simply due to chance.

$$\begin{aligned} P(S) = 1 - (1 - \alpha_{\mathrm{TH}})^{\binom{N}{2}} \simeq 1 - (1 - \alpha_{\mathrm{TH}})^{N^2} &\simeq 1 - 10^{-9} \quad \text{for } N = 20,\ \alpha_{\mathrm{TH}} = 5\% \\ &\simeq 0.98 \qquad \text{for } N = 20,\ \alpha_{\mathrm{TH}} = 9\% \end{aligned} \tag{3.10}$$

One possible solution to multiple testing was proposed by Bonferroni [41].
The Bonferroni correction sets the significance cut-off at α/n. For example, in the example

[2]This kind of error is called a type I error (false positive) and is sometimes called an error of the first kind.

above, with 20 tests and α=0.05, you'd only reject a null hypothesis if the p-value is less than 0.0025.

$$P(S) = 1 - P(S^c) = \\ = 1 - (1 - 0.0025)^{20} \simeq 0.0488.$$

However depending on the correlation structure of the tests, the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives.

3.3 Introduction to the Fuzzy network approach

Though the SRNM model in section 3.2 appears to be trivial, section 3.2.3 reports some significant observations to pay attention to in many computational frameworks and, specifically, to our inverse problem. With the knowledge of multiple testing deficiencies, we do not opt for the thresholding procedure, neither we choose for the conservative Bonferroni correction. From now on, we only deal with the so-called fuzzy network approach.

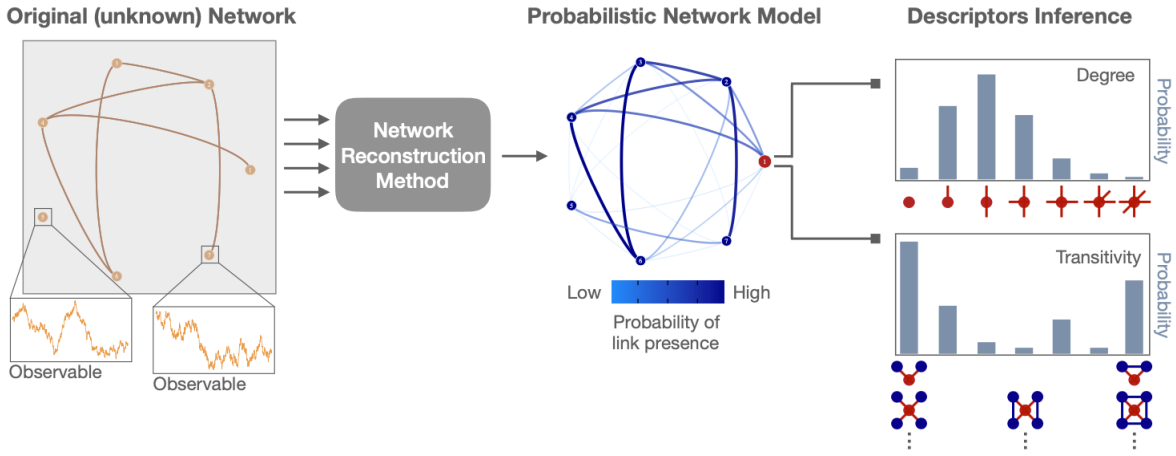


Figure 3.4: The figure from Ref. [26] synthesize the fuzzy network reconstruction method. In many practical problems we cannot observe directly the structure of a system, while we observe the time course of physical variables from each units of the system. The fuzzy reconstruction method allows us to obtain, within the framework of Bayesian inference, the probabilistic network model (central picture). Such a model provides us with the probability of link existence which uncertainty reflects on the topological descriptors of our model. E.g. degree and transitivity on the right side of the picture.

The fuzzy network method allows revealing the structural features of a complex system from the observed collective dynamics [42] (Figure 3.4). The proposed methodological framework allows us to analyze structural features of a complex network by taking into account all p-values, the strength of the evidence against the null hypothesis of lack of connectivity in our problem, without imposing a significance level/ cut-off (the thresholding criterion α_{TH} indeed).

Employing a Bayesian procedure [43, 44], reference [42] allows us to compute π_{ij}, the existence probability of an edge between node i and j, from the above-mentioned p_{ij}-values.

$$\pi_{ij} = 1 - \left[1 + \left(\frac{B_{ij} P(H_{ij}^0)}{1 - P(H_{ij}^0)}\right)\right], \tag{3.11}$$

where

$$B_{ij} = \begin{cases} -e p_{ij} \ln p_{ij} & \text{for } p_{ij} < e^{-1}, \\ 1 & \text{for } p_{ij} > e^{-1} \end{cases} \tag{3.12}$$

The only free parameter is the prior probability for the null hypothesis $P(H_{ij}^0)$ of the lack of connectivity between nodes i and j. In the following, we set it equal to $1 - \rho$ for all the edges, where ρ is the average density of the networks considered. With this choice, we are allowed to write the prior probability as $P(H^0)$, without the subscript ij.

The probabilities π_{ij} of existence of the edge between nodes i and j can be rearranged in a matrix Π, to obtain the probabilistic counterpart of the adjacency matrix:

$$\Pi = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1N} \\ \vdots & \ddots & \vdots \\ \pi_{N1} & \cdots & \pi_{NN} \end{bmatrix} \tag{3.13}$$

The matrix Π reminds us of a weighted adjacency matrix, but it has a different meaning: the value π_{ij} represents the existence probability of the (i, j) link. Therefore we are left with an ensemble of networks in which adjacency matrices are sampled from the Π.

Hence, each network in the fuzzy ensemble is considered a realization of the possibilities encoded in the probabilistic model Π. Besides avoiding the introduction of any arbitrary choice in the process and dealing with multiple testing deficiencies 3.2.3, the fuzzy network provides some theoretical insights based on a new formulation where uncertainty about the existence of the edges reflects the uncertainty about the topological descriptors (Chapter 4).

3.4 Fuzzy network modelling of the SKMN dynamics

Let us apply the same method of the previous section 3.2 to a toy model dynamics: the Stochastic Kuramoto model on a synthetic network.

$$\dot{\theta}_i(t) = \omega_i + \frac{\sigma}{k_i} \sum_{j=1}^{N} A_{ij} \sin(\theta_j(t) - \theta_i(t)) + \xi_i(t) \tag{3.14}$$

Similarly to model in 3.5, we start with N time series, but now we want to end up with the fuzzy network matrix Π. The procedure we follow is the same one described in the first two paragraphs of the previous section 3.2. However, we use the p-value information to compute Π from eq. (3.11) and eq. (3.12).

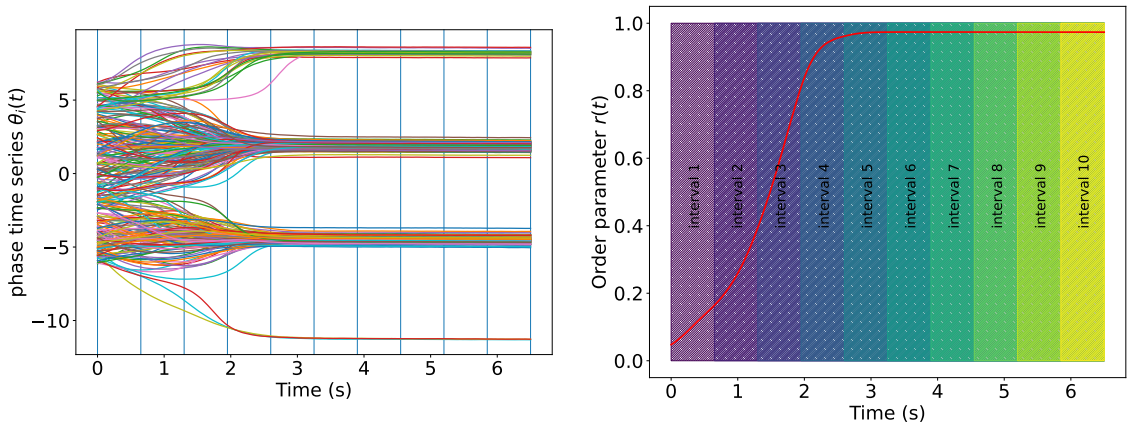
At this point, we can justify the choice in 3.2.1 for Pearson's coefficient as for the Kuramoto oscillators' dynamics the cross-correlation is usually the best performing similarity measure [45]. From results in Chapter 2, we set the following parameters:

Network	degree $\langle k \rangle$	σ	ϵ	T	dt	$P(H^0) = 1 - \rho$
Erdos-Renyi	~ 12.65	5	0.0001	6.5	0.00025	0.95

Table 3.1: Parameters in the SKMN dynamics (eq.(3.14)) to compute the fuzzy matrix Π.

where ρ is the density link of the synthetic Erdos-Renyi network.

We compute for each pair of oscillators the Pearson's coefficient $CC_{ij}^{(obs)}$ as in equation (3.6). We first choose to integrate the entire time series as shown in Figure 3.5A. The time considered is significant as we can tell we involve in the computation all the phases of synchronization dynamics Kuramoto model implies (see Figure 3.5B). As a second step, we use the time interval subdivision (shown in Figure 3.5B) to compute the fuzzy matrix for each time region involved in the synchronization process. Moreover, we remind that we create an ensemble $N_{it} = 20$ of single SKMN realization for a fixed coupling and noise, as prescripted in section 1.4.1.



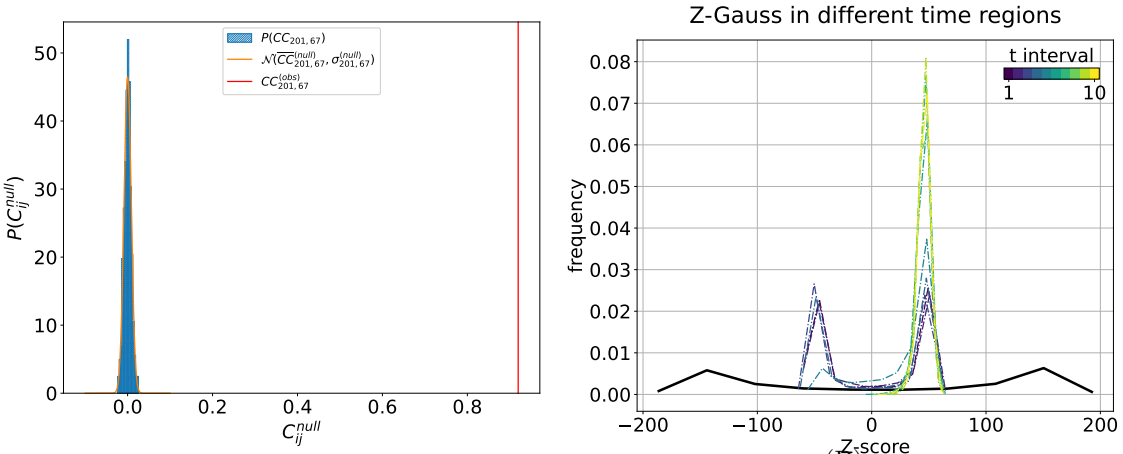
(A) (B)

Figure 3.5: Simulating stochastic Kuramoto model on Erdos-Renyi network, with $\langle k \rangle = 12$, $\sigma = 5$, $\epsilon = 0.0001$ (A) $\theta(t)$ phase time series for each node in the Kuramoto network. (B) Order parameter function in time computed as in eq. (1.22).

At this point, we have to test statistics from the SKMN dynamics against a null hypothesis. At first, we proceed with the same testing in the SRNM exercise in section 3.2.2 (RP surrogates). When realizing that data cannot be uncorrelated noise because of synchronization, we opt for IAFFT surrogates.

3.4.1 RP surrogates testing

As prescripted in section 3.1.1, we generate $M = 100$ RP surrogates for each time series of the Kuramoto model with the parameters in Table 3.1. $M = 100$ is sufficient to have significant statistics for the null model hypothesis.



(A) (B)

Figure 3.6: (A) Null hypothesis distribution for link $(i, j) = (6, 8)$ The orange line is the Gaussian distribution with $\mu = \overline{CC}^{(null)}_{ij}(0)$ and $\sigma = \sigma^{(null)}_{ij}$. The red vertical line individuates our observed case for the specific edge. (B) Z-score distribution for all the edges from the single SKMN realization. The solid black line represent the Z-score result obtained from the $CC^{(obs)}_{ij}$ of the full time series; whereas the dashed coloured lines are come from the analysis of the time series in different and growing in time temporal intervals of the original time series.

Therefore, we test the observed correlation coefficient $CC^{(obs)}_{ij}$ against the distribution of $CC^{(null,m)}_{ij}$, where $m = 1, 2, \ldots, M$ coming from the RP surrogates.
What we obtain is that for every link in the network, the correlation coefficient is very high (see Figure 3.6A).
Hence the Z-score (Figure 3.6B), for the single SKMN realization, obtained after the null hypothesis comparison, is very high with respect to the SRNM in the previous section. Therefore we reject the hypothesis of uncorrelated noise time series.
However, we notice a change in the Z-score distribution for time regions above the 4^{th} temporal interval. This interval coincides with the onset of synchronization (see Figure 3.5B), we explain the change as the time series from non-coherent, after 4^{th} temporal interval, became more and more synchronized until stationarity is reached.

We try to reduce the rejection of the null hypothesis for the single realization by instead relying on the average result in the $N_{it} = 20$ statistical ensemble due to noise variability. For each noise seed realization of SKMN we compute the $Z^{(k)}_{ij}$-scores.

Since $Z_{ij}^{(k)}$-scores come from a Gaussian distribution (see Figure 3.6A), we then can calculate the Z-Stouffer [46] to then proceed with the computation of p_{ij}-values.

$$Z_{Stouffer} = \frac{\sum_{k=1}^{N_{it}} Z^{(k)}}{\sqrt{N_{it}}} \tag{3.15}$$

$$p_{ij} - \text{value} = 1 - \text{erf}(Z_{ij}^{Stouffer}/\sqrt{2}) \tag{3.16}$$

where erf stands for error function[3] and $\text{erf}(Z/\sqrt{2})$ is the probability that the error of a single measurement lies between $-Z$ and $+Z$.
However this averaging procedure makes no difference. The consequence is that the p-value to reject the hypothesis of no link between two oscillator is mainly zero (Figure 3.7B). Therefore the fuzzy network method provides us with a link existence probability matrix π_{ij} mainly composed by ones. This coincides with the result for which we have a quasi-fully connected network, that gives us no information about the topology of the underneath system, which is indeed sparse.

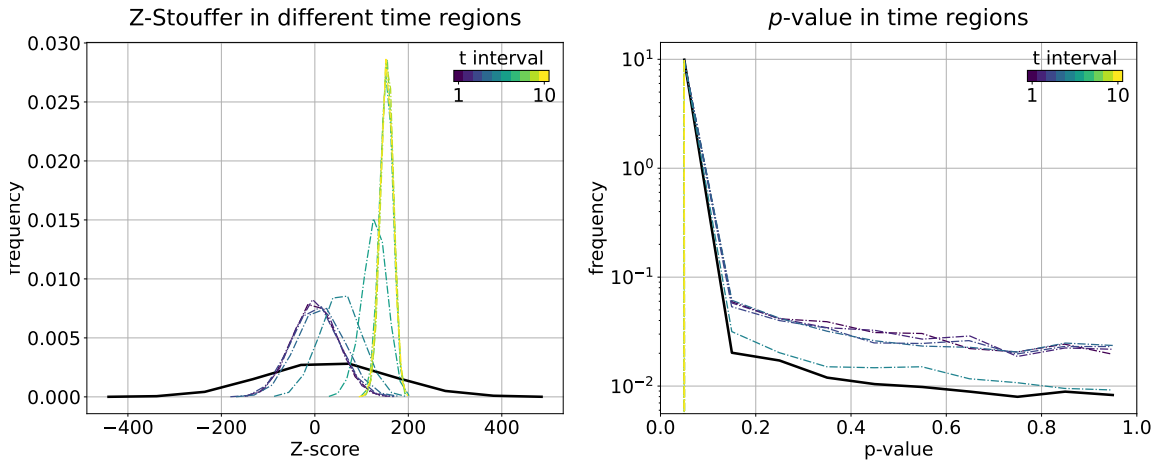


Figure 3.7: (A) Z_{ij}-Stouffer distribution for all the edges from all the single SKMN realizations in the N_{it} noise ensemble. The solid black line represents the Z_{ij}-Stouffer scores result obtained from the $CC_{ij}^{(obs)}$ of the full time series; whereas the dashed coloured lines come from the analysis of the time series in different and growing in time temporal intervals of the original time series. (B) p_{ij}-values distribution for all the edges from from all the single SKMN realizations in the N_{it} noise ensemble. p_{ij}-values are calculated by means of eq. (3.16) from the Z_{ij}-Stouffer results found from eq. (3.15).

The answer to this result is shown in the following plot 3.8 where we point out the effect

[3] $\text{erf}\, x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$

of the reshuffling on a single time series θ_i.

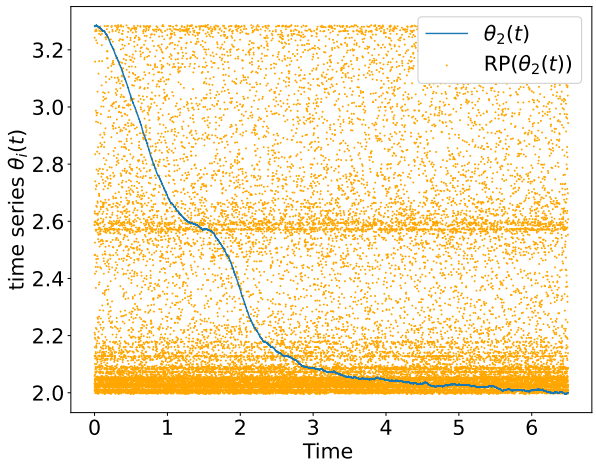
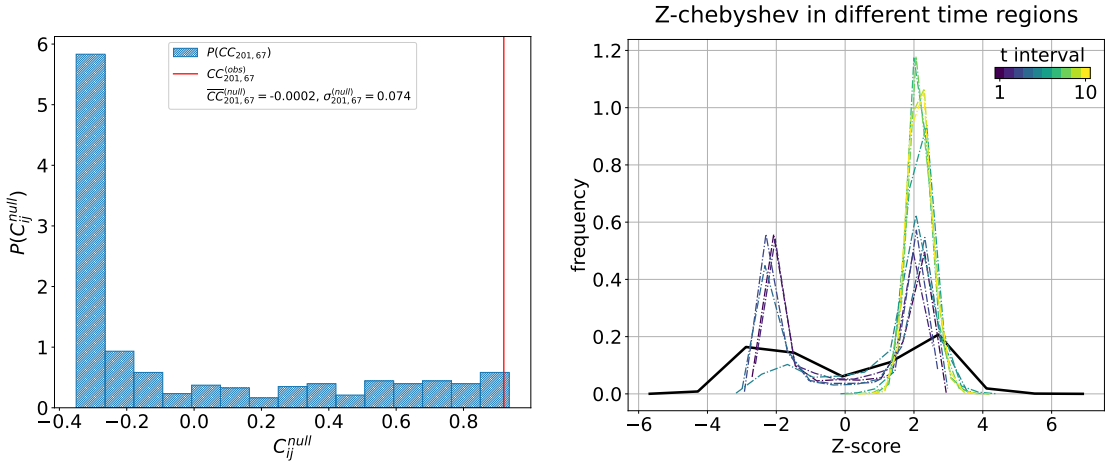


Figure 3.8: The blue line is a single time series of the Kuramoto model see Figure 3.5A, the orange plot is the time series reshuffled for the null hypothesis comparison.

3.4.2 IAFFT surrogates testing

Since the failure of the previous null hypothesis testing, we try a more specific statistical test: the non linear testing discussed in section 3.1.2 and, more precisely, we create $M = 100$ IAFFT surrogates by the use of `pyunicorn.py` package developed in [47] with 5 iterations.



(A) (B)

Figure 3.9: (A) Null hypothesis distribution for link $(i, j) = (6, 8)$ the distribution is non-Gaussian with $\mu = \overline{CC}^{(null)}_{ij}$ and $\sigma = \sigma^{(null)}_{ij}$. The red vertical line individuates our observed case for the specific edge. (B) Z-score distribution for all the edges from the single SKMN realization. The solid black line represent the Z-score result obtained from the $CC^{(obs)}_{ij}$ of the full time series; whereas the dashed coloured lines are come from the analysis of the time series in different and growing in time temporal intervals of the original time series.

As the distribution of the null-IAFFT is non-Gaussian (see Figure 3.9A), we then proceed with Chebishev inequality [48]. In probability theory, Chebyshev's inequality guarantees that no more than a certain fraction of values can be more than a certain distance from the mean. Inequality has great utility because it can be applied for a wide class of probability distributions. Let $CC_{ij}^{(null)}$ be the random variable with finite expected value $\mu = \overline{CC}_{ij}^{(null)}$ and finite non-zero variance $\sigma_{ij}^{2\ (null)}$. Then for any real number $z > 0$,

$$\Pr(|CC_{ij}^{(obs)} - \overline{CC}_{ij}^{(null)}| \geq z\sigma_{ij}^{(null)}) \leq \frac{1}{z^2}. \tag{3.17}$$

Therefore, we are left with an ensemble of $Z_{ij}^{(k),Cheb}$ (Figure 3.9B), the z calculated in eq. (3.17). We first find the p_{ij}-value for each $Z_{ij}^{(k),Cheb}$-score by using Chebishev inequality:

$$p_{ij}^{(k)} - \text{value} = \frac{1}{\left[Z_{ij}^{(k),Cheb}\right]^2}. \tag{3.18}$$

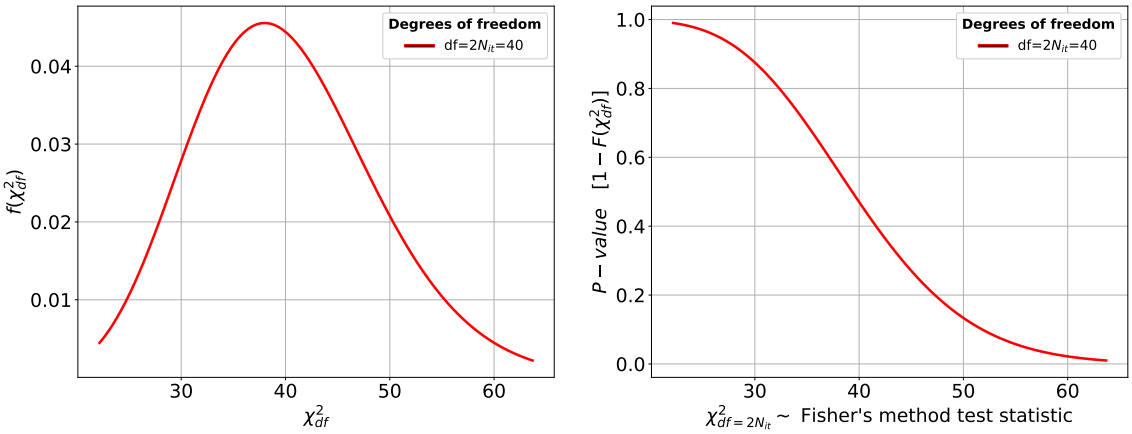


Figure 3.10: (A) Probability density function of χ^2_{df}, with $df = 2N_{it} = 40$ (B) P-value distribution from cumulative density function of χ^2_{df}, with $df = 2N_{it} = 40$.

We then apply the Fisher-test [49,50]. Fisher method combines extreme value probabilities from each test, commonly known as p-values, into one test statistic (χ^2) using the formula:

$$\chi^2_{2N_{it},ij} \sim -2\sum_{k=1}^{N_{it}} \log(p_{ij}^{(k)} - \text{value}), \tag{3.19}$$

As final step we compute the p-value of the empirical $\chi^2_{2N_{it}}$ found (Figure 3.11A).
To do it we compare the values with the $1 - CDF(\chi^2_{df})$ by imposing $df = 2N_{it} = 40$ degrees of freedom[4]. The following plots (Figure 3.10A and 3.10B) show both the χ^2_{df}

[4]PDF and CDF are respectively the probability and the cumulative density function

distribution $f(\chi^2_{df})$ and the $1 - CDF(\chi^2_{df})$ function to asses the p-value of our observable. In the following formulas, we set $x := \chi^2_{df}$

$$f^{df}(x) = PDF^{df}(x) = \frac{1}{2^{df/2}\Gamma(df/2)}\, x^{df/2-1} e^{-x/2} \tag{3.20}$$

$$F^{df}(x) = CDF^{df}(x) = \frac{1}{\Gamma(df/2)}\, \gamma\left(\frac{df}{2}, \frac{x}{2}\right) \tag{3.21}$$

$$p - value = 1 - F^{df}(x) \tag{3.22}$$

By comparing the $\chi^2_{2N_{it}}$ result with the function in Figure 3.10B, we then derive the p_{ij}-value distribution for each edge in the Network (Figure 3.11B). By the use of these p_{ij}-values, we the recover the fuzzy probability matrix Π.

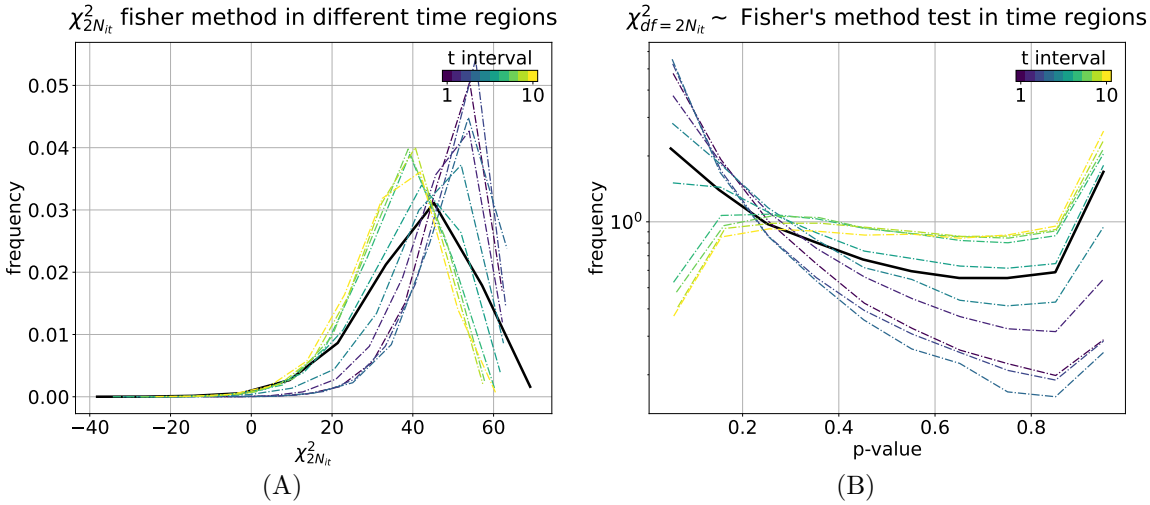


Figure 3.11: (A) $\chi^2_{2N_{it}}$ distribution for all the edges from the single SKMN realization. The solid black line represents the $\chi^2_{2N_{it}}$ result obtained from the full time series; whereas the dashed coloured lines come from the analysis of the time series in different time temporal intervals of the original time series. (B) p_{ij}-values distribution for $\chi^2_{2N_{it}}$ results in panel A. p_{ij}-values are calculated by means of eq. (3.22) from the $\chi^2_{2N_{it}}$ results found for Fisher test eq. (3.19).

The result for the IAFFT is the opposite of the RP surrogates null hypothesis results. Indeed, the probability of rejecting the null hypothesis is higher in the first part of the dynamics since the SKMN dynamics shows an highly non-coherent collective behaviour. Whereas after the r_∞ global steady state is reached all time-series are linear and therefore the associated p_{ij}-values are higher on average ($int\ d$, with $d > 4$ in Figure 3.11B).
Therefore to calculate a significant fuzzy matrix Π, we use the values found from analyzing the time region interval before synchronization is reached. We then plot the link existence probability distribution (red line in Figure 3.12), that represents the fuzzy matrix we

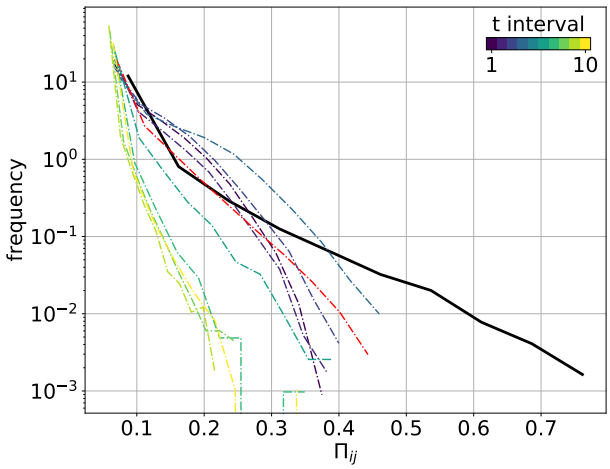


Figure 3.12: Fuzzy matrix π_{ij} distribution. The solid black line represents the result obtained from the full time series; whereas the dashed coloured lines come from the analysis of the time series in different time temporal intervals of the original time series. The line red line highlight the π_{ij} distribution at the onset of synchronization.

choose to compute robustness properties in the next chapter.

Now that we have found a meaningful fuzzy network linked to the collective dynamics of SKMN, we are ready to extend the network percolation theory to the fuzzy network approach and plot the results for the robustness properties of our inferred network.

Chapter 4

Inferring network robustness from network dynamics

In the previous chapter, we explained why introducing the fuzzy network approach is relevant to network topology inference. Besides, by looking at the SKMN time series, we extract the fuzzy probability matrix Π, where an existence probability describes each edge in our reconstructed network. In the following discussion, we at first extend the arguments in [42] by reformulating some topological descriptors of a complex network having information about each edge's existence π_{ij}.
Specifically, we focus on excess degree distribution and then the percolation problem. At last, we use the theoretical results about percolation in this new approach to assessing the robustness properties of the reconstructed fuzzy network Π. We then compare the results with the ones for the original network underneath.

4.1 Building the node excess degree distribution

4.1.1 Standard case

Let us start by defining the standard degree distribution $p(k)$ and $q(k)$.
$p(k)$ is the probability that a randomly chosen node in our network has exactly k links attached or, alternatively, $p(k)$ can be read as the fraction of nodes in the network with k-degree.

Instead, $q(k)$ is the probability distribution of the excess degree at the end of an edge. Where the excess degree is the number of edges attached to that vertex other than the one we arrived along. Alternatively $q(k)$ is the fraction of nodes in the network with excess degree k.

Moreover the unique mapping between $p(k)$ and $q(k)$, that is intrinsic in $q(k)$ definition:

$$q(k) = \frac{(k+1)p(k+1)}{\langle k \rangle} \tag{4.1}$$

holds for the configurational model networks.

Single fixed network

In the following we denote index (G) to indicate the single network G we are referring to. When dealing with G network, $p_i^{(G)}(k)$ is the probability that i node has k neighbours. That is a delta-dirac function on $d_i^{(G)}$, the value of neighbours node i has in (G) specific network:

$$p_i^{(G)}(k) = \delta(k - d_i^{(G)}) \tag{4.2}$$

We define $q_i^{(G)}(k)$ the excess probability distribution for a specific node in the network, what we will call the node excess degree distribution.
We first have to find an analogous definition coherent with the standard degree distribution, that is: $q_i^{(G)}(k)$ is the probability that by following one of i's links we find a neighbour whose excess degree is k (Figure 4.1).

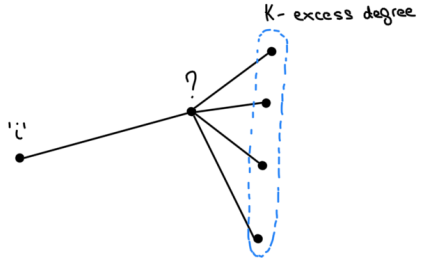


Figure 4.1: Node excess degree definition.

Given that, we first try to make the previous definition more precise for a single fixed network. This definition will be fundamental both to computational use and as intermediate step to the theoretical part, where we then try to extend the reasoning for the fuzzy network approach.

By taking a specific node i, we can have k_i^* neighbours. However if we are dealing with a unique network G, individuated by its adjacency matrix A, the number of neighbors of node i will be fixed by the value $d_i^{(G)}$. That could be easily represented by a delta Dirac on $d_i^{(G)}$ fixed value.

Once knowing the exact number of first neighbours of i node, we have two cases:

- Either the i node has $d_i^{(G)} = 0$ (i is isolated) and therefore ${}^{(G)}q_i(k) = 0 \quad \forall k = 0, N-2$.

- Either $d_i^{(G)} \neq 0$.

Let us define $F_{k_i^*}$, that is the ensemble of all the sets, N_i, of k_i^* neighbours of node i. $|F_{k_i^*}| = \binom{N}{k_i^*}$. Each $N_i \in F_{k_i^*}$ contains the specific neighbouring nodes of i, $|N_i| = k_i^*$ is its cardinality.
As the network is fixed, $d_i^{(G)}$ is a unique numerical value individuating a unique configuration of $d_i^{(G)}$ first neighbours, that we will call N_i^{fix}. Therefore among all the combinations of k_i^* nodes in G, $F_{k_i^*}$, the only significative is N_i^{fix}. In N_i^{fix} set we then count for all the neighbours who have excess degree exactly equal to k. We then obtain the following formula:

$$\begin{aligned} q_i(k) &= \sum_{k_i^*=1}^{N-1} \delta(d_i = k_i^*) \left[\sum_{N_i \in F_{k_i^*}} \delta(N_i^{\text{fix}} = N_i) \left[\frac{\sum_{j \in N_i} \delta(d_j^{(-i)} = k)}{k_i^*} \right] \right] \\ \sum_{k=0}^{N-2} q_i(k) &= 1 \quad \forall i \in G, d_i \neq 0 \end{aligned} \tag{4.3}$$

where $\delta(d_j^{(-i)} = k)$ returns 1 if node j (neighbour of i) has exactly k neighbours other than i.
Equation (4.3) is a complicated way of writing the following:

$$\begin{aligned} q_i(k) &= \sum_{N_i \in F_{d_i}} \delta(N_i^{\text{fix}} = N_i) \left[\frac{\sum_{j \in N_i} \delta(d_j^{(-i)} = k)}{k_i^*} \right] = \\ &= \frac{\sum_{j \in N_i^{\text{fix}}} \delta(d_j^{(-i)} = k)}{k_i^*} \end{aligned} \tag{4.4}$$

The last equation is of immediate meaning: that is the number of neighbours among the ensemble N_i^{fix}, the actual neighbour of i-th node, which have k links without considering the node i from which we start.

4.1.2 Theoretical formula for fuzzy networks

In reference [42], the probability distribution linked to the i-th node degree is derived as:

$$p_i(k) = \sum_{N_i \in F_{k_i}} \prod_{j \in N_i} \pi_{ij} \prod_{l \in N_i^c} [1 - \pi_{il}] \tag{4.5}$$

We would like to recover an analogous formula from the excess degree distribution. We start our reasoning from eq. (4.3):

$$q_i(k) = \sum_{k_i^*=1}^{N-1} \delta(d_i = k_i^*) \left[\sum_{N_i \in F_{k_i^*}} \delta(N_i^{\text{fix}} = N_i) \left[\frac{\sum_{t \in N_i} \delta(d_j^{(-i)} = k)}{k_i^*} \right] \right],$$

where we have to substitute all the deltas, to include the variability the fuzzy network approach provides.
We will rewrite it in the following form:

$$q_i(k) = \sum_{k_i^*=1}^{N-1} \left[\sum_{N_i \in F_{k_i^*}} P(N_i|k_i^*) \cdot \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right], \tag{4.6}$$

The $\delta(d_i = k_i^*)$ and $\delta(N_i^{\text{fix}} = N_i)$, that univocally identify the single network have to be subsituted by considering all the possible numbers of neighbours and the consequent configurations each node i could assume, $P(N_i|k_i^*)$. $P(N_i|k_i^*)$ is of immediate derivation by:

$$P(N_i|k_i^*) = \prod_{j \in N_i} \pi_{ij} \prod_{l \in N_i^c} [1 - \pi_{il}] \tag{4.7}$$

where N_i^c is the set containing all the other nodes in the network other than i and the k_i^* neighbours, N_i^c is the complement of N_i.
$\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle = \langle n \rangle$ is the average number of first neighbours, in the N_i configuration, which have degree k excluding the link with i node. The average is due to the fact that each first neighbour j allows for two possibilities:

$$\begin{cases} d_j = k \text{ with prob } p_j^{(-i)}(k) \\ d_j \neq k \text{ with prob } 1 - p_j^{(-i)}(k) \end{cases} \quad \forall j \in N_i \tag{4.8}$$

n.b. $p_j^{(-i)}(k)$ is the same as the node degree distribution computed by formula (4.5) except the fact we remove the neighbour node i from the calculation as we are interested in the excess degree.
Therefore we can have k_i^* different possibilities for n, the number of first neighbours in the N_i configuration which have degree k excluding the link with i node:

- $n = 0$ no first neighbour with excess degree k, this happen with probability $\prod_{t \in N_i}[1 - p_t^{(-i)}(k)]$;
- $n = 1$ just one first neighbour with excess degree k, this happen with probability $\sum_{j \in N_i} p_j^{(-i)}(k) \prod_{t \in N_i \setminus \{j\}}[1 - p_t^{(-i)}(k)]$;
- ...
- ...
- ...
- $n = k_i^*$ all first neighbours with excess degree k, this happen with probability $\prod_{t \in N_i}[p_t^{(-i)}(k)]$.

We then must take into account all the different possibilities for each $j \in N_i$

$$q_i(k) = \sum_{k_i^*=1}^{N-1} \sum_{N_i \in F_{k_i^*}} \left[\prod_{j \in N_i} \pi_{ij} \prod_{l \in N_i^c} [1-\pi_{il}] \cdot \frac{\sum_{n=0}^{k_i^*} n \sum_{\Lambda \in N_i^{(n)}} \prod_{a \in \Lambda} p_a^{(-i)}(k) \prod_{b \in \Lambda^c} [1 - p_a^{(-i)}(k)]}{k_i^*} \right] \tag{4.9}$$

This will bring us to the following formula:

$$\langle n \rangle = \langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle = \sum_{n=0}^{k_i^*} n \sum_{\Lambda \in N_i^{(n)}} \prod_{a \in \Lambda} p_a^{(-i)}(k) \prod_{b \in \Lambda^c} [1 - p_a^{(-i)}(k)], \tag{4.10}$$

where we take into account all the possibilities in the summation $\sum_{\Lambda \in N_i^{(n)}}$, in which Λ considers all the possible combination of neighbours in N_i, $|N_i^{(n)}| = \binom{k_i^*}{n}$, for which exactly n neighbours among the k_i^* have excess degree k. The formula is valid $\forall k = 0, 1, ..., N-2$. We can notice that in the following reasoning we have not considered the case in which i node is isolated. Following the fuzzy network approach reasoning in Raimondo paper this happens with a frequency equal to:

$$p_i(k=0) = \prod_{j \in G \setminus \{i\}} [1 - \pi_{ij}] \tag{4.11}$$

where $j \in G \setminus \{i\}$ means j can assume be any node in the network apart from i-th vertex. Therefore it makes sense to normalize only on the not isolated i possibilities which is computed as $p_i(k \neq 0) = 1 - p_i(k=0)$.
$\pi_{ij} = 0, \quad \forall j \in G$, is the pathological case in which node i is isolated for all the networks in the fuzzy ensemble. Since there is no link between nodes j and i, we can not define a node excess degree distribution for this pathological case.
The final formula is then:

$$q_i(k) = \frac{\sum_{k_i^*=1}^{N-1} \left[\sum_{N_i \in F_{k_i^*}} P(N_i | k_i^*) \cdot \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right]}{p_i(k \neq 0)}, \tag{4.12}$$

Comparison between theoretical and computational approach

When dealing with the fuzzy network approach we start from $\Pi = [\pi_{ij}] \quad \forall i, j \in G$ which inform us about the existence probability of each link e_{ij}, $P(e_{ij}) = Bern(\pi_{ij})$.
Hence our computational framework will deal with an ensemble of simulated networks $\{A_{ij}^{(G)}\}_{G=1,...,N_G}$, in which each $A_{ij}^{(G)} = P(e_{ij})$. The cardinality of the ensemble, N_G has to be sufficiently numerous to correctly represent our problem $\sim 10^5$ (we set N_G so high to better visualize that eq. (4.12) correctly match the computational results in Figure 4.2). Therefore for each $A_{ij}^{(G)}$ we compute $q_i^{(G)}(k)$ by using Eq (4.4).
The average of the node excess degree distribution over the whole ensemble will provide

us with a computational value of this fuzzy network property, which will be compared to the same theoretical measure to test it.

$$q_i^{\text{comp}}(k) = \frac{1}{N_G - N_{\text{isolated}}(d_i = 0)} \sum_{G=1}^{N_G} q_i^{(G)}(k) \tag{4.13}$$

$N_{\text{isolated}}(d_i = 0)$ is the number of networks in the ensemble for which i node is isolated. The $N_s - N_{\text{isolated}}(d_i = 0)$ normalization is therefore due to the fact that $q_i(k)$ has no meaning for an isolated node.

The following Figure shows the result for the toy network of $N = 5$ nodes with a synthetic fuzzy matrix. The panel on the left is the same picture presented in [42], node degree distribution follows the same curve both by averaging $p_i^{(G)}(k)$ value in eq. (4.2) on every network of the fuzzy ensemble ($p_i^{\text{comp}}(k) = \frac{1}{N_G} \sum_{G=1}^{N_G} q_i^{(G)}(k)$) and by the use of eq. (4.5). This is not a new result. However if we look at the panel on the right in Figure 4.2, we notice that the node excess degree distribution in formula (4.12) perfectly matches the behaviour of the fuzzy ensemble (4.13). That reveals our analytical reasoning is correct.

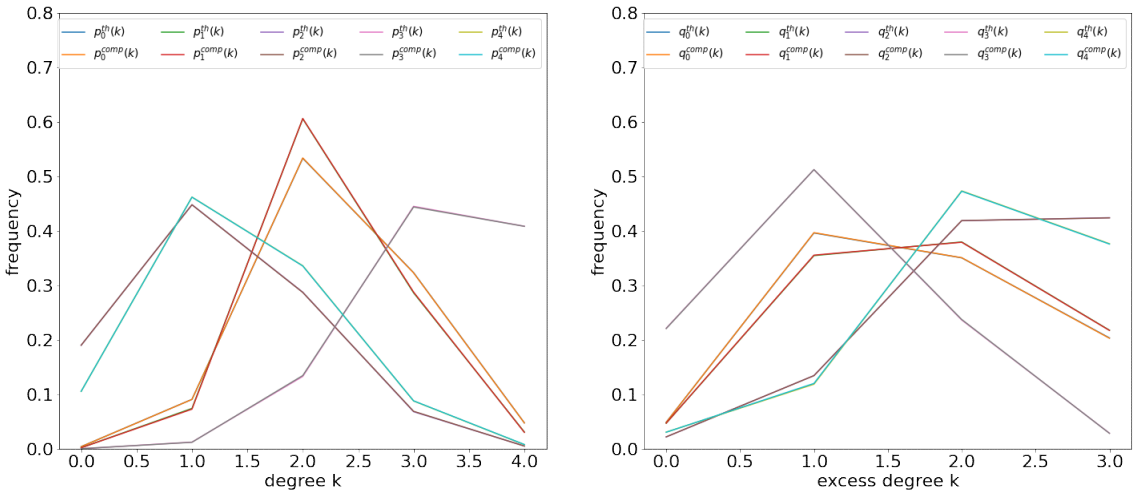


Figure 4.2: On the left we plot the figure for the $N = 5$ toy network in [42] paper where theoretical formula (4.5) perfectly recovers the simulation results of the fuzzy ensemble where 10^5 network are sampled from π_{ij}.
The figure on the right shows the node excess degree distribution in formula (4.12) that perfectly recovers the behaviour of the fuzzy ensemble (4.13).

4.2 Recovering the standard excess degree distribution

In Figure 4.3, we try to recover the standard degree and excess degree results by computing $p^{(G)}(k)$ and $q^{(G)}(k)$ for each network G of the ensemble $\{A_{ij}^{(G)}\}_{G=1,\dots,N_G}$ and then plotting

the average values $\bar{p}(k) = \sum_{G=1}^{N_G} p^{(G)}(k)$ and $\bar{q}(k) = \sum_{G=1}^{N_G} q^{(G)}(k)$.
If standard excess degree distribution $p(k)$ can be easily recovered from the node excess degree distribution $p_i(k)$ (Figure 4.3 on the left):

$$p(k) = \frac{1}{N}\sum_{i\in G} p_i(k) = \frac{1}{N}\sum_{i\in G}\sum_{N_i\in F_{k_i}}\prod_{j\in N_i}\pi_{ij}\prod_{l\in N_i^c}[1-\pi_{il}]; \tag{4.14}$$

there are some problems in recovering the standard excess degree distribution $q(k)$ from formula (4.12) (not perfect match in two histograms in right panel of Figure 4.3).

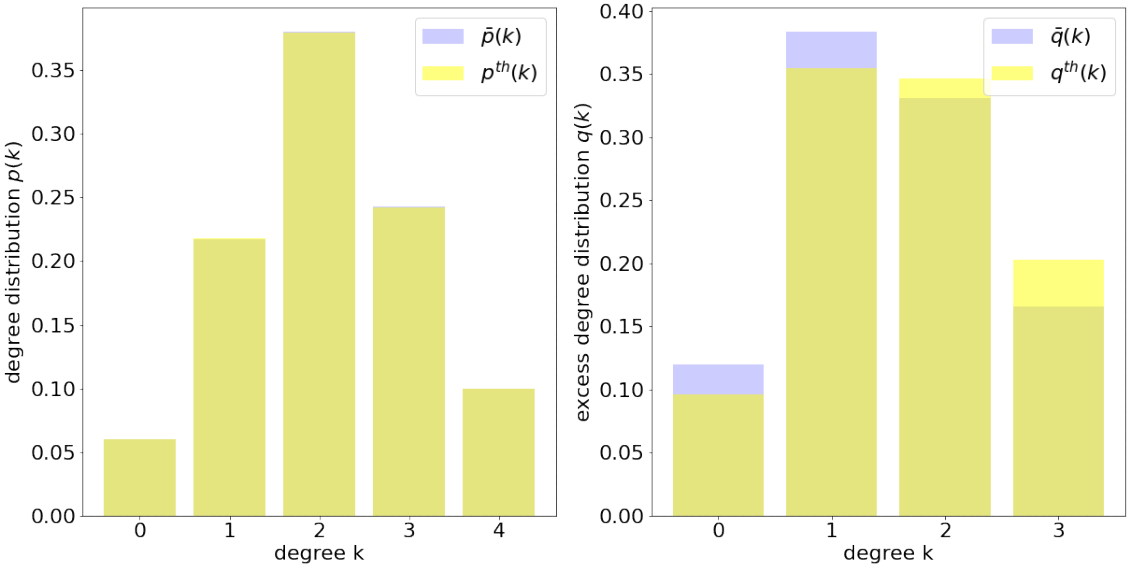


Figure 4.3: For the $N = 5$ toy network in [42] we plot the degree distribution $p(k)$. $\bar{p}(k)$ and $\bar{q}(k)$ labels in legend stands for the average computation of degree/excess degree distribution on every network in the fuzzy ensemble. $p^{th}(k)$ and $q^{th}(k)$ are the values computed from eq. (4.14) and eq. (4.15), respectively. On the left panel we show the results for degree distribution. On the right panel we show the results for excess degree distribution.

Indeed, we tried the following ansatz:

$$q(k) = \sum_{i\in G} q_i(k)\cdot\frac{\mu_{d_i}}{2\langle m\rangle} \tag{4.15}$$

where μ_{d_i} and $\langle m\rangle$ are defined respectively the mean average degree for node i ($\sum_j \pi_{ij}$) and the average number of edges ($\sum_i\sum_j \pi_{ij}/2$). We then dimostrate that this formula

provides with a normalized excess degree distribution.

$$\sum_{k=0}^{N-2} q(k) = \sum_{k=0}^{N-2} \sum_{i \in G} q_i(k) \cdot \frac{\mu_{d_i}}{2\langle m \rangle} = \sum_{i \in G} \underbrace{\sum_{k=0}^{N-2} q_i(k)}_{=1} \cdot \frac{\mu_{d_i}}{2\langle m \rangle} = 1 \tag{4.16}$$

If this ansatz is correct and, therefore, we recover the standard definition of excess degree distribution from $q_i(k)$, that would be the ideal case. Since, in network theory, $q(k)$ can be recovered from $p(k)$ and, moreover $p(k) = \frac{1}{N} \sum_{I \in G} p_i(k)$ (eq. (4.3)), this means there should be a unique mapping between $p_i(k)$ and $q_i(k)$. But as we will see in the following discussion this is not our case.
In fact finding a dependence of $q_i(k)$ on $p_i(k)$, $q_i(k, p_i(k))$ would mean rewrite eq. (4.6) in a way that allows us showing the $p_i(k)$ term.
This is correct if we find that the sum over the product of two successions is equal to the product of the separate sum of the same two successions.
If this is not true in general, we can computationally show it is still not the case of our formula (4.6).

$$q_i(k) = \sum_{k_i^*=1}^{N-1} \left[\sum_{N_i \in F_{k_i^*}} P(N_i | k_i^*) \cdot \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right], \tag{4.17}$$

$$\begin{aligned} q_i(k) &\neq \sum_{k_i^*=1}^{N-1} \left[\sum_{N_i \in F_{k_i^*}} P(N_i | k_i^*) \cdot \sum_{N_i \in F_{k_i^*}} \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right] = \\ &= \sum_{k_i^*=1}^{N-1} \left[p_i(k_i^*) \cdot \sum_{N_i \in F_{k_i^*}} \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right] \neq \\ &\neq \sum_{k_i^*=1}^{N-1} p_i(k_i^*) \cdot \sum_{k_i^*=1}^{N-1} \left[\sum_{N_i \in F_{k_i^*}} \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right] = \\ &= \sum_{k_i^*=1}^{N-1} p_i(k_i^*) \cdot \sum_{k_i^*=1}^{N-1} \left[\sum_{N_i \in F_{k_i^*}} \frac{\langle \sum_{j \in N_i} \delta(d_j^{(-i)} = k) \rangle}{k_i^*} \right] \end{aligned} \tag{4.18}$$

Therefore we deduce the ansatz in which $q(k) = \sum_{I \in G} q_i(k) * \frac{\mu_i}{2\langle m \rangle}$ is not true. There is no mapping between the standard excess degree distribution and our new formulation.

Now let us discuss the main point of our work: we want to look at the robustness properties of networks described by the fuzzy model Π derived by observing the SKMN collective dynamics. We test whether the reconstructed network resilience properties are representative of the synthetic network underneath the SKMN.

4.3 Percolation on networks

The main aim is to find robustness properties by extending the generating function formalism to the fuzzy network case since existing techniques are only valid for physical links. Now that we have found a precise reformulation for the excess degree distribution on each node of the fuzzy network, we exploit the result or, at least, the same reasoning to face the percolation problem.
We choose to measure network robustness properties in terms of percolation theory since it provides the most complete way to assess the system's response to failure through node removal. We start by considering some general network theory results to gain some understanding of the percolation transition and the giant cluster.
Let us consider the behaviour of the site percolation process on networks generated using the configuration model (chapter 13 and 16 in [2]), a simple but useful model of a network with a specified degree distribution. The mathematical formalism of the generating function is here used.
Being ϕ, the occupation probability of a vertex, u is the average probability that a vertex is not connected to the giant cluster via its connection to some particular neighbour.

$$u = \sum_k q(k)(1 - \phi + \phi u^k) \tag{4.19}$$

Then the probability of belonging to the giant cluster is:

$$S = 1 - \sum_k p(k)u^k \tag{4.20}$$

Vertex i is not connected to the giant cluster if is not connected to the giant cluster via all its connections. Supposing its degree is k, averaging over all nodes in the network we obtain $\sum_k p(k)u^k$.

All these results are well-known in literature. Let us deal with the new framework [42].
In the fuzzy network setting, the uncertainty about the existence of the edges is reflected in the uncertainty about the topological descriptors. Therefore we must reformulate the network descriptor u, with the one at the level of the single node, u_i.
u_i is the probability that vertex i is not connected to the giant cluster GC via its connection to some particular neighbouring vertex. Vertex i is not connected to the giant cluster if its not connected to the giant cluster via all its connections.
Therefore we can define the node quantity s_i as the probability that the i-th node belongs to the Giant Cluster. The average property is then recovered by:

$$S = \frac{1}{N} \sum_{i \in G} s_i \tag{4.21}$$

The formula for u_i and s_i follows the same reasoning of the previous section on the node

excess degree distribution.

$$u_i = 1 - \phi + \phi \cdot \sum_{k_i^*=1}^{N-1} \sum_{N_i \in F_{k_i^*}} \left[\prod_{j \in N_i} \left[\pi_{ij} \cdot \sum_{k=0}^{N-2} \sum_{N_j \in F_k^{(-i)}} \prod_{t \in N_j} \pi_{jt} u_t \prod_{s \in N_j^c} [1 - \pi_{js}] \right] \prod_{l \in N_i^c} [1 - \pi_{il}] \right] \tag{4.22}$$

To make eq. 4.22 clearer we rewrite it in the following way:

$$u_i = 1 - \phi + \phi \cdot \sum_{k_i^*=1}^{N-1} \sum_{N_i \in F_{k_i^*}} \left[\prod_{j \in N_i} \left[\pi_{ij} \cdot f(\text{neighbours of i node}) \right] \prod_{l \in N_i^c} [1 - \pi_{il}] \right] \tag{4.23}$$

where

$$f(\text{neighbours of i node}) = \sum_{k=0}^{N-2} \sum_{N_j \in F_k^{(-i)}} \prod_{t \in N_j} \pi_{jt} u_t \prod_{s \in N_j^c} [1 - \pi_{js}] \tag{4.24}$$

$f(\text{neighbours of i node})$ is a function depending on the neighbours of node i. However every neighbour of node i, let us call it j has k excess degree.
Summatory $\sum_{k=0}^{N-2}$ takes in account all the possible excess degrees, node j can have. Therefore we further look at each neighbour of j, let call it t. Node t belongs to N_j, the set that contains a specific combination of $\binom{N-2}{k}$ nodes among all possible sets in the ensemble $F_k^{(-i)}$. Obviously both node i and j are excluded from the $N-2$ in the combinatorial term $\binom{N-2}{k}$.
As final step we multiply π_{jt} the existence probability of the link between node j and node t to the probability u_t that t does not belong to the Giant Cluster. Our formula is therefore computed by successive iteration as in the message passing algorithm [51]. In fact we start our algorithm by assigning to u_i random real numbers between $[0, 1]$ interval, we iterate formula (4.22) until all u_i converge to stable values. Then we compute s_i from u_i to the recover the Giant Cluster S.

$$s_i = \phi \left[1 - \sum_{k=0}^{N-1} \sum_{N_i \in F_k} \prod_{j \in N_i} \pi_{ij} \cdot u_j \prod_{l \in N_i^c} [1 - \pi_{il}] \right]. \tag{4.25}$$

4.4 Computational results for robustness properties

In Chapter 3, we showed that the probabilistic network, the so-called fuzzy network, relies on finding a probability matrix, Π, in which entries π_{ij} define the existence probability of each edge. In previous sections we exploit the theoretical insights based on this new formulation where uncertainty about the existence of the edges reflects the uncertainty about the topological descriptors (node excess degree distribution and giant cluster component). At the same time, the fuzzy network allows for a computational approach where we compute the topological descriptors of each complex network of the fuzzy ensemble and then average them to obtain the information about the fuzzy network.

In this last section, we want to look at the robustness properties of the network inferred starting from the SKMN collective dynamics. We test whether the reconstructed network resilience properties are representative of the synthetic network underneath the SKMN, the Erdos-Renyi graph in table 4.2.
The first step is to derive the fuzzy "adjacency" matrix, Π. Indeed, we use the result found in the last part of third chapter. To make the derivation clear, we resume the procedure. The inference method involves the SKMN collective dynamics analysis. In chapter 2, we find the model setting parameters for which the best reconstruction is shown by help of the standard thresholding procedure. Our results find that the topological inference is better achieved for coupling higher than the critical coupling σ_c and low noise intensities. Moreover, by the time region dynamics subdivision (section 2.3), we find an optimal reconstruction region just before the transient time is off.
As a second step, we fix this optimal setting for (σ, ϵ) parameters and we extract time series from the time region before the collective dynamics results in the order parameter reaching its steady state, $r(t) \to r_\infty$.
With this information we can calculate the probability matrix of edge existence Π.

Let us now concentrate on the computational part of the fuzzy network approach. We proceed by sampling an ensemble from the Π matrix. Hence our computational framework will deal with an ensemble of N_G simulated networks $\{A_{ij}^{(G)}\}_{G=1,\dots,N_G}$, in which each $A_{ij}^{(G)} = P(e_{ij})$. Where $P(e_{ij})$ informs us about the existence probability of each link e_{ij}, $P(e_{ij}) = Bern(\pi_{ij})$.
We choose the cardinality of this ensemble to enumerate 10^3 networks. This value is set in order to make the algorithm feasible and at the same time consider a numerous enough statistics.

For each graph in the fuzzy ensemble, univocally individuated by its adjacency matrix $A^{(G)}$, we test robustness by using percolation network theory. We opt for the uniform random removal method. The algorithm for each network $A^{(G)}$ is the following, where ϕ is the probability of a node in the network to be removed, which is equal to the fraction of removed vertices.
We start by $\phi = 0$ value (no removed nodes) we randomly select the ϕN nodes in our analyzed network and we delete them, cancelling out all the connections with the still existing vertices. After the removal, we calculate the value of the Giant Cluster, which is the largest connected component in our network. We then proceed by removing more and more nodes in the network until $\phi = 1$ (see Table 4.1). We remind that the first step of this algorithm allows the computation of the *Giant Component* since no nodes are removed.
For each network $A^{(G)}$, we iterate the random removal 100 times to avoid our algorithm being affected by the order in which we remove the nodes.
We then obtain 100 sequences of $\{\mathrm{GC}^\phi\}_{\phi=0,0.1,\dots,1}$. For each GC^ϕ we compute the average over the 100 values: $\overline{\mathrm{GC}}(\phi) = \sum_{i=1}^{100} \mathrm{GC}_i^\phi / 100$.

We repeat the above procedure for each $A^{(G)}$ in the fuzzy ensemble. We proceed by com-

Algorithm: GC node random removal
0: $\phi = 0$, $A^{\phi} = A^{(G)}$
1: While $\phi \neq 1$:
2: select $N\phi$ nodes labels to be removed: $[i_1,\ldots,i_{N\phi}]$
3: we impose $A^{\phi}_{i_k j} = 0 \quad \forall j \in G, \forall k$ in previous list
4: calculate GC^{ϕ} of the new A^{ϕ}
5: $\phi = \phi + d\phi$

Table 4.1: GC node random removal procedure.

puting the average of $GC(\phi)$ over the network ensemble. The hope is that robustness properties mimic the original network underneath. We obtain the results in Figure 4.4.

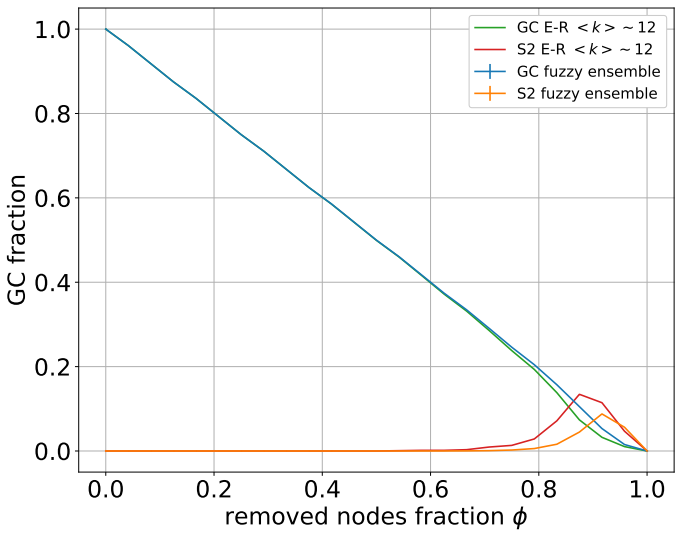


Figure 4.4: The figure shows the comparison between the Erdos-Renyi network with topological descriptors in table 4.2 and the ensemble of 1000 network belonging to the fuzzy ensemble. S2 curves, the second giant components, are multiplied by a 10 factor.

The behaviour of the Giant Cluster by varying the ϕ probability is similar between the reconstruction method and the original Erdos-Renyi network. However, the critical threshold of the uniform removal process is not very well recovered. Then for both networks, we look at the second giant cluster to better analyse percolation process differences. Indeed, percolation theory tells us that we find the critical threshold ϕ_c, where the second giant cluster reaches its maximum.
The fact that the correspondence is not perfect is probably due to the fact that the in-

ferred topological descriptors such as the average degree, the average path length, the assortativity and the clustering are not too well recovered (Table 4.4).

Network	degree $< k >$	d	average path lenght	ass	clustering
Erdos-Renyi	~ 12.65	0.05	2.46	0.034	0.048
1000 ensemble	~ 18.97	0.07	2.21	0.0784	0.088

Table 4.2: Properties of the original Erdos-Renyi network vs the average properties of the reconstructed fuzzy ensemble.

Finally, we are interested in finding a correspondence between the theoretical and computational approaches. We cannot use the "theoretical" algorithm since the computational cost for a network of 256 nodes is too expensive.
Though we cannot compute the $GC(\phi)$ function from eq. (4.21), for our experimental fuzzy network, we can test the correctness of our formula by comparing both the computational and theoretical fuzzy methods for the giant cluster on a smaller synthetic fuzzy network. The theoretical formula to calculate the giant cluster for the fuzzy network ap-

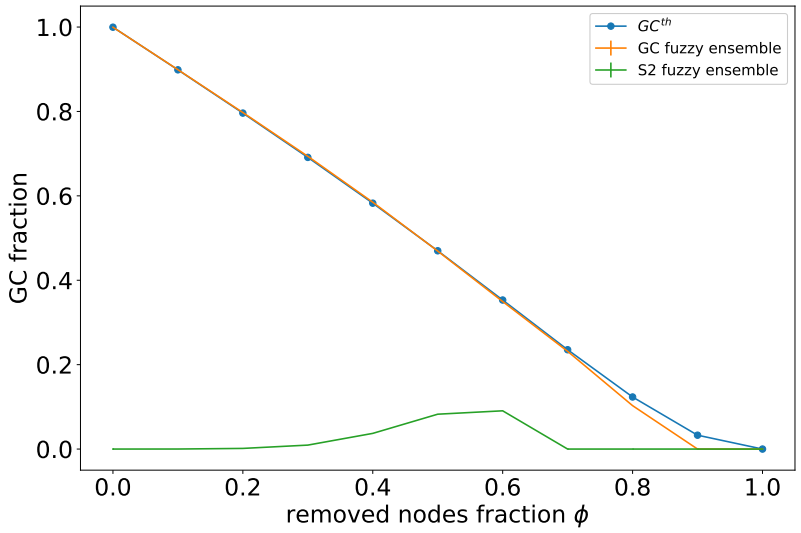


Figure 4.5: Comparison between the random removal of nodes from each network in the fuzzy ensemble and the one given by an Erdos-Renyi network with link probability defined above. The network is composed by 10 vertices and the fuzzy ensemble has 10^3 networks.

proach gives a promising result for a smaller synthetic fuzzy network of 10 nodes (Figure 4.5).
Still, we must work on our algorithm to make it feasible. As a future step, we will look for some approximations that leave out most of the redundant information in Π and, at

the same time, correctly describe the properties of the fuzzy network ensemble by making the computation more efficient.
A possible procedure to start such approximation is taking the fuzzy probability matrix $\Pi = [\pi_{ij}] \quad \forall i, j \in G$ and plotting the distribution of π_{ij}. Then, we can deal with the redundant information in Π and how to approximate the fuzzy network model depending on such distribution shape.

Conclusions

The main goal of this thesis is to analyze the robustness of an interconnected system when we have available information only about its collective dynamics. State of the art methods allow to infer network structure from multivariate time series with several limitations. Here, we explore the new probabilistic network framework proposed in Ref. [42], the so-called fuzzy network model, for which robustness is still an unexplored feature.
We analyze collective dynamics simulated from the stochastic Kuramoto model (SKMN) on Erdos-Renyi networks to compute the fuzzy matrix Π, whose entries define the existence probability of each link of the inferred networks.
From the Π matrix, we extract a statistical ensemble of inferred networks. Our goal is to prove that the percolation process, averaged over this probabilistic ensemble, is representative of the robustness properties of the synthetic network underneath SKMN dynamics. In our discussion, we choose to measure network robustness properties in terms of percolation theory since it provides the most complete way to assess the system's response to failures and perturbations. Our results are promising but, since we dealt only with the Erdos-Renyi networks, we have to test our framework on other network models to gain further insights into the quality of robustness quantification.

The most considerable advantage of fuzzy network modelling is that it avoids the introduction of arbitrary choice to threshold connections in the network inference process. While the most widely used procedure applies a thresholding criterion to decide whether the quantified interdependency between nodes is significant or not, the fuzzy network model overcomes the need to impose a significance level on connection strength and computes π_{ij}, the existence probability of the edge between node i and j.
In this framework, the p_{ij}-values computation is extremely relevant and care must be paid to choosing the null hypothesis to compare.
In the SKMN collective dynamics analysis, we notice that null hypothesis testing involving random permutation (RP) surrogates is too weak in suppressing the pairwise connection among nodes. Conversely, we find that iterative amplitude adjusted Fourier transform (IAFFT) surrogates testing is way too effective in cancelling out the interaction. A possible improvement of this work would involve finding a more coherent hypothesis testing relative to the SKMN dynamics.

Despite the limitations we highlight in our discussion, the thresholding reconstruction

procedure is an important step in assessing the quality of inferred networks depending on parameters set for the SKMN dynamics. In fact, in chapter 2, we set the threshold criterion to obtain reconstructed networks that follow the sparsity ansatz coming from the allometric scaling result. Since the collective dynamics we analyze is characterized by a stochastic term, we use the additional information due to noise to propose two different methods: the before and after reconstruction, respectively, BR and AR. Both strategies involve averaging the noise ensemble dynamics information to reconstruct the topology of interactions. Comparing properties of the BR and AR inferred networks with the equivalent topological descriptors for the original network, results in a clear preference for the BR method over the AR. This result reflects the expected outcome since the BR procedure better suppresses the spurious correlation among time series. BR thresholding procedure applied on the SKMN dynamics on top of Erdos-Renyi networks gives optimal reconstruction parameters for low noise intensity and coupling above the critical σ_c. Besides from the analysis on different time regions of the SKMN dynamics (transient time, onset of synchronization and stationarity), we find that a better reconstruction is achieved just before the collective dynamics has reached synchronization. These results were a fundamental step to the fuzzy network modelling inference.

Several improvements to our approach are possible and worth investigating. For instance, we could refine the statistical similarity measure we use to quantify interdependencies between nodes in our network: in fact, if we opt for a statistical similarity measure different from the basic correlation coefficient, we probably enhance some dissimilarities in the inferred topological structure.
Besides avoiding the introduction of any arbitrary choice, the fuzzy network provides some theoretical insights based on the new formulation where uncertainty about the existence of the edges reflects the uncertainty about the topological descriptors. In our work, we exploit the fuzzy framework to define the excess degree distribution for the single node and to reformulate the network theory percolation problem.
For small networks, the analytical formulas we derive perfectly capture the topological descriptors in the fuzzy model framework. However, in the case of large networks, the computational costs reveal to be prohibitive. Then the next step in our method would involve finding some approximations that leave out most of the redundant information in Π, to correctly describe the properties of the fuzzy network ensemble in a more efficient way.
Concerning the analysis of the stochastic Kuramoto network model discussion, we found an interesting critical behaviour: the order parameter r_∞ dependence on both the coupling and noise intensity. At first sight, this landscape highlights a metacritical point in the portrayed phase transition curve, requiring further analysis.
A future research line could be to investigate further the analytic framework that could include noise, coupling and network topology to explain such phase transition behaviour.
Finally, we conclude by stating the interdisciplinary scope of this study. An example of application of this framework could involve ecology: e.g. we can replace our toy model SKMN dynamics with the generalized Lotka-Volterra function, allowing for the analysis of population evolution on networks. In the next future, a new experimental setting will

provide us with time series involving populations of interacting bacteria: we can then infer the robustness properties of bacteria communities to perturbation through the developed framework.

Nonetheless, the purpose of our framework is wide-ranging since it focuses on the interplay between network structure and dynamics. Indeed, we can substitute the stochastic Kuramoto model on a network with a more general description. Referring to discussion in Chapter 1, let us take the general dynamics of an interconnected system:

$$\dot{\mathbf{x}}_i(t) = \boldsymbol{f}_i(\boldsymbol{x}_i(t)) + \sum_{j=1}^{N} A_{ij}\boldsymbol{g}_{ij}(\boldsymbol{x}_i(t), \boldsymbol{x}_j(t)) + \boldsymbol{I}_i(t) + \boldsymbol{\xi}_i(t) \tag{4.26}$$

Here $\mathbf{x}_i(t)$ is the internal D-dimensional state associated with the i-th unit of a complex system described in the network theory framework. The vector $\mathbf{x}_i(t)$ consists of several observables individuating meaningfully descriptors of the analyzed system.
The possibility of considering $D > 1$ opens quite an exciting future research perspective where we extend our framework to the case of multi-layer networks.

Bibliography

[1] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

[2] Mark Newman. *Networks*. Oxford university press, 2018.

[3] Danielle S Bassett, Perry Zurn, and Joshua I Gold. On the nature and use of models in network neuroscience. *Nature Reviews Neuroscience*, 19(9):566–578, 2018.

[4] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

[5] Ed Bullmore and Olaf Sporns. The economy of brain network organization. *Nature reviews neuroscience*, 13(5):336–349, 2012.

[6] Jingfang Fan, Jun Meng, Yosef Ashkenazy, Shlomo Havlin, and Hans Joachim Schellnhuber. Climate network percolation reveals the expansion and weakening of the tropical component under global warming. *Proceedings of the National Academy of Sciences*, 115(52):E12128–E12134, 2018.

[7] Niklas Boers, Bedartha Goswami, Aljoscha Rheinwalt, Bodo Bookhagen, Brian Hoskins, and Jürgen Kurths. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566(7744):373–377, 2019.

[8] A Ghavasieh and M De Domenico. Statistical physics of network structure and information dynamics. *Journal of Physics: Complexity*, 3(1):011001, 2022.

[9] Changsong Zhou, Adilson E Motter, and Jürgen Kurths. Universality in the synchronization of weighted random networks. *Physical review letters*, 96(3):034101, 2006.

[10] Jürgen Kurths, D Maraun, CS Zhou, G Zamora-Lopez, and Y Zou. Dynamics in complex systems. *European Review*, 17(2):357–370, 2009.

[11] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

[12] Yoshiki Kuramoto. Chemical turbulence. In *Chemical oscillations, waves, and turbulence*, pages 111–140. Springer, 1984.

[13] Steven H Strogatz. From kuramoto to crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1-4):1–20, 2000.

[14] Michael Plischke and Birger Bergersen. *Equilibrium statistical physics*. World scientific, 1994.

[15] Julia M Yeomans. *Statistical mechanics of phase transitions*. Clarendon Press, 1992.

[16] Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International symposium on mathematical problems in theoretical physics*, pages 420–422. Springer, 1975.

[17] Edward A Guggenheim. The principle of corresponding states. *The Journal of Chemical Physics*, 13(7):253–261, 1945.

[18] Nigel Goldenfeld. *Lectures on phase transitions and the renormalization group*. CRC Press, 2018.

[19] Juan A Acebrón, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler. The kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137, 2005.

[20] Alex Arenas, Albert Díaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou. Synchronization in complex networks. *Physics reports*, 469(3):93–153, 2008.

[21] Chen Chris Gong, Chunming Zheng, Ralf Toenjes, and Arkady Pikovsky. Repulsively coupled kuramoto-sakaguchi phase oscillators ensemble subject to common noise. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(3):033127, 2019.

[22] Juan G Restrepo, Edward Ott, and Brian R Hunt. Onset of synchronization in large networks of coupled oscillators. *Physical Review E*, 71(3):036151, 2005.

[23] Mason A Porter and James P Gleeson. Dynamical systems on networks. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, 4, 2016.

[24] Marc Timme and Jose Casadiego. Revealing networks from dynamics: an introduction. *Journal of Physics A: Mathematical and Theoretical*, 47(34):343001, 2014.

[25] Delio Mugnolo. *Mathematical Technology of Networks: Bielefeld, December 2013*, volume 128. Springer, 2015.

[26] Leto Peel, Tiago Peixoto, and Manlio De Domenico. Statistical inference links data and theory in network science. *To appear in Nature Communications*, 2022.

[27] Arsham Ghavasieh and Manlio De Domenico. *To be submitted*, 2023.

[28] Ulrich Brose, Richard J Williams, and Neo D Martinez. Allometric scaling enhances stability in complex food webs. *Ecology letters*, 9(11):1228–1236, 2006.

[29] Daniel M Busiello, Samir Suweis, Jorge Hidalgo, and Amos Maritan. Explorability and the origin of network sparsity in living systems. *Scientific reports*, 7(1):1–8, 2017.

[30] Fabrizio De Vico Fallani, Vito Latora, and Mario Chavez. A topological criterion for filtering information in complex brain networks. *PLoS computational biology*, 13(1):e1005305, 2017.

[31] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

[32] Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3-4):346–382, 2000.

[33] Rainer Hegger, Holger Kantz, and Thomas Schreiber. Practical implementation of nonlinear time series methods: The tisean package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, 1999.

[34] James Theiler and Dean Prichard. Constrained-realization monte-carlo method for hypothesis testing. *Physica D: Nonlinear Phenomena*, 94(4):221–235, 1996.

[35] James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1-4):77–94, 1992; Reprinted in [53].

[36] Thomas Schreiber and Andreas Schmitz. Improved surrogate data for nonlinearity tests. *Physical review letters*, 77(4):635, 1996.

[37] Leonard Mandel and Emil Wolf. Spectral coherence and the concept of cross-spectral purity. *JOSA*, 66(6):529–535, 1976.

[38] Claude E. Shannon and Warren Weaver. *A Mathematical Theory of Communication, Vol. 97.* University of Illinois Press, 1949.

[39] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

[40] Gemma Lancaster, Dmytro Iatsenko, Aleksandra Pidde, Valentina Ticcinelli, and Aneta Stefanovska. Surrogate data for hypothesis testing of physical systems. *Physics Reports*, 748:1–60, 2018.

[41] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[42] Sebastian Raimondo and Manlio De Domenico. Measuring topological descriptors of complex networks under uncertainty. *Physical Review E*, 103(2):022311, 2021.

[43] Thomas Sellke, MJ Bayarri, and James O Berger. Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

[44] Leonhard Held. A nomogram for pvalues. *BMC Medical Research Methodology*, 10(1):1–7, 2010.

[45] Giulio Tirabassi, Ricardo Sevilla-Escoboza, Javier M Buldú, and Cristina Masoller. Inferring the connectivity of coupled oscillators from time-series statistical similarity analysis. *Scientific reports*, 5(1):1–14, 2015.

[46] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 1949.

[47] Jonathan F Donges, Jobst Heitzig, Boyan Beronov, Marc Wiedermann, Jakob Runge, Qing Yi Feng, Liubov Tupikina, Veronika Stolbova, Reik V Donner, Norbert Marwan, et al. Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(11):113101, 2015.

[48] Pafnutij L'vovič Tchébychef. Des valeurs moyennes (traduction du russe, n. de khanikof. *Journal de Mathématiques Pures et Appliquées*, pages 177–184, 1867.

[49] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[50] Frederick Mosteller and R. A. Fisher. Questions and answers. *The American Statistician*, 2(5):30–31, 1948.

[51] Filippo Radicchi and Claudio Castellano. Beyond the locally treelike approximation for percolation on real networks. *Physical Review E*, 93(3):030302, 2016.

[52] Bnaya Gross and Shlomo Havlin. Percolation in spatial networks: Spatial network models beyond nearest neighbours structures. *Elements in Structure and Dynamics of Complex Networks*, 2022.

[53] Edward Ott, Tim Sauer, and James A Yorke. Coping with chaos. analysis of chaotic data and the exploitation of chaotic systems. *Wiley Series in Nonlinear Science*, 1994.