Università degli Studi di Salerno

.DIEM

Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata

Dottorato di Ricerca in Ingegneria dell'Informazione
Ciclo 35

TESI DI DOTTORATO / PH.D. THESIS

Learning Preferential Attachment Graphs Under Partial Observability

MICHELE CIRILLO

SUPERVISOR: **PROF. VINCENZO MATTA**

PHD PROGRAM DIRECTOR: **PROF. PASQUALE CHIACCHIO**

Anno Accademico 2022/2023

Dedicated to my family

Contents

Abstract

This thesis work deals with the problem of learning the topology of a network starting from the signals emitted by the network nodes while executing some distributed processing task. In particular, these signals are generated over time through a vector autoregressive process, which is a linear diffusion process where each node exchanges messages with its neighbors (therefore, according to the underlying network graph) and aggregate them according to a certain combination policy. We consider the demanding setting of graph learning under partial observability, where only part of the nodes can be probed, and we study under which conditions the subgraph relative to the probed nodes can be correctly estimated. This is a challenging problem since the observed signals are also influenced by the presence of latent nodes, whose signals act as noise and can in principle prevent faithful graph reconstruction. In particular, we consider two fundamental questions. The first question is about achievability: Under which conditions the graph learning problem can be solved? Namely, for meaningful classes of graphs, is there a graph estimator which, starting from signals of the probed nodes, is able to recover the related subgraph? Usually, a positive answer for achievability only says that one or more estimators exist, and that it works for a sufficiently large number of samples. Therefore, the second question arises in the context of sample complexity: Given a graph estimator found by the achievability analysis, how many samples it needs to work properly in practice? Recent results in the literature examined this problem when the underlying graph is generated according to an Erdős-Rényi random model. The main limitation of this assumption is that Erdős-Rényi graphs use a simple construction mechanism that produces (sometimes unrealistic) homogeneous networks with independent edges. We overcome this issue by solving the problem of graph learning over preferential attachment graphs, which are characterized by a large heterogeneity, featuring both very connected nodes (which model real-world network "hubs") and peripheral nodes having few connections. Moreover, preferential-attachment graphs enforce a strong dependence between the edges of the graph. These are important properties that can be observed in real-world networks. In particular, our main contribution examines the case where a first-order vector autoregressive process, equipped with a stable Laplacian combination matrix, is run over the a graph drawn according to the popular Bollobás-Riordan preferential attachment model. In this thesis we first introduce a unifying framework for graph learning under partial observability. This framework covers in particular the previous results on Erdős-Rényi graphs and our

novel results on Bollobás-Riordan graphs. The main achievability result established in the present thesis is that a combination matrix estimator known as Granger estimator achieves graph learning under partial observability. We also characterize the sample complexity over Bollobás-Riordan graphs, establishing that it is essentially linear in the network size. Comparing this result with what was observed before for Erdős-Rényi graphs, we obtain the following interesting classification: i) dense Erdős-Rényi graphs require the highest sample complexity rate, where the number of samples scales quadratically as the network size grows; ii) the intermediate sample complexity rate is given by Bollobás-Riordan graphs, which require an almost-linear sample scaling law; finally, iii) sparse Erdős-Rényi graphs have a lighter, sublinear sample scaling law.

Notation

Matrices are denoted by upper-case letters, vectors by lower-case letters. We use boldface font to denote random variables, and normal font for their realizations. Sets and graphs are denoted by upper-case calligraphic letters. For an $N \times N$ matrix Z, the submatrix spanning the rows of Z indexed by set $\mathcal{P} \subseteq \{1, 2, \ldots, N\}$ and the columns indexed by set $\mathcal{T} \subseteq \{1, 2, \ldots, N\}$, is denoted by $Z_{\mathcal{P}\mathcal{T}}$, or alternatively by $[Z]_{\mathcal{P}\mathcal{T}}$. When $\mathcal{P} = \mathcal{T}$, the submatrix $Z_{\mathcal{P}\mathcal{T}}$ is abbreviated as $Z_{\mathcal{P}}$. Moreover, in the indexing of a submatrix we keep the index set of the corresponding full matrix. For example, if $\mathcal{P} = \{2, 3\}$ and $\mathcal{T} = \{2, 4, 5\}$, the submatrix $M = Z_{\mathcal{P}\mathcal{T}}$ is a 2×3 matrix, indexed as follows:

$$M = \begin{pmatrix} z_{22} & z_{24} & z_{25} \\ z_{32} & z_{34} & z_{35} \end{pmatrix} = \begin{pmatrix} m_{22} & m_{24} & m_{25} \\ m_{32} & m_{34} & m_{35} \end{pmatrix}. \tag{1}$$

For a graph $\mathcal{G}$, the corresponding capital letter G is used to denote its adjacency matrix, which has zero diagonal, and whose off-diagonal (k, ℓ)-entry $g_{k\ell}$ is equal to 1 if a directed edge from ℓ to k exists, and is zero otherwise. The symbol $\|\cdot\|_{\max}$ computes the maximum absolute entry of its matrix argument, whereas the symbol $\|\cdot\|_{\text{max-off}}$ computes the maximum absolute *off-diagonal* entry of its matrix argument. The symbol $\xrightarrow{\text{p}}$ denotes convergence in probability *as the network size scales to infinity.* Likewise, the symbol $\xrightarrow{\text{a.s.}}$ denotes almost-sure convergence.

Chapter 1

Introduction and Problem Formulation

The present work deals with distributed systems made by a large number of units, such as devices of a communication network, sensors in a monitoring system, or individuals in a social network. These units are allowed to interact over time and, when considered together, can give rise to sophisticated dynamics. Usually, an individual unit is not able to interact directly with all the other units in the system. Instead, the single units form a network wherein each of them is allowed to reach a limited number of units, i.e., its *neighbors*. Thus, the interactions across the network consists of *local* information exchanges, and the ensemble of local effects gives rise over time to a global, decentralized system dynamics.

There are many notable examples of complex systems that derive their sophistication from coordination among simpler units and from the aggregation and processing of decentralized pieces of information. Relevant examples of these systems are telecommunication and computer networks. These types of systems are the natural environment to run distributed algorithms. Nature itself provides beautiful examples of distributed systems. Discoveries in biological sciences have revealed remarkable patterns of organization and structured complexity in the behavior of animal groups [1] and in the dynamics of brain connectivity [5]. Motivated by the aforementioned reasons, in recent years, many efforts have been devoted to achieve a deeper understanding of information processing, adaptation, and learning over complex networks in several disciplines, including machine learning, optimization, control, economics, biological sciences, information sciences, and social sciences [103].

Multi-agent networks can be employed to solve complex problems that would be unaffordable by a stand-alone processing unit (e.g., for lack of resources) performing a centralized algorithm. In particular, they have been considered to solve demanding problems in the context of optimization, learning, and inference [14, 30, 33, 55, 83, 102, 103, 117, 120].

Moreover, decentralized solutions exhibit undisputed advantages in terms of scalability, resilience to failures and robustness.

In a decentralized system, the network topology plays a critical role in enabling the interactions among individuals: while each unit in these systems is not capable of sophisticated behavior on its own, it is the interaction among the constituents that leads to systems able to accomplishing complex tasks, with even an impressive capability of adjusting their behavior in response to changes in the environment [103].

Many literature works examine the *direct* learning problem, i.e., given a certain topology, how the network agents are able to solve the assigned learning task [13, 19, 20, 59, 83–85, 87, 94, 102–104, 110, 115–119]. In this thesis we focus instead on the fundamental *inverse* learning problem: Given the output signals produced by the network agents during the accomplishment of their (direct) learning task, can we infer the network structure? Our final aim is to reconstruct the underlying *network graph* determining the interaction pattern among the probed agents. Due to the emphasis on the network structure, in our treatment we will refer to the individual units of the distributed system as *network nodes*. Since the addressed problem arises across multiple disciplines, it is referred to in different ways. The various terminologies include graph learning, topology inference, network tomography, graph reconstruction, and graph estimation. In this work we will mostly use "graph learning."

This is a problem of fundamental importance, which can provide answers to many useful questions arising across several disciplines. For example, by observing the local dynamics at a subset of the nodes, can one establish how the information is shared across the network? Or how privacy is reflected in the nodes' signals? Can one reconstruct how a given information propagates across the network nodes? Can one discover whether there are some "influential" nodes which directly affect the behavior of large portions of the network?

Providing answers to these questions would be beneficial for a large number of applications. For example, discovering who is communicating with whom over the Internet is crucial in several cybersecurity applications [41, 66, 92, 113]. As another example, one can study the mechanism of opinion formation over a social network, or attempting to locate the source of fake news [64, 72]. Moreover, with a graph learning tool it is possible to characterize the evolution of urban traffic in large cities [31], learning the synchronized cognitive behavior of a school of fish which is escaping from a predator [29, 89], and investigating the connectivity patterns within the brain [60].

In this work we address the graph learning task under some demanding conditions. First, we consider the setting of *partial observability*, where only a limited subset of nodes can be accessed. The goal is to infer the topology linking the probed nodes This setting arises very often, especially over large networks, as it is usually not possible to gather information from all network nodes.

A second important element of novelty of this thesis is that we consider *preferential attachment* graphs to model the network structure. These are *random* graph models that, at the price of introducing significant sophistication in the network formation process and in the technical analysis, are able to capture important features observed over real-world

graphs. For example, they exhibit node heterogeneity (i.e., "hubs" with many connections as opposed to peripheral nodes with few connections) and edge dependence (i.e., edges are not drawn independently as is the case for the popular Erdős-Rényi random graphs).

In the remainder of this chapter we will present the relevant modeling assumptions in full detail. In Section 1.1 we formalize the dynamical system in terms of a discrete-time, linear diffusion system, namely, a first-order vector autoregressive model. Section 1.2 introduces the partial observability setting. In Section 1.3 we introduce the two fundamental questions addressed in the analysis: Is graph learning *achievable* in the considered setting? If yes, how many samples are required, i.e., which is the *sample complexity* of the graph learning task? Finally, in Section 1.4 we describe the overall organization of the thesis.

1.1 First-Order Vector Autoregressive Dynamics

A graph is defined by an ensemble of nodes and edges. In the general formulation, we consider *directed* graphs. Later, when focusing on specific graph models in Chapters 4 and 5, we will consider *undirected* graphs. Over a directed graph, given any two nodes k and ℓ, they can be disconnected (no edge between them), connected in one direction (e.g., from k to ℓ or from ℓ to k), or they can be connected in both directions (two directed edges). We consider a *random* graph defined over N nodes and denoted by $\boldsymbol{\mathcal{G}}(N)$ (bold notation highlights graph randomness). The qualification "random" signifies that connections between nodes are drawn according to some probabilistic mechanism. The graph structure can be conveniently encoded into an *adjacency matrix* $\boldsymbol{G}(N)$. This matrix has all zeros on the main diagonal, whereas its (k,ℓ) entry $\boldsymbol{g}_{k\ell}(N)$ is equal to 1 if there is an edge from ℓ to k, and is 0 otherwise.

Given a certain graph, the actions of the network nodes are described by a *distributed* linear dynamical system. Every node k, at time $t = 1, 2, \ldots$, is driven by a random input source $\boldsymbol{x}_{k,t}(N)$ and produces the output signal $\boldsymbol{y}_{k,t}(N)$ according to the following diffusion model, a.k.a. first-order *vector autoregressive* model [63]:

$$\boldsymbol{y}_{k,t}(N) = \sum_{\ell=1}^{N} \boldsymbol{a}_{k\ell}(N)\boldsymbol{y}_{\ell,t-1}(N) + \boldsymbol{x}_{k,t}(N), \tag{1.1}$$

which can be conveniently recast in matrix form as:

$$\boldsymbol{y}_t(N) = \boldsymbol{A}(N)\boldsymbol{y}_{t-1}(N) + \boldsymbol{x}_t(N), \tag{1.2}$$

where $\boldsymbol{x}_t(N)$ and $\boldsymbol{y}_t(N)$ stack the entries $\boldsymbol{x}_{k,t}(N)$ and $\boldsymbol{y}_{k,t}(N)$ into $N \times 1$ column vectors, and where matrix $\boldsymbol{A}(N) = [\boldsymbol{a}_{k\ell}(N)]$ collects the nonnegative combination weights $\boldsymbol{a}_{k\ell}(N)$.[1] This model has many applications in several fields. For example, in economics

[1] Making explicit the dependence of the combination weights upon N is critical in our treatment, since we need to examine the properties of these weights and related network descriptors as functions of N.

it is used for time-series forecasting of financial data [48]. It is also employed in bioinformatics and biostatistics for estimating gene-regulatory networks from gene expression data [40]. Moreover, the vector autoregressive model governs several distributed algorithms over networks aimed at solving inference tasks, such as distributed detection problems [16, 76].

In our setting, the eigenvalues of $\boldsymbol{A}(N)$ are assumed to lie strictly inside the unit circle to ensure that system (1.1) is Schur stable. The combination matrix $\boldsymbol{A}(N)$ reflects the interconnections dictated by graph $\mathcal{G}(N)$. Thus, weight $\boldsymbol{a}_{k\ell}(N)$ is strictly positive if there is an edge from ℓ to k, and is zero otherwise. In view of (1.1), this structure implies that node k at time t updates its state $\boldsymbol{y}_{k,t}(N)$ by incorporating only previous-time signals $\boldsymbol{y}_{\ell,t-1}(N)$ received from nodes ℓ for which $\boldsymbol{a}_{k\ell}(N) > 0$. In general, $\boldsymbol{A}(N)$ need not be symmetric. For example, we could have $\boldsymbol{a}_{k\ell}(N) > 0$ and $\boldsymbol{a}_{\ell k}(N) = 0$. Accordingly, when we talk of "connected/disconnected pairs", we refer to *ordered* pairs with (k, ℓ) being distinct from (ℓ, k).

The stochastic dynamical system in (1.1) contains different sources of randomness. All involved random variables are assumed to lie in a common probability space $(\Omega, \mathcal{F}, P)$. One source of randomness is given by the sequence of random graphs $\mathcal{G}(N)$, for $N = 1, 2, \ldots$, while the combination matrix $\boldsymbol{A}(N)$ is a deterministic function of the graph $\mathcal{G}(N)$. Thus, once a graph realization is fixed, the combination matrix becomes deterministic, and the system in (1.1) evolves according to the randomness of the input signals $\boldsymbol{x}_{k,t}(N)$, which are independent and identically distributed (i.i.d.) w.r.t. to node index k, time index t, and network size N. These signals are statistically independent of the sequence of graphs, and, without loss of generality, are assumed to have zero mean and unit variance. The vectors $\boldsymbol{y}_0(N)$ that initialize the recursion (1.1) are assumed to be square-integrable random vectors with arbitrary distribution, independent of all input signals $\boldsymbol{x}_{k,t}(N)$. They are allowed to depend only on $\mathcal{G}(N)$ and, conditionally on $\mathcal{G}(N)$, they are independent of all other graphs in the sequence. The particular distribution of $\boldsymbol{y}_0(N)$ will be mostly immaterial for our results, since we will be dealing with the steady-state regime where the number of samples goes to infinity and the initial state does not play a role. Only when we study the sample complexity, we will assume a specific distribution for the initial state — see Theorem 8 in Chapter 6.

1.2 Partial Observability

Most of the earlier works on graph learning assume that all nodes in a network are monitored. We will refer to this condition as *full observability*. However, this condition is seldom verified. For example, in probing signals from the brain, usually only certain regions can monitored. Also, monitoring a social network with millions of members is possible only by limiting the observation scope. Over these networks, due to different forms of physical limitations, it is not practical to assume that data can be collected from all nodes. We refer to this condition as *partial observability*. Partial observability makes graph learning task more demanding. In fact, the observations collected at the monitored

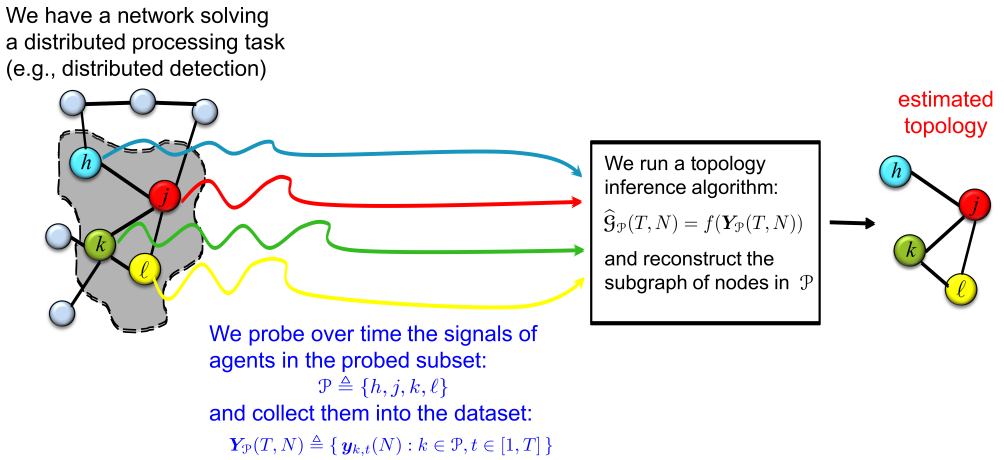

Figure 1.1: A graphical sketch of the graph learning problem under partial observability. The input of the problem is constituted by the signals generated by the subset of probed nodes, and the learning problem amounts to use this information to discover the underlying topology governing the interactions of the observed nodes.

nodes are influenced (through information propagation) by the unobserved nodes, which act as source of noise. It is then natural to ask whether the graph learning problem is well-posed under partial observability, namely, if we can collect sufficient information to learn the underlying graph linking the probed nodes. This is a hard problem, which could be unfeasible in general.

Figure 1.1 provides a schematic illustration of graph learning under partial observability. Given a network graph, a signal evolving over time is associated with each node of the graph. The time evolution of the signals is dictated by an exogenous random mechanism (driving noisy source) and by the local interactions between neighboring nodes. In particular, in our case we consider the first-order vector autoregressive dynamics in (1.2). An inferential engine collects signal samples from a limited set of probed nodes, since we work in the regime of partial observability. The inference goal is to learn the *subgraph* linking the probed nodes.

Letting $\mathcal{P}$ be the *probed subset,* i.e., the subset containing only the nodes that can be probed, and given an observation time window $t = 1, 2, \ldots, T$, the collection of signals available to perform graph learning will be compactly denoted by:

$$\boldsymbol{Y}_{\mathcal{P}}(T,N) \triangleq \{\, \boldsymbol{y}_{k,t}(N) : k \in \mathcal{P}, t \in [1,T] \,\}. \tag{1.3}$$

Our goal is to estimate the interconnections between nodes in $\mathcal{P}$, namely, the topology of the partial graph $\mathcal{G}_{\mathcal{P}}(N)$ relative to $\mathcal{P}$, starting from the signals in (1.3). Formally, we

need to build a *graph estimator*:[2]

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T,N) = f(\boldsymbol{Y}_{\mathcal{P}}(T,N)), \tag{1.4}$$

where in the notation we emphasized that the properties of the graph estimator will depend on the number of samples and the network size. We will judge the goodness of a graph estimator in the asymptotic framework where one specifies the functional dependence of the number of samples T upon the network size N through a sample law T_N, and lets the network size N go to infinity. In this analysis, we allow the probed subset to depend on N, namely, we introduce a deterministic *sequence* of subsets:

$$\mathcal{S}_N \subseteq \{1,2,\dots,N\}, \tag{1.5}$$

where $\mathcal{S}_N$ is the probed subset of the graph of size N. For ease of notation, when a graph on N nodes or an $N \times N$ matrix is evaluated over subset $\mathcal{S}_N$, subscript N will be omitted, for example, we will write $\mathcal{G}_{\mathcal{S}}(N)$ in place of $\mathcal{G}_{\mathcal{S}_N}(N)$.

1.3 Achievability and Sample Complexity Analysis

The evolution of the signals $\boldsymbol{y}_{k,t}(N)$ is dictated by repeated interactions between neighboring nodes, and these interactions are determined by the graph topology. Thus, it is legitimate to ask whether the topology can be inferred from observing the evolution of the signals at the nodes.

We say that a graph estimator is *consistent* for the family of random graphs $\mathcal{G}(N)$ and for the sequence of probed subsets $\mathcal{S}_N$ if:

$$\lim_{N\to\infty} \mathbb{P}\left[\widehat{\mathcal{G}}_{\mathcal{S}}(T_N,N) = \mathcal{G}_{\mathcal{S}}(N)\right] = 1, \tag{1.6}$$

for some sample scaling law T_N. Accordingly, we say that the graph learning problem is *achievable* if there exists a consistent graph estimator.

The law T_N characterizes the so-called *sample complexity* of the estimator, namely, how the number of samples scales with the network size to provide faithful graph learning *using that estimator*. The study of the sample law T_N, and in particular the quest for finding the slowest growth rate ensuring achievability (possibly considering a fixed graph estimator), will be referred to as *sample complexity analysis*. This analysis is relevant in practical applications, since the amount of available data is determined by several physical constraints. For example, the dynamical system can be observed only over a certain time interval, the data acquisition rate is limited, and there are limitations in terms of energy, bandwidth or storage capacity.

[2]Formally, function f in (1.3) is allowed to depend on T, N, and $\mathcal{P}$. This dependence is left implicit for ease of notation.

1.4 Thesis Overview

This thesis is organized as follows. In the next chapter we will make an analysis of the related work. In Chapter 3 we propose a unifying framework to describe and certify when graph learning is achievable. We will introduce relevant descriptors and notions such as the bias and the identifiability gap, which will be then useful to address the two fundamental questions of achievability and sample complexity. This framework paves the way for examining specific scenarios of interest. More precisely, by "specific scenario" we mean a set of assumptions that *i*) fully characterize the vector autoregressive system (1.2) by specifying the random model for the network graph $\mathcal{G}(N)$ and the combination policy to build the matrix $\boldsymbol{A}(N)$; and *ii*) describe the graph learning strategies $\widehat{\mathcal{G}}_{\mathcal{P}}(T, N)$ that can be adopted, and the particular observability scenario in terms of the sequence of probed subsets $\mathcal{S}_N$.

One relevant scenario that has been recently considered is based on the Erdős-Rényi random graph model [73–75, 77, 78, 100]. The achievability and sample-complexity results for this model will be described in Chapter 4. As said before, this thesis focuses instead on a more sophisticated graph construction, the preferential attachment construction. In particular, we consider the Bollobás-Riordan graph model, which will be examined in detail in Chapters 5 and 6. Finally, in Chapter 7, we apply the theoretical analysis to both synthetic and real data. Moreover, we also test two useful extensions not covered by the theoretical analysis, namely, directed preferential-attachment graphs and dynamic graphs. The main publications relative to the content of the present thesis are [25–27].

Chapter 2

Related Work

There is a large body of literature that examines the problem of graph learning. In this chapter, we examine the works most relevant to our treatment. In order to facilitate the illustration, it is useful to classify the pertinent works in terms of three main features.

- *Full vs. partial observability.* Namely, whether the input of the graph learning problem is a function of all network nodes, or of a limited subset thereof.
- *Static vs. dynamical systems.* Some works examine graph learning over static systems, like graphical models, where the data collected from the graph nodes do not evolve according to a dynamical model. They have a joint distribution that does not depend on time, and the inferential engine is assumed to collect independent realizations of these data [2]. Other works focus instead on dynamical situations, like the vector autoregressive model in (1.2) or typical systems studied in graph signal processing [98, 106].
- *Sample complexity.* Many works consider the achievability issue, by establishing whether the graph can be faithfully estimated provided that an unspecified, sufficiently large number of samples can be collected. However, only some of these works additionally address the sample complexity issue, by providing formal results on how many samples are required to estimate the graph. Typically, these results establish the number of samples necessary to get some prescribed accuracy as the network size grows, and therefore they typically involve a *large-scale,* asymptotic analysis. This is coherent with our definition of sample complexity provided in Chapter 1, whose aim is to characterize the sample law T_N required by a graph estimator to achieve graph learning as $N \to \infty$.

In Table 2.1 we report the works that will be examined in this chapter, organized in a convenient taxonomy based on the aforementioned three features. Along the rows of the table, the references are divided in two groups according to whether they assume full observability or partial observability. Along the columns, the references are sliced according

	Static networked systems	**Dynamical networked systems**	
	With sample complexity analysis		**Without sample complexity analysis**
Full observ.	Gaussian graphical models with local separation [2]	Graph signal processing [80, 90, 106] Parameter estimation in autoregressive systems [49, 95] Continuous-time stochastic differential systems [6]	Stochastic processes interrelated by self-kin networks [69] Nonlinear dynamic systems [44]
Partial observ.	Gaussian graphical models with "sparsity & low-rank" assumption [18] Restricted Boltzmann machines with bounded degree [15] Locally tree-like graphs with bounded girth [3] Link prediction strategies with similarity indices [62, 123]	Latent graphical models with polytrees [37] Continuous-time stochastic differential systems with local-global structure [51] *Erdős-Rényi model* [22, 73–75, 77, 78, 100] *Bollobás-Riordan model* [25–27] Link prediction strategies on dynamic graphs [52, 88]	Parameter estimation in autoregressive systems [42] Stochastic processes interrelated by polytree [70] or loopy [71] networks

Table 2.1: Taxonomy of existing works addressing the graph learning problem. Across rows, the works are separated according to the observability regime: the first row contains works examining graph learning under the full observability regime, whereas the second row contains works assuming the partial observability regime. Across columns, the same works are divided in works on static and dynamical networked systems, and are further sliced according to whether they include a sample complexity analysis or only study the problem feasibility. We use italic font to identify the works relative to the setting adopted in this thesis. These include both previous literature works based on the Erdős-Rényi model, and the works collecting the results of the present thesis, which consider the Bollobás-Riordan model.

to two criteria, which are identified by the two pairs of columns headers. According to the first criterion, the works are partitioned in two groups, depending on the kind of dynamics involved in the assumed model: i) A group of works where no dynamics occurs, which

includes some works in the context of high-dimensional graphical models; ii) a group of works assuming a networked system evolving according to a prescribed dynamics. According to the second criterion, Table 2.1 organizes the works in other two groups. It distinguishes between: i) Works that carry out a sample complexity analysis in addition to achievability; ii) works where only the achievability issue is examined.

In this taxonomy, we use italic font to identify the works that use the setting illustrated in Chapter 1. Each work in the table is identified by its bibliography item and a brief text description. The text description tries to capture the most peculiar features of the work, for example, the specific graph model and/or the kind of dynamics.

The remainder of this chapter is organized as follows. In Sections 2.1 and 2.2 we provide an essential survey of the literature works in Table 2.1. In particular, in the former section we discuss the works assuming full observability and in the latter we do the same for the works assuming partial observability. Finally, in Section 2.3 we highlight the main elements of novelty of our work.

2.1 Graph Learning Under Full Observability

Even if the focus of our work is on partial observability, we start with an illustration of relevant works on graph learning under full observability. This is useful since the works on full observability introduce some fundamental principles and methods that constitute a useful basis/reference necessary before tackling the more challenging setting of partial observability.

Gaussian graphical models. Given a graph $\mathcal{G}$ of N nodes, a Gaussian graphical model on $\mathcal{G}$ is the family of multivariate Gaussian distributions on $\mathbb{R}^N$ with concentration matrix (i.e., the inverse of the covariance matrix) $J_{\mathcal{G}}$ satisfying:

$$[J_{\mathcal{G}}]_{k\ell} = 0 \iff \text{nodes } k \text{ and } \ell \text{ are disconnected in } \mathcal{G}. \tag{2.1}$$

Hence, the topology of $\mathcal{G}$ describes the sparsity pattern of the concentration matrix. The graph learning problem addressed in the context of graphical models can be described as follows. One observes a stream of i.i.d. realizations of a multivariate Gaussian satisfying (2.1) for a certain underlying graph $\mathcal{G}$. By exploiting the fact that the data distribution depends on $\mathcal{G}$, the problem of estimating $\mathcal{G}$ from the collected samples is considered.

In [2], the graph learning problem is addressed for a specific class of graphical models fulfilling a structural constraint named *local separation*. This constraint enforces a homogeneous sparsity across the graph by imposing high distances (path lengths) among the nodes of the graph. Technically speaking, a graph $\mathcal{G}$ satisfies the local separation property with parameters η and γ when for any couple of disconnected nodes (k, ℓ) in $\mathcal{G}$ the distance between k and ℓ can be made higher than γ by removing at most η nodes from $\mathcal{G}$. The parameter η is strictly related to the number of paths from k to ℓ with length less or equal to γ.

Under some technical assumptions including the local separation property, the Authors of [2] prove the existence of an algorithm which is able to recover $\mathcal{G}$ provided that a

sufficiently high number of samples is available. The analysis is conducted under the asymptotic setting where the graph size N tends to infinity, and establishes how the number of collected samples must scales with N.

The aforementioned problem falls in the class of full observability since the algorithm receives as inputs the complete (N-dimensional) realizations, and tries to recover the entire graph $\mathcal{G}$. Additionally, the observations collected for graph recovery are assumed to be i.i.d. rather than arising from a dynamical model with memory, as in the model illustrated in Chapter 1. In this respect, there are several other works that address the graph learning problem for graphical models in the setting of partial observability, and we will discuss these results in Section 2.2.

Graph signal processing. Works [90,106] arise in the context of graph signal processing [67,68,91,98,99,107,111], a recent framework that extends classical signal processing tools and operations (e.g., Fourier transform, sampling, filtering) to signals defined on a graph. In [106] the Authors study the discrete-time dynamics:

$$\boldsymbol{y}_t = (I - \alpha_t L_{\mathcal{G}})\, \boldsymbol{y}_{t-1}, \tag{2.2}$$

where $L_{\mathcal{G}}$ is the $N \times N$ Laplacian matrix of graph $\mathcal{G}$ and α_t is a time-varying scalar function. The proposed approach consists in learning the graph $\mathcal{G}$ by first estimating the Laplacian matrix $L_{\mathcal{G}}$. The input at disposal for this task is obtained by observing several *different* runs of the dynamics (2.2), all evolving over the same network, and collecting one sample for each run. Each run may have different values of the function α_t and the relative sample is taken at different, unknown times. The problem of recovering matrix $L_{\mathcal{G}}$ from these "snapshots" is underdetermined. To overcome this issue the Authors of [106] propose a strategy based on a regularized convex optimization problem. The idea is to find, among the candidate Laplacian matrices that are consistent with the observations, the one that maximizes an objective function promoting sparsity. It is shown in [106] that the proposed strategy ensures consistent estimation of the graph as the number of measurements grows.

In [90] the Authors examine the discrete-time dynamics:

$$\boldsymbol{y}_t \;=\; W_{\mathcal{G}}\, \boldsymbol{y}_{t-1}, \tag{2.3}$$

where $W_{\mathcal{G}}$ is a $N \times N$ matrix with nonnegative entries and spectral radius equal to 1; of particular interest in this analysis is the case:

$$W_{\mathcal{G}} = D_{\mathcal{G}}^{-\frac{1}{2}} G D_{\mathcal{G}}^{-\frac{1}{2}}, \tag{2.4}$$

where $D_{\mathcal{G}}$ is the diagonal matrix whose nonzero entries are the degrees of the nodes in $\mathcal{G}$. The formulation of the graph learning problem is as follows. We are allowed to observe several runs of the dynamical system defined by (2.3), and for each run we sample only one value, at a time that is unknown and that can differ across the experiments. The final aim is to estimate $W_{\mathcal{G}}$. Again, this problem is highly undetermined since, even assuming an infinite number of samples at disposal, the set of matrices $W_{\mathcal{G}}$ is infinite. In particular,

it is possible to show that the set of admissible matrices can be expressed in closed form as a convex polytope. Therefore, a two-step inferential process is proposed. First, the aforementioned polytope is approximated using the (finite) samples at disposal. Then, the final estimator for $W_{\mathcal{G}}$ is selected from this candidate pool according to a suitable selection criterion. For example, the *sparsity criterion* extracts from the polytope the matrix with minimum total sum of its entries, and this consists in practice in solving a linear programming problem.

A common feature of the works examined in this section is that the considered graph estimation problem is undetermined, namely, the data at disposal are not sufficient to determine a unique estimate of $\mathcal{G}$. In other words, there are several candidate solutions explaining the measurements. To overcome this issue, the strategies in [90, 106] exploit some prior knowledge about the graph structure under examination, which is then translated into appropriate structural constraints. We have already mentioned sparsity constraints. Another typical constraint is on the *smoothness* of the graph signal [53, 106]. In a nutshell, a graph signal is smooth when the signals indexed by nearby nodes are close to each other. In contrast, the framework considered in the present thesis do not impose such constraints.

A more general system has been proposed in [80], which considers the following autoregressive dynamics:

$$\boldsymbol{y}_t = \sum_{j=1}^{m} p_j(W_{\mathcal{G}})\, \boldsymbol{y}_{t-j} + \boldsymbol{x}_t, \qquad \text{for } t = 1, 2, \ldots, \tag{2.5}$$

where $p_j(W_{\mathcal{G}})$ are matrix polynomials of an $N \times N$ matrix $W_{\mathcal{G}}$ having support graph $\mathcal{G}$, and $\boldsymbol{x}_t$ is a random noise process.

In [80] the aim is to estimate matrix $W_{\mathcal{G}}$ from the output signals $\boldsymbol{y}_t$ collected over a certain observation window.

Parameter estimation in vector autoregressive systems. There are several works that try to estimate the parameters of vector autoregressive systems. Even if these works are not originally focused on graph learning, they can be in principle adopted for this purpose on systems like the one considered Chapter 1. In fact, the support graph (i.e., the nonzero entries) of the combination matrix A coincide with the underlying network graph. Therefore, once we have an estimated combination matrix $\widehat{A}$, it makes sense to estimate the network graph by classifying the entries of $\widehat{A}$ in some suitable way to distinguish connected/disconnected node pairs.

Before starting our discussion, it is worth recalling that for first-order vector autoregressive systems, there exists a well-known relation that relates the combination matrix A, the steady-state covariance matrix R_0 and the one-lag covariance matrix of the system [63]:

$$R_1 = AR_0. \tag{2.6}$$

Therefore, when addressing the estimation of A in the full observability regime a matrix estimator is promptly available:

$$\widehat{A} \triangleq \widehat{R}_1 \widehat{R}_0^{-1}, \tag{2.7}$$

where $\widehat{R}_j$ is the sample covariance relative to R_j, for $j \in \{0, 1\}$. This estimator is also known as Granger estimator or predictor, a terminology that arises in the context of Granger causality [45].[1]

In [49] the Authors focus on estimating the transition matrices $A_1, A_2, \ldots, A_m$, of the general stationary vector autoregression of order m:

$$\boldsymbol{y}_t = \sum_{j=1}^{m} A_j \boldsymbol{y}_{t-j} + \boldsymbol{x}_t, \tag{2.8}$$

were $\boldsymbol{x}_t$ is a Gaussian process. The Authors provide results for both achievability and sample complexity analysis. When considering the case $m = 1$, the proposed estimator $\widehat{A}_1$ is the solution to an optimization problem whose constraints enforce an approximate version of relation (2.6), by imposing an upper bound:

$$\|A\widehat{R}_0 - \widehat{R}_1\|_{\max} < \lambda \tag{2.9}$$

where A is the generic candidate solution for $\widehat{A}_1$, and $\widehat{R}_0$ and $\widehat{R}_1$ are sample versions of R_0 and R_1, respectively. In particular, under the assumption that A_1 belongs to the following class:

$$\max_{\ell=1,2,\ldots,N} \sum_{k=1}^{N} |[A_1]_{k\ell}|^q \leq \beta, \qquad \text{and} \qquad \|A_1\|_1 \leq \gamma(T, N), \tag{2.10}$$

for some constants $0 \leq q < 1$ and $\beta > 0$, and some function $\gamma(T, N)$, the Authors are able to characterize the estimation errors $\|\widehat{A}_1 - A_1\|_1$ and $\|\widehat{A}_1 - A_1\|_{\max}$ — relative to the ℓ_1 matrix norm and the max matrix norm, respectively — in terms of T, N and $\gamma(T, N)$, and consequently are able to derive the sample complexity of their estimation algorithm.

In [95], the problem of learning the parameters of a vector autoregressive process is addressed under the assumption that the probed data are *corrupted*. In particular, the considered system is:

$$\boldsymbol{y}_t = A\boldsymbol{y}_{t-1} + \boldsymbol{x}_t, \tag{2.11}$$

with $\boldsymbol{x}_t$ a random noise process, and the collected data are not realizations of the output process $\boldsymbol{y}_t$ but instead they are realizations of:

$$\boldsymbol{z}_t = \boldsymbol{P}_t \left(\boldsymbol{y}_t + \boldsymbol{u}_t \right), \tag{2.12}$$

where $\boldsymbol{P}_t$ is a random diagonal matrix and $\boldsymbol{u}_t$ is a random additive noise process. The mapping $\boldsymbol{y}_t \mapsto \boldsymbol{z}_t$ models either missing data or distortion effects due to imperfect measurements. For example, the diagonal entries of $\boldsymbol{P}_t$ can be binary random variables indicating whether or not a node is probed at time t; also, they can act as multiplicative noise for modeling highly inaccurate measurements.

[1] This estimator will play a key role in our analysis, and in Section 3.5 we will present it in more detail.

There are two proposed estimators. The first one is the Granger estimator (2.7), which is adopted when matrix A is *dense*. The second one, employed when A is sparse, is a regularized version similar to the solution proposed by [49]. The peculiarity of this work is that the sample covariance matrices $\widehat{R}_0$ and $\widehat{R}_1$ are not computed in the standard way (i.e., by means of sample averages), but new ways are adopted to account for the corruption effects introduced by (2.12).

Also the kind of theoretical guarantees provided in [95] are of the same type as in [49]. The Authors show that the effect of data corruption is equivalent to a reduction of the number of collected samples, leading to the following conclusion: learning using corrupted data does not impair achievability, but only worsens the sample complexity of the system.

The assumption of corrupted data is a key difference that distinguishes our work from [95]. As a matter of terminology, we remark that in [95] the intermittent measurements at some nodes are referred to as "partial observations," but the meaning is different from the one adopted here. In fact, in [95] all nodes can be probed, even if not continuously at any time step, and the qualification "partial" refers to intermittence of observations at each node. Instead, in our work we assume that some nodes can be always probed (without corruption) and other nodes are completely inaccessible.

Continuous-time linear systems of stochastic equations. In [6] the Authors study the following N-dimensional *continuous-time* linear system described by the stochastic vector differential equations:

$$\dot{\boldsymbol{y}}_t = W_{\mathcal{G}} \boldsymbol{y}_t + \boldsymbol{x}_t, \qquad \text{for } t \geq 0, \tag{2.13}$$

where $W_{\mathcal{G}}$ is an $N \times N$ matrix with support $\mathcal{G}$, and $\boldsymbol{x}_t$ is a N-dimensional standard Brownian motion. System (2.13) is the continuous-time counterpart of (1.2), where the rate of change of the system state $\boldsymbol{y}_{k,t}$ relative to node k is driven by the current state of the neighbors of k, corrupted by an additive noise.

The topology inference algorithm proposed in [6] solves N independent regularized least squares problems, each one estimating a row of the adjacency matrix G of $\mathcal{G}$. The regularizers of the optimization problem encode sparsity assumptions on the topology.

The algorithm is fed by monitoring the trajectory of $\boldsymbol{y}_t$ over a time window $[0, W]$. In particular, the data are collected by sampling the trajectory with a suitable sampling period η. The work also develops a sample complexity analysis, and it shows that the resulting sample scaling laws critically depend on the value of η.

Frequency-domain approach. In [69] the Authors consider a set of N discrete-time, wide-sense stationary stochastic processes whose z-transforms $\boldsymbol{y}_1(z), \boldsymbol{y}_2(z), \ldots, \boldsymbol{y}_N(z)$ are mutually dependent in view of the following relations:

$$\boldsymbol{y}_k(z) = \sum_{\ell=1}^{N} h_{k\ell}(z) \boldsymbol{y}_\ell(z) + \boldsymbol{x}_k(z), \qquad \text{for } k = 1, 2, \ldots, N, \tag{2.14}$$

where $\boldsymbol{x}_1(z), \boldsymbol{x}_2(z), \ldots, \boldsymbol{x}_N(z)$ are the z-transforms of another set of random processes employed to model additive noise, while the function $h_{k\ell}(z)$ is equal to zero for all z if

process k is not influenced by process ℓ, and is a (possibly non-causal) transfer function otherwise. We can rewrite (2.14) in matrix form as:

$$\boldsymbol{y}(z) = H(z)\boldsymbol{y}(z) + \boldsymbol{x}(z), \tag{2.15}$$

where $\boldsymbol{x}(z)$ and $\boldsymbol{y}(z)$ stack the entries $\boldsymbol{x}_k(z)$ and $\boldsymbol{y}_k(z)$ into $N \times 1$ column vectors, and where the matrix $H(z) = [h_{k\ell}(z)]$ collects the transfer functions $h_{k\ell}(z)$.

From the sparsity pattern of matrix $H(z)$ the Authors define the directed support graph $\mathcal{G}_H$ having N nodes and directed edges such that:

$$\text{the edge from node } \ell \text{ to node } k \text{ is in } \mathcal{G}_H \iff h_{k\ell}(z) \neq 0. \tag{2.16}$$

By probing the nodes signal sequences across a finite time window, the Authors propose a reconstruction algorithm based on Wiener filter to estimate $\mathcal{G}_H$. The proposed strategy is shown to guarantee exact reconstruction for the so-called *self-kin networks*. For more general kinds of networks the strategy computes the smallest self-kin network embodying the true network. In a nutshell, self-kin networks are oriented graphs with the following transitivity property: if three nodes k, ℓ and h are such that the edges (k, h) and (ℓ, h) both exist (i.e., we have a directed edge from k to h and another directed edge from ℓ to h) then there must exist either edge (k, ℓ) or (ℓ, k). The self-kin assumption can be an important limitation in practical applications. In fact, it imposes the requirement that the aforementioned property holds for *any* triple of nodes, which is seldom verified over real networks, especially over large-scale networks.

Nonlinear dynamical systems. Most of the existing works on graph learning deal with models where the nodes' signals are linearly combined through a suitable weighting matrix. The extension of the existing results and tools to the nonlinear case is highly nontrivial. Before concluding this section, it is useful to mention some recent works that focus on topology inference over nonlinear dynamical systems.

In [44] the Authors consider a nonlinear dynamical model running on a graph $\mathcal{G}$ of N which satisfies some structural constraints like:

$$\boldsymbol{y}_t = f(\boldsymbol{y}_t, t; \mathcal{G}) + \boldsymbol{x}_t, \qquad \text{for } t = 1, 2, \ldots, \tag{2.17}$$

for some nonlinear function f. Note also that (2.17) does not represent an iterative evolution like (1.2), but instead it represents an *instantaneous* constraint enforced on $\boldsymbol{y}_t$. Let $\boldsymbol{y}_{k,t}$ and $\boldsymbol{y}_{\ell,t}$ be the signals of nodes k and ℓ, respectively. To determine if a directed edge exists from node ℓ to node k in $\mathcal{G}$, the Authors propose the following strategy, which extends to the nonlinear case the method of partial correlations. They define a mapping:

$$\widehat{\boldsymbol{y}}_{k,t} \triangleq \widehat{f}\Big(\{\boldsymbol{y}_{h,t} \,:\, h \in \{1, 2, \ldots, N\} \setminus \{k, \ell\}\}\Big). \tag{2.18}$$

arising from solving a nonlinear (possibly regularized) regression problem. In particular, the work considers kernel-based nonlinear regression models for $\widehat{f}$, namely, the function $\widehat{f}$ is modeled as a linear combination of nonlinear kernel functions. This mapping is an

estimator for $\boldsymbol{y}_{k,t}$ starting from the values of the other processes, except for $\boldsymbol{y}_{\ell,t}$ and $\boldsymbol{y}_{k,t}$. Define the *residual*:

$$\boldsymbol{\rho}_{k\ell} \triangleq \boldsymbol{y}_{k,t} - \widehat{f}\Big(\{\boldsymbol{y}_{h,t} \,:\, h \in \{1,2,\ldots,N\} \setminus \{k,\ell\}\}\Big), \tag{2.19}$$

and define symmetrically $\boldsymbol{\rho}_{\ell k}$. The proposed strategy consists in comparing against a threshold the correlation coefficient between the residuals $\boldsymbol{\rho}_{k\ell}$ and $\boldsymbol{\rho}_{\ell k}$, and an edge between nodes ℓ and k is declared iff the threshold is exceeded.

2.2 Graph Learning Under Partial Observability

Gaussian graphical models. In [18] the Authors address graph learning over Gaussian graphical models. Thus, we have a multivariate Gaussian $\boldsymbol{y} \in \mathbb{R}^N$ with concentration matrix $J_{\mathcal{G}}$ satisfying (2.1) for a certain underlying graph $\mathcal{G}$ of N nodes. One observes a stream of i.i.d. realizations from the *marginal* Gaussian distribution relative to a subset $\mathcal{P} \subset \{1,2,\ldots,N\}$ of the vector entries. Recalling that by definition $\Sigma_{\mathcal{G}} \triangleq J_{\mathcal{G}}^{-1}$ is the covariance matrix of the entire multivariate Gaussian, then by using the Schur complement formula we get (let Σ and J denote $\Sigma_{\mathcal{G}}$ and $J_{\mathcal{G}}$, respectively):

$$\Sigma_{\mathcal{P}} = J_{\mathcal{P}} - J_{\mathcal{P}\mathcal{P}'}(J_{\mathcal{P}'})^{-1}J_{\mathcal{P}'\mathcal{P}}, \tag{2.20}$$

where $\mathcal{P}'$ is the complement set $\{1,2,\ldots,N\} \setminus \mathcal{P}$, relative to the unobserved variables. Recalling (2.1), we have that the first term on the RHS of (2.20) encodes the subgraph topology $\mathcal{G}_{\mathcal{P}}$ of the graphical model relative to the observed nodes. The Authors assume that this term is sparse. The second term is observed to be a low-rank matrix, provided that the number of unobserved variables is small relative to the number of observed variables.

Since matrix $\Sigma_{\mathcal{P}}$ in (2.20) can be estimated from the data, we can in principle estimate the two terms described above and, therefore: *i*) recover the subgraph $\mathcal{G}_{\mathcal{P}}$, and *ii*) compute the number of unobserved variables. However, this "inverse problem" is highly underdetermined in general, and the aforementioned *sparsity & low-rank assumption* is essential to turn it into a feasible problem. Under the aforementioned assumptions, it is shown in [18] that the graph learning problem can be formulated as a regularized convex optimization problem. In particular, the Authors adopt an ℓ_1-norm regularization to account for sparsity of the matrix associated with the probed nodes, and a nuclear-norm regularization to control the rank of the matrix associated with the latent nodes.

Restricted Boltzmann machines. In [15], the Authors consider a special kind of graphical model called *restricted Boltzmann machine*. It is a special kind of bipartite graph, namely, a graph where the nodes can be divided into two partitions such that any edge of the graph connects a node of one partition with a node of the other partition (that is, each partition represents a fully disconnected subgraph). In particular, a restricted Boltzmann machine is a bipartite graph equipped with a weighted adjacency matrix $W_{\mathcal{G}}$, namely, a matrix such that for each pair of nodes k and ℓ:

$$[W_{\mathcal{G}}]_{k\ell} = 0 \iff \text{nodes } k \text{ and } \ell \text{ are disconnected in } \mathcal{G}. \tag{2.21}$$

The nodes in the two partitions of the graph encode two random vectors $\boldsymbol{y}'$ and $\boldsymbol{y}''$ with binary entries, whose joint distribution is parametric in $W_{\mathcal{G}}$ and such that the entries of $\boldsymbol{y}'$ are mutually independent given $\boldsymbol{y}''$ and, conversely, the entries of $\boldsymbol{y}''$ are mutually independent given $\boldsymbol{y}'$. The vector entries of $\boldsymbol{y}'$ and $\boldsymbol{y}''$ are called observed variables and latent variables, respectively. The aim of the work is to learn the distribution of the observed variables $\boldsymbol{y}'$. Since this distribution is parametrized by $W_{\mathcal{G}}$, this strategy is also useful for a graph learning task. The proposed strategy has been proved to work well for a specific class of Boltzmann machines that are named *ferromagnetic*. It is composed of two steps. The first step consists of a greedy algorithm for learning the two-hop neighborhood of an observed node. In this algorithm a key role is played by a statistical descriptor called *empirical influence*. Once the two-hop neighborhoods are determined, a further technique is proposed to learn the distribution of the observed variables by means of a regression algorithm involving the two-hop neighbors.

Graphical models with tree-like topology. In [3] the problem of structure estimation in graphical models with latent variables is considered for the family of locally tree-like graphs with a bound on the girth, which is the length of the shortest cycle in the graph. The proposed algorithm operates in two stages. The first stage is based on reconstructing acyclic local parts of the graph, and the second stage merges them together. In order to learn efficiently the local acyclic pieces of the graph, a further assumption is required, referred to as *correlation decay*, which guarantees that the correlation of "far" nodes in the graph is small.

However, assumptions like the bounded girth and the tree-like structure can be hard to be met in practice, especially for large-scale networks. In fact, these assumptions enforce some *local constraints* on the network, which must be verified at *any* portion on the underlying graph. For example, the tree-like assumption imposes that there not even a single cycle in the graph. As a result, as the considered network gets large, such type of conditions become more and more difficult to be met.

Moreover, graphical models such as the ones used in the aforementioned references do not assume that there are signals evolving over time at the network nodes. For this reason, the results obtained in the aforementioned references on graph learning in the presence of latent variables do not apply to the dynamical system considered in our work. A relevant exception is [37], where the Authors consider N random processes whose joint distribution, when represented with a graphical model, exhibits a *polytree* topology. A polytree is a directed acyclic graph whose corresponding undirected graph (obtained by replacing its directed edges with undirected edges) is a tree. It is interesting to note that the polytree topology fulfills the self-kin constraint used in [69], with the only exceptions at the roots of the shared subtrees.

In [37], graph learning must be performed from data samples collected by monitoring a subset $\mathcal{P}$ of the N network nodes. No information is available about the unobserved nodes (neither their number is known). The goal is to discover the hidden nodes of the polytree along with the connections between the hidden and observed nodes. The rationale of the algorithm is as follows. The first step is to discover the set of roots of the underlying polytree. For each root found at the previous step, the second step consists in

reconstructing the relative tree. The last step builds an estimated polytree by merging the trees estimated at the previous step. The Authors also carry out a sample complexity analysis, determining the size of the time horizon required to successfully retrieve the polytree.

Parameter estimation in vector autoregressive systems. In the context of full observability, we discussed some works that focus on estimating the parameters of a vector autoregressive system by probing the system nodes over time [49, 95]. Now we discuss another work on the same topic which assumes instead partial observability [42].

In [42], the following N-dimensional vector autoregressive system is considered:

$$\boldsymbol{y}_t = A\boldsymbol{y}_{t-1} + \boldsymbol{x}_t, \qquad \text{for } t = 1, 2, \dots. \tag{2.22}$$

Given a subset $\mathcal{P} \subset \{1, 2, \dots, N\}$ of probed nodes, let $\mathcal{P}' \triangleq \{1, 2, \dots, N\} \setminus \mathcal{P}$ be its complement set, i.e., the set of unobserved nodes. Thus, the equations in (2.22) can be conveniently partitioned in the following block representation:

$$\begin{pmatrix} [\boldsymbol{y}_t]_{\mathcal{P}} \\ [\boldsymbol{y}_t]_{\mathcal{P}'} \end{pmatrix} = \begin{pmatrix} A_{\mathcal{P}\mathcal{P}} & A_{\mathcal{P}\mathcal{P}'} \\ A_{\mathcal{P}'\mathcal{P}} & A_{\mathcal{P}'\mathcal{P}'} \end{pmatrix} \begin{pmatrix} [\boldsymbol{y}_{t-1}]_{\mathcal{P}} \\ [\boldsymbol{y}_{t-1}]_{\mathcal{P}'} \end{pmatrix} + \begin{pmatrix} [\boldsymbol{x}_t]_{\mathcal{P}} \\ [\boldsymbol{x}_t]_{\mathcal{P}'} \end{pmatrix}. \tag{2.23}$$

The aim is to estimate a submatrix $A_{\mathcal{P}}$ from observing $[\boldsymbol{y}_t]_{\mathcal{P}}$ over time. The addressed problem is solved by assuming that the number of probed nodes, $|\mathcal{P}|$, is smaller than the number of unobserved nodes, $|\mathcal{P}'| = N - |\mathcal{P}|$. This is a restrictive assumption that is often not met in practice, especially over large networks. We remark that, in contrast, our analysis will hold for any number of unobserved nodes (ruling out the trivial cases that the probed subset is either fully connected or fully disconnected). Two algorithms are proposed in [42], one using a variational expectation-maximization approach and another exploiting the second-order moments of the random process $[\boldsymbol{y}_t]_{\mathcal{S}}$. These algorithms are proved to work well under some technical "identifiability" conditions.

Continuous-time systems. In the context of full observability we have seen that the works in [49, 95] and the work in [6] adopt a similar model, with the key difference that the former works consider a continuous-time dynamics, whereas the latter work considers discrete-time signals. In a similar fashion we can now present [51] as the continuous-time counterpart of [42]:

$$\begin{pmatrix} [\dot{\boldsymbol{y}}_t]_{\mathcal{P}} \\ [\dot{\boldsymbol{y}}_t]_{\mathcal{P}'} \end{pmatrix} = \begin{pmatrix} A_{\mathcal{P}\mathcal{P}} & A_{\mathcal{P}\mathcal{P}'} \\ A_{\mathcal{P}'\mathcal{P}} & A_{\mathcal{P}'\mathcal{P}'} \end{pmatrix} \begin{pmatrix} [\boldsymbol{y}_t]_{\mathcal{P}} \\ [\boldsymbol{y}_t]_{\mathcal{P}'} \end{pmatrix} + \begin{pmatrix} [\boldsymbol{x}_t]_{\mathcal{P}} \\ [\boldsymbol{x}_t]_{\mathcal{P}'} \end{pmatrix}, \qquad \text{for } t \geq 0. \tag{2.24}$$

The graph learning strategy proposed in [51] is similar to the one proposed in [18], and amounts to find an optimal sum decomposition of a matrix into a sparse matrix and a low-rank matrix, by means of a regularized convex optimization problem.

The key assumption made here is that the considered models have the so-called *local-global structure,* where each of the observed random processes is explicitly influenced by only a few observed ones, while at the same time the unobserved processes interact with

many observed ones. As in [42], the results in [51] are proved under the assumption that the number $|\mathcal{P}'|$ of latent variables is smaller than the number $|\mathcal{P}|$ of observed ones.

Frequency-domain approach. References [70] and [71] consider the same model adopted in [69], with the difference that only the processes from a subset of nodes can be observed. As in [37], the underlying topology is assumed to be a polytree. The proposed algorithm is based on the computation of the so-called *log-coherence distance,* which requires the knowledge of the cross-spectral densities of the observed processes.

In [71] the transfer functions can be learned by a generalization of the so-called *door criterion*, a powerful parameter identification tool for structural equations models. This generalization is important since it allows to consider graphs having loops, unlike other works [37, 70] that rule out this possibility.

Link Prediction problems. Another interesting problem related to graph learning is the *link prediction* problem. Consider a network graph $\mathcal{G}$ with N nodes and assume that, for a fixed pair of nodes k and ℓ, one is interested to know whether an edge connecting them exists. The link prediction problem attempts to estimate the likelihood of the existence of such an edge (i.e., a "link") starting from some relevant information on the graph $\mathcal{G}$, such as the set of known edges, some attributes of nodes k and ℓ (and possibly of their neighbors), or some structural properties of the graph (e.g., sparsity). The simplest family of link prediction methods uses the similarity-based framework, where each pair of nodes k and ℓ is assigned a score defined as the similarity score $s_{k,\ell}$. All non-observed links are ranked according to their scores, and the links connecting more similar nodes are supposed to have higher likelihoods. The method based on similarity scores can be very simple or very complicated and it may work well for some networks while fail for some others.

Node similarity can be defined by using some essential attributes/descriptors of nodes or, if such attributes are unavailable, by considering some forms of structural kinships. One of the simplest similarity index is the *common neighbors* index: two nodes are more likely to have a link if they have many common neighbors. For instance, this quantity has been used to examine collaboration networks, showing a positive correlation between the number of common neighbors and the probability that two scientists will collaborate in the future [86]. In [56], the neighbors counting index has been exploited to confirm the common intuition that two students having many mutual friends are very probable to be friends in future. Many variants of this index have been proposed, each one favoring different kinds of similarity criteria, and a systematic survey can be found in [62].

Anyway, similarity heuristics make strong assumptions on when two nodes are likely to be connected, which limits their effectiveness on networks where these assumptions fail. For example, consider again the common neighbors heuristic. Its assumption may be correct in social networks [56], but was shown to fail in protein-protein interaction networks, where two proteins sharing many common neighbors are actually less likely to interact [57]. In this regard, a more rewarding approach could be to learn a suitable heuristic from a given network instead of using predefined rules. This can be done by using modern deep learning techniques. By extracting a local subgraph around each target link, the aim is to learn a function mapping the subgraph patterns to link existence, thus

automatically learning a heuristic suited to the current network. An important work in this field is [123], which proposes a solution based on the so-called *SEAL* framework. In a nutshell, this solution consists in extracting *local enclosing subgraphs* around links as training data, and use a neural network to learn which enclosing subgraphs correspond to link existence. More formally, the enclosing subgraph for a node pair (k, ℓ) is the subgraph induced from the network made by the neighbors of k and ℓ up to a certain number of hops.

An approach different from similarity-based techniques is proposed in [28], which is useful for many networks when some hierarchical organization exists among nodes. The Authors define a general technique to infer the underlying hierarchy of the network and predict the missing links. In particular, the procedure adopts a maximum likelihood method, where the likelihood is a function of both the *network dendogram* $\mathcal{D}$ (i.e., a possible network hierarchy) and of the conditional probability that, given $\mathcal{D}$, a certain link exists. The hierarchical structure model provides a smart way to predict missing links, and is able to capture the hidden hierarchical structure of the network. However, this algorithm is very slow. Its time complexity is usually quadratic in the network size and, in the worst case, it takes exponential time. In comparison, the hierarchical structure model cannot manage a network with tens of thousand nodes, while the algorithms based on local similarity indices can deal with networks with tens of million nodes [62]. Another noticeable remark is that this model may give poor predictions for networks without clear hierarchical structures.

In [39, 122], a class of graph learning strategies is proposed, based on the following steps: first, learn suitable regression or classification models that best represent an observed real network, and then predict the missing links by using the learned model. Given a target network graph $\mathcal{G}$, this kind of probabilistic models will optimize a target function to learn a network model $\mathcal{M}$ which can fit the observed data of the target network; then the probability of existence of a link (k, ℓ) is estimated through the conditional probability:

$$\mathbb{P}\left(k \text{ and } \ell \text{ connected} \mid \text{model } \mathcal{M}\right). \tag{2.25}$$

The link prediction works presented so far do not assume any dynamics in the underlying graph, whereas in [52, 88] link prediction is settled on networks whose topology evolves over time. In [88], a time sequence of graphs snapshots is given and classical similarity scores are computed on each snapshot, so as to get a time sequence for each score. Then, a forecasting model is used on these time series to predict the future similarity scores. In [52], given a snapshot sequence, the similarity scores are computed in dependence of the so-called *temporal events*, which encode the formation/disappearance time pattern of a link.

By comparison, we note that the graph learning problem addressed in the present thesis work and the link prediction both devise a reverse-engineering approach on some data collected from a networked apparatus to determine whether two nodes are directly connected or not. Anyway, there are some important differences between the two approaches.

The first difference is about the nature of the available input data. The considered graph learning problem aims at retrieving the graph from some kind of signals emitted by the nodes, and no topological information is at disposal. On the other side, in the context of link prediction the proposed strategies assume the knowledge of part of the graph interconnections. Remarkable examples are the common neighbors similarity index and, even more evidently, the use of the enclosing graphs for feeding graph neural networks. The second difference regards the kind of considered dynamics. In the graph learning setting considered in this work, the networked system is inherently dynamical, since the signals emitted by the nodes evolve over time. On the other side, the link prediction problem generally does not allow for dynamical systems. Some works consider networks with a time-evolving topology, and the information considered for the vertices can be considered time-varying only in the sense that it reflects the structural changes of the networks. Another difference pertains to the scope of the prediction. In graph learning problems, even the ones facing with partial observability, the goal is to estimate an entire (sub)graph at once, which may be arbitrarily large. On the other side, in link prediction settings the focus is to detect the existence of a single edge at a time. Even if one can learn large portions of a network one link at a time, this is usually more expensive than directly estimating all the links with techniques tailored to this task.

2.3 Main Contributions

In the previous section, several literature works in the field of graph learning have been discussed. In particular, for any work we introduced the key assumptions of the addressed problem, the network model, the kind of dynamics (if any) evolving over it, and the ideas underlying the proposed solutions. We have also organized these works in a convenient taxonomy (see Table 2.1) to shed light on the similarities among them and the considered problem. Now, we are ready to pinpoint the exact position of the proposed work in this big picture, by systematically describing the research advances achieved with the results collected in the present thesis. This analysis is reported in Table 2.2, which highlights the main elements of novelty of the present thesis: we work under *partial observability, without limitations in the number of unobserved nodes, considering large-scale network models that do not enforce local constraints that are hardly met in large-scale networks and examining network systems equipped with a dynamics.* This setting was first proposed in [73–75,77,78,100], where we can find results on both achievability and sample complexity under the assumption that the underlying graph is drawn according to the Erdős-Rényi model. We present a summary of these results in Chapter 4. As anticipated in the previous chapter, in this thesis we extend the existing results in a twofold direction: *i*) We introduce a unifying framework useful for a general class of random graph models (Chapter 3); *ii*) we characterize the achievability and sample complexity of Bollobás-Riordan graphs (Chapter 5).

Assumptions of other works...	**...which differ in our work.**
The graph learning problem is examined under the *full observability* assumption, where all the network nodes are known to exist, can be accessed and are probed [2, 6, 44, 49, 69, 80, 90, 95, 106].	The graph learning problem is examined under the *partial observability* assumption. This assumption naturally takes into account many real-world applications, when one faces time/space complexity constraints, privacy concerns, etc.
A *"weak" partial observability* regime is considered, where the number of unobserved nodes must be less than the number of observed ones [18].	A *"strong" partial observability* regime is considered, where the number of unobserved nodes is arbitrary, ruling out the trivial case that the probed subgraph is either fully connected or fully disconnected.
The graph learning problem is examined for *fixed, finite-size* networks. Only the achievability issue is considered, while the sample complexity analysis is not carried out [42, 44, 69–71].	The graph learning problem is examined in the *doubly-asymptotic framework* where both the network size and the number of samples grow. This provides analytic tools for studying both achievability and sample complexity issues for *large-scale* networks.
The examined networked system *is not dynamical* [2, 3, 15, 18, 62, 123].	The examined networked system *runs the autoregressive dynamics* (1.2).
The topology of the target graph must fulfill some *local constraints* like bounded girth plus correlation decay [3], self-kin structures [69] or tree-like constraints [37], which must be satisfied at *any* specific piece of the network. These conditions are hard to be met over large networks. For example, in polytree networks, not even a single cycle is admitted in the graph.	The considered network graph topology arises from the Bollobás-Rordan model. This model does not enforce local constraints. Moreover, it produces heterogeneous networks featuring real-world properties like dependence across edges and the formation of "hubs" with many connections, as opposed to peripheral nodes with few connections.
The topology of the target graph is assumed to be drawn according the Erdős-Rényi model, which leads to homogeneous networks [73–75, 77, 78, 100].	,

Table 2.2: Summary table reporting the key differences between the present work and the existing works.

Chapter 3

Achievable Graph Learning

Over a networked dynamical system, the problem of retrieving the graph from signals collected at the nodes is meaningful since the (partially) observed network dynamics is dependent on the underlying graph $\mathcal{G}(N)$. However, it is not obvious how the graph learning task should be implemented and whether the underlying graph can be learned faithfully. The achievability analysis addresses these questions. Specifically, when we say that *graph learning is achievable*, we mean that there exists a strategy that ensures faithful reconstruction of the probed network topology in the asymptotic regime where the network size goes to infinity, and the number of samples is allowed to grow with the network size.

For the considered dynamics (1.2), this dependence on the graph is conveyed by the combination matrix $\boldsymbol{A}(N)$. For this reason, a promising approach to build a graph estimator consists in the following two steps: *i*) first compute a *matrix estimator*:

$$\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N) = g\Big(\boldsymbol{Y}_{\mathcal{P}}(T,N)\Big), \tag{3.1}$$

and *ii*) then provide a suitable strategy to classify the entries of the matrix estimator as connected or disconnected pairs.

This approach has been successfully followed for scenarios employing the Erdős-Rényi model [73–75, 78] and the Bollobás-Riordan model [25, 26] to generate the random graph $\mathcal{G}(N)$. However, the validity of this approach is not limited to these specific models, and could be adopted in future works to other topologies. In this chapter we will describe this approach in its general form, i.e., without assuming a specific model for $\mathcal{G}(N)$, while in the next chapters we will derive from the general analysis presented here some fundamental results on graph learning over specific graph models.

The forthcoming notions of identifiability, universal local structural consistency, identifiability gap and bias where introduced in [73–75, 78]. These notions cover the case of *deterministic* identifiability gap and bias. However, there are relevant classes of graphs that do not match this assumption. For example, as we will see later, in the case of

preferential attachment graphs the identifiability gap is *random* [25, 26]. In this chapter, we will accordingly extend the original definitions by including the possibility that both the identifiability gap and the bias are random quantities, and identify three general conditions to ensure achievability: *i*) *asymptotic concentration* of the combination matrix (Definition 1), which ensures that, as $N \to \infty$, the (scaled) matrix entries corresponding to the graph edges stay well separated from the zero entries, and are all clustered around a strictly positive value; *ii*) the existence of a *limiting matrix estimator* as $T \to \infty$, i.e., for an ideally infinite number of samples (Definition 4); *iii*) the *regularity* of the limiting estimator (Definition 5), which ensures that this estimator is similar, for large network sizes, to (a possibly biased version of) the true combination matrix, thus preserving useful information to detect the graph edges. When these three conditions hold, the resulting sample estimator fulfills universal local structural consistency (Definition 2), a fundamental property that will be shown to imply achievability of the graph learning problem under partial observability — see Theorems 1 and 2 further ahead. With this general framework in mind, we are able to state general conditions for achievability, which are also used to organize conveniently the proofs of the results pertaining to Erdős-Rényi and Bollobás-Riordan graphs, reported in Chapters 4 and 5, respectively.

3.1 Useful Class of Combination Matrices

Throughout our treatment, we will focus on combination matrices satisfying the following asymptotic concentration property.

Definition 1 (Asymptotic Concentration of the Combination Matrix). *Consider a family of random graphs $\mathfrak{G}(N)$ with adjacency matrices $\boldsymbol{G}(N)$. If there exist a positive sequence c_N and a positive random variable $\boldsymbol{\gamma}$ defined on the probability space $(\Omega, \mathcal{F}, P)$, such that for any deterministic sequence of probed subsets $\mathcal{S}_N$:*

$$\|c_N \boldsymbol{A}_{\mathcal{S}}(N) - \boldsymbol{\gamma}\, \boldsymbol{G}_{\mathcal{S}}(N)\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{3.2}$$

then we say that the combination matrix $\boldsymbol{A}(N)$ is asymptotically concentrated for the family of graphs $\mathfrak{G}(N)$, with scaling sequence c_N and identifiability gap $\boldsymbol{\gamma}$. □

In principle, the nature of the randomness of $\boldsymbol{\gamma}$ looks rather abstract from the definition. Its practical meaning will become clearer in the Chapter 5, where we will study the Bollobás-Riordan preferential attachment model. Over this model, the random variable $\boldsymbol{\gamma}$ arises as a specific limiting value associated to the sequence of maximum degrees of the graphs obtained during the preferential attachment procedure.

Let us now give some insight on the practical meaning of Definition 1. Since $\boldsymbol{\gamma}$ is positive, from (3.6) we can write, for $\varepsilon > 0$:[1]

$$\lim_{N\to\infty} \mathbb{P}\left[\|c_N \boldsymbol{A}_{\mathcal{S}}(N) - \boldsymbol{\gamma}\, \boldsymbol{G}_{\mathcal{S}}(N)\|_{\text{max-off}} > \varepsilon\boldsymbol{\gamma}\right] = 0, \tag{3.4}$$

[1] Given a sequence of random variables $\boldsymbol{e}(N)$ vanishing in probability with N, and a strictly positive

which means that, with high probability as $N \to \infty$, for all $k \neq \ell$ in $\mathcal{S}_N$:

$$c_N \, \boldsymbol{a}_{k\ell}(N) \in \begin{cases} [(1-\varepsilon)\boldsymbol{\gamma}, (1+\varepsilon)\boldsymbol{\gamma}], & (k,\ell) \text{ connected}, \\ \\ [-\varepsilon\boldsymbol{\gamma}, \varepsilon\boldsymbol{\gamma}], & (k,\ell) \text{ disconnected}. \end{cases} \tag{3.5}$$

We see that, for small ε, the combination matrix entries are tightly clustered: *i*) around a positive value $\boldsymbol{\gamma}$ for connected node pairs, and *ii*) around zero for disconnected node pairs — see Figure 3.1 (top panel) for a graphical illustration.

3.2 Universal Local Structural Consistency

If the combination matrix $\boldsymbol{A}_{\mathcal{S}}(N)$ is asymptotically concentrated according to Definition 1, we expect that a good estimator $\widehat{\boldsymbol{A}}_{\mathcal{S}}(T,N)$ emulates this behavior. In particular, we introduce the following definition.

Definition 2 (Universal Local Structural Consistency of the Matrix Estimator). *Consider a combination matrix $\boldsymbol{A}(N)$ with support graph $\mathcal{G}(N)$, having adjacency matrix $\boldsymbol{G}(N)$. If there exist a positive sequences c_N, and two random variables $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, with $\boldsymbol{\gamma}$ strictly positive, defined on the probability space $(\Omega, \mathcal{F}, P)$, such that for some T_N and for the deterministic sequence of probed subsets $\mathcal{S}_N$:*

$$\|c_N \widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N) - \boldsymbol{\gamma}\boldsymbol{G}_{\mathcal{S}}(N) - \boldsymbol{\beta}\|_{\text{max-off}} \xrightarrow{\text{P}} 0, \tag{3.6}$$

then we say that the matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ achieves universal local structural consistency for the sequence $\mathcal{S}_N$, with sample law T_N, scaling sequence c_N, identifiability gap $\boldsymbol{\gamma}$ and bias $\boldsymbol{\beta}$. □

As a matter of terminology: *i*) the adjective *universal* is used because, as we will promptly show, Eq. (3.6) automatically enables the possibility of recovering the topology by means of *unsupervised* clustering, i.e., without prior knowledge as regards the network size and other system parameters; *ii*) the adjective *local* comes from the observation that the structure of the topology connecting nodes in the probed subset will be faithfully recovered by probing *only these particular nodes*; and *iii*) the adjective *structural* is used because we estimate only the structure (i.e., the support graph) underlying the combination matrix.

---

random variable $\boldsymbol{z}$, for an arbitrary $\delta > 0$ we have:

$$\begin{aligned} \mathbb{P}[\boldsymbol{e}(N) > \boldsymbol{z}] &\leq \mathbb{P}[\boldsymbol{e}(N) > \boldsymbol{z}, \boldsymbol{z} > z_{\delta/2}] + \mathbb{P}[\boldsymbol{z} \leq z_{\delta/2}] \\ &\leq \mathbb{P}[\boldsymbol{e}(N) > z_{\delta/2}] + \delta/2 < \delta, \end{aligned} \tag{3.3}$$

where $z_{\delta/2} > 0$ is chosen such that $\mathbb{P}[\boldsymbol{z} \leq z_{\delta/2}] \leq \delta/2$ (a condition that can be met for any δ since $\mathbb{P}[\boldsymbol{z} \leq 0] = 0$), and the last inequality holds for sufficiently large N since $\boldsymbol{e}(N)$ vanishes in probability. The arbitrariness of δ implies that $\lim_{N\to\infty} \mathbb{P}[\boldsymbol{e}(N) > \boldsymbol{z}] = 0$, and (3.4) follows from (3.6) by setting $\boldsymbol{e}(N) = \|c_N \boldsymbol{A}_{\mathcal{S}}(N) - \boldsymbol{\gamma}\boldsymbol{G}_{\mathcal{S}}(N)\|_{\text{max-off}}$ and $\boldsymbol{z} = \varepsilon\boldsymbol{\gamma}$.

Using the same arguments that, starting from (3.2), led to the configuration (3.5) illustrated in Figure 3.1 (top panel), we can also conclude that the universal local structural consistency (3.6) implies the following asymptotic dichotomy, illustrated in Figure 3.1 (bottom panel), with high probability as $N \to \infty$:

$$c_N\, \widehat{\boldsymbol{a}}_{k\ell}(T_N, N) \in \begin{cases} \Big[\boldsymbol{\beta} + (1-\varepsilon)\boldsymbol{\gamma},\ \boldsymbol{\beta} + (1+\varepsilon)\boldsymbol{\gamma}\Big], & (k,\ell) \text{ connected}, \\ \Big[\boldsymbol{\beta} - \varepsilon\boldsymbol{\gamma}, \boldsymbol{\beta} + \varepsilon\boldsymbol{\gamma}\Big], & (k,\ell) \text{ disconnected}. \end{cases} \tag{3.7}$$

By comparing the two illustrations in Figures 3.1 we observe that the off-diagonal entries of the matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N)$ preserve the same pattern of the corresponding entries of matrix $\boldsymbol{A}_{\mathcal{S}}(N)$, with the difference that the estimator has a *bias* quantified by the random variable $\boldsymbol{\beta}$. Notably, this bias does not impair the possibility of recovering the support graph. What matters to achieve faithful graph recovery is the presence of the *identifiability gap* $\boldsymbol{\gamma}$ that separates the matrix entries corresponding to connected and disconnected node pairs. This property automatically enables the possibility of *clustering* the off-diagonal entries of matrix $\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N)$ so as to classify connected vs. disconnected pairs, ruling out the trivial case that the probed subgraph is either fully connected or fully disconnected, i.e., that we have only one cluster. Once a clustering procedure graphclu$(\cdot)$ is devised, we obtain the consistent graph estimator:

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T, N) = \text{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)\Big). \tag{3.8}$$

This result is established in the next section.

3.3 Clustering the Matrix Estimator

It is not difficult to envisage clustering algorithms that can achieve correct classification of the node pairs under condition (3.7). In order to show that such an algorithm actually exists, we need to introduce first a formal definition of correct clustering.

Definition 3 (Correct Clustering). *Let* $\mathsf{A} = [\mathsf{a}_{k\ell}]$ *be an* $S \times S$ *matrix with nonnegative entries and let* G *the adjacency matrix corresponding to the support graph of* A*. The diagonal entries of* G *are zero by convention (since we are not interested in self-loops).*

Let $\widehat{\mathsf{A}} = [\widehat{\mathsf{a}}_{k\ell}]$ *be an estimated matrix fulfilling, for some positive values* ε *and* γ*, and for value* β*, the following condition for all* $k \neq \ell$*:*

$$\widehat{\mathsf{a}}_{k\ell} \in \begin{cases} \Big[\beta + (1-\varepsilon)\gamma,\quad \beta + (1+\varepsilon)\gamma\Big], & \text{if } \mathsf{a}_{k\ell} > 0, \\ & \text{(i.e., if } (k,\ell) \text{ is connected in } \mathsf{G}) \\ \Big[\beta - \varepsilon\gamma,\quad \beta + \varepsilon\gamma\Big], & \text{otherwise}, \end{cases} \tag{3.9}$$

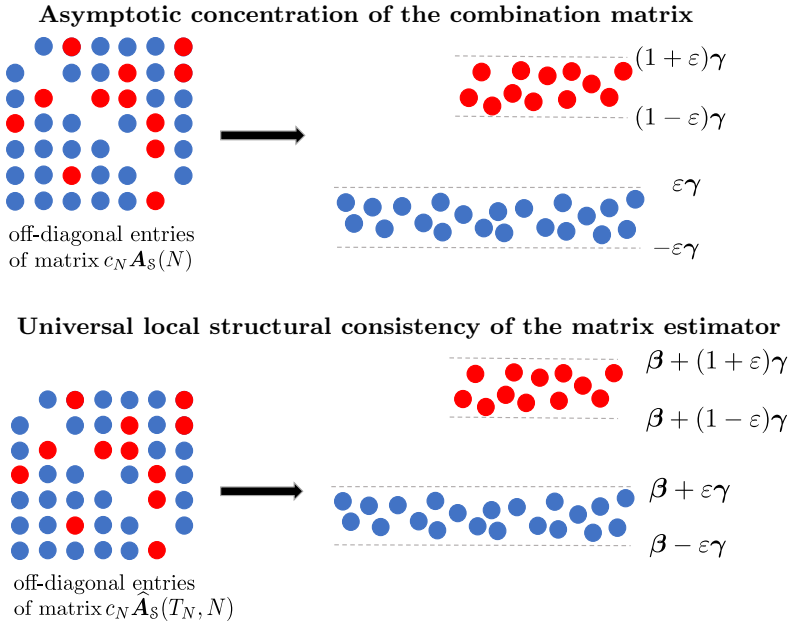

Figure 3.1: Graphical illustration of Definition 1 (top panel) and Definition 2 (bottom panel). Blue circles denote disconnected node pairs, whereas red circles denote connected pairs. The essential difference between the two illustrations is that the entries of the matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N)$, while still preserving the same dichotomy as the corresponding entries of matrix $\boldsymbol{A}_{\mathcal{S}}(N)$, converge to biased limit values.

and let

$$\widehat{\mathsf{G}} = \text{graphclu}(\widehat{\mathsf{A}}) : \mathbb{R}^{S\times S} \to \{0,1\}^{S\times S} \tag{3.10}$$

be a clustering algorithm that computes an estimated adjacency matrix $\widehat{\mathsf{G}}$*. Then, algorithm* graphclu$(\cdot)$ *will be said to be correct if, when the sets of connected and disconnected pairs in* G *are both non-empty, there exists a sufficiently small* ε *such that we have:*

$$\widehat{\mathsf{G}} = \mathsf{G} \tag{3.11}$$

independently of the values of γ *and* β*.* □

As an example of clustering that matches Definition 3, let us consider an algorithm that computes the midpoint between the maximum and minimum off-diagonal entries of the estimated matrix. Using the bounds in (3.9) we can write:

$$\beta + \left(\frac{1}{2} - \varepsilon\right)\gamma \leq \frac{\max_{k,\ell} \widehat{\mathsf{a}}_{k\ell} + \min_{k,\ell} \widehat{\mathsf{a}}_{k\ell}}{2} \leq \beta + \left(\frac{1}{2} + \varepsilon\right)\gamma. \tag{3.12}$$

Accordingly, correct clustering will be surely performed if the lowest admissible value $\beta + (1-\varepsilon)\gamma$ for the connected pairs lies above the threshold, namely if:

$$\beta + (1-\varepsilon)\,\gamma > \beta + \left(\frac{1}{2} + \varepsilon\right)\gamma \Leftrightarrow \varepsilon < \frac{1}{4}, \tag{3.13}$$

and if the highest admissible value $\beta + \varepsilon\gamma$ for the disconnected pairs lies below the threshold, namely if:

$$\beta + \varepsilon\gamma < \beta + \left(\frac{1}{2} - \varepsilon\right)\gamma \Leftrightarrow \varepsilon < \frac{1}{4}. \tag{3.14}$$

In summary, the simple algorithm that employs an intermediate threshold to separate the clusters matches Definition 3, provided that $\varepsilon < 1/4$.

We now prove that universal local structural consistency of the sample estimator implies the existence of a consistent graph estimator according to (1.6), which in turn implies achievability. This is established in the next theorem.

Theorem 1 (Sufficient Conditions for Consistency of the Graph Estimator). *Consider a graph model $\mathcal{G}(N)$ and a sequence of probed subsets $\mathcal{S}_N$ such that the probability that $\mathcal{G}_{\mathcal{S}}(N)$ is either fully connected or fully disconnected vanishes as $N \to \infty$. Let $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ be an estimator of $\boldsymbol{A}_{\mathcal{P}}(N)$, which achieves universal local structural consistency according to Definition 2 with sample law T_N, and let* graphclu$(\cdot)$ *be a correct clustering algorithm according to Definition 3. Then, the graph estimator:*

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T,N) = \text{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)\Big) \tag{3.15}$$

satisfies the consistency property (1.6), namely,

$$\lim_{N\to\infty} \mathbb{P}\left[\widehat{\mathcal{G}}_{\mathcal{S}}(T_N,N) = \mathcal{G}_{\mathcal{S}}(N)\right] = 1. \tag{3.16}$$

for the family of graphs $\mathcal{G}(N)$, the sequence of probed subsets $\mathcal{S}_N$ and the sample law T_N.

Proof: See Appendix A. ■

From a theoretical standpoint, the midpoint rule is enough to conclude that correct clustering is possible. However, the midpoint rule is not necessarily the best option to be used in practice, as we will carefully explain in Section 3.6, where we will also introduce another clustering rule inspired by the k-means algorithm.

3.4 Limiting Matrix Estimators and Consistent Graph Estimators

It is useful to introduce the *limiting* estimator, which corresponds to the ideal estimator obtained when an infinite number of samples is available, namely, with $T \to \infty$ and N fixed.

Definition 4 (Limiting Matrix Estimator). *We say that a matrix estimator* $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ *converges to a limiting estimator* $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ *if, for any* $\varepsilon > 0$*:*

$$\lim_{T\to\infty} \mathbb{P}\Big[\|\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N) - \widehat{A}_{\mathcal{P}}(N)\|_{\max} > \varepsilon | \boldsymbol{A}(N) = A\Big] = 0. \tag{3.17}$$

□

We notice that: *i*) the conditional probability in (3.17) depends on the *overall* combination matrix $\boldsymbol{A}(N)$ corresponding to the entire network; and *ii*) the limiting estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ inherits the randomness in $\boldsymbol{A}(N)$, but in (3.17) we used normal font to denote it since the realization of $\boldsymbol{A}(N)$ is fixed due to conditioning. Condition (3.17) is a standard condition of consistency achieved by many empirical estimators, which are convergent as the number of samples scales to infinity.

If a limiting estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ exists, it makes sense to examine whether it allows us to recover faithfully the true combination matrix. In particular, it is useful to introduce the following notion of regularity.

Definition 5 (Regular Limiting Estimator). *Consider an asymptotically concentrated combination matrix* $\boldsymbol{A}(N)$*, with scaling sequence* c_N*. If there exist a random variable* $\boldsymbol{\beta}$ *defined on the probability space* $(\Omega, \mathcal{F}, P)$*, such that for the deterministic sequence of probed subsets* $\mathcal{S}_N$ *we have:*

$$\Big\| c_N \Big(\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N) \Big) - \boldsymbol{\beta} \Big\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{3.18}$$

then we say that the limiting estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ *is regular for the sequence* $\mathcal{S}_N$*, with bias* $\boldsymbol{\beta}$*.* □

The next theorem establishes useful connections among the combination matrix, the sample estimator and the limiting estimator.

Theorem 2 (Sufficient Conditions for Universal Local Structural Consistency). *Let* $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ *be an estimator of the combination submatrix* $\boldsymbol{A}_{\mathcal{P}}(N)$ *such that:*

i) $\boldsymbol{A}(N)$ *asymptotically concentrates according to Definition 1, with scaling sequence* c_N *and identifiability gap* $\boldsymbol{\gamma}$*,*

ii) $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ *converges to the limiting matrix estimator* $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ *as* $T \to \infty$ *according to Definition 4,*

iii) the limiting estimator is regular in the sense of Definition 5, with bias $\boldsymbol{\beta}$*, for the sequence* $\mathcal{S}_N$*.*

Then, there exists a sample law T_N*, with* $T_N \overset{N\to\infty}{\longrightarrow} \infty$*, such that the matrix estimator* $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ *achieves universal local structural consistency for the sequence* $\mathcal{S}_N$ *according to Definition 2, with scaling sequence* c_N*, identifiability gap* $\boldsymbol{\gamma}$ *and bias* $\boldsymbol{\beta}$*.*

Proof: See Appendix A. ■

We remark that in Theorem 2 the law T_N is unspecified. In other words, the theorem ensures that the exists a suitable sample law ensuring consistent learning, but it does not establish how fast the number of samples must grow with the network size. The aim of the sample complexity analysis is to fill this gap.

We can summarize the results of the present analysis as follows. Assume that the combination matrix $\boldsymbol{A}(N)$ fulfills Definition 1 and assume that we have a matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ of $\boldsymbol{A}_{\mathcal{P}}(N)$ converging according to Definition 4. With these premises, in Theorem 2 we proved that the regularity property of the limiting estimator of Definition 5 implies universal local structural consistency of the sample estimator according to Definition 2, and therefore it implies achievability in view of Theorem 1. This is a useful result, since it allows us to establish whether graph learning is achievable by focusing only on the properties of the *limiting* estimator, i.e., disregarding sample complexity issues.

3.5 Granger Estimator

So far we have dealt with the abstract concepts of matrix estimators and limiting matrix estimators. In this section we introduce a concrete example of estimator that will play a fundamental role in the analysis conducted in Chapters 4 and 5.

Preliminarily, it is necessary to introduce the steady-state covariance matrix and the one-lag covariance matrix corresponding to model (1.1), which are, respectively [63]:

$$\boldsymbol{R}_0(N) = \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{y}_i(N)\boldsymbol{y}_i^\top(N)\middle|\boldsymbol{A}(N)\right], \tag{3.19}$$

$$\boldsymbol{R}_1(N) = \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{y}_i(N)\boldsymbol{y}_{i-1}^\top(N)\middle|\boldsymbol{A}(N)\right], \tag{3.20}$$

where bold notation for the covariance matrices is used to encompass randomness of the underlying graph. Under model (1.1) (with a stable matrix $\boldsymbol{A}(N)$), it is known that the covariance matrix is the solution to the discrete-time Lyapunov equation $\boldsymbol{A}(N)\boldsymbol{R}_0(N)\boldsymbol{A}^\top(N) - \boldsymbol{R}_0(N) + I = 0$, which is [103]:

$$\boldsymbol{R}_0(N) = \sum_{i=0}^{\infty} \boldsymbol{A}^i(N)[\boldsymbol{A}^i(N)]^\top. \tag{3.21}$$

Moreover, by exploiting (1.1) it is readily seen that:

$$\boldsymbol{R}_1(N) = \boldsymbol{A}(N)\boldsymbol{R}_0(N), \tag{3.22}$$

which implies the following inversion formula:

$$\boldsymbol{A}(N) = \boldsymbol{R}_1(N)\boldsymbol{R}_0^{-1}(N) \Rightarrow \boldsymbol{A}_{\mathcal{P}}(N) = [\boldsymbol{R}_1(N)\boldsymbol{R}_0^{-1}(N)]_{\mathcal{P}}. \tag{3.23}$$

Since covariance matrices can be faithfully estimated through *sample* covariance matrices as the number of samples increases, Eq. (3.23) suggests that the *true* matrix $\boldsymbol{A}_{\mathcal{P}}(N)$ can

be actually estimated from the samples, which would imply that consistent graph learning is trivially possible.

However, under partial observability we can only compute the covariance matrices *over the probed subset*, $[\boldsymbol{R}_0(N)]_{\mathcal{P}}$ and $[\boldsymbol{R}_1(N)]_{\mathcal{P}}$. As a result, computation of the inversion formula (3.23) is impaired by the unavailability of signals from the latent nodes. Nevertheless, the limiting estimator (which actually depends only on the covariances over the probed subset):

$$\widehat{\boldsymbol{A}}_{\mathcal{P}}(N) = [\boldsymbol{R}_1(N)]_{\mathcal{P}}([\boldsymbol{R}_0(N)]_{\mathcal{P}})^{-1} \tag{3.24}$$

is a meaningful choice that, in the context of Granger causality [45], is referred to as the *Granger estimator or predictor*, and attempts to provide the best linear prediction of the future samples from the past one-lag samples over the probed subset. In a nutshell, Granger causality refers to the relationships between time series. With reference to our example, assume that we regress $\boldsymbol{y}_{k,t}(N)$ on the past one-lag time series available in the network, $\boldsymbol{y}_{\ell,t-1}(N)$, for $\ell \in \mathcal{P}$. It has been proved that the optimal predictor of $\boldsymbol{y}_{k,t}(N)$ (i.e., the one minimizing the regression error) does not exploit the time series $\boldsymbol{y}_{\ell,t-1}(N)$ if $\boldsymbol{a}_{k\ell}(N) = 0$. Thus, one says that k is "Granger-caused" by those ℓ such that $\boldsymbol{a}_{k\ell}(N) \neq 0$.

On the other hand, from elementary matrix algebra we know that:

$$[\boldsymbol{R}_1(N)]_{\mathcal{P}}([\boldsymbol{R}_0(N)]_{\mathcal{P}})^{-1} \neq [\boldsymbol{R}_1(N)\boldsymbol{R}_0(N)^{-1}]_{\mathcal{P}}, \tag{3.25}$$

namely, the Granger estimator (3.24) constructed from the probed subset is different from the Granger estimator (3.23) constructed from the entire network and then projected onto the probed subset. This difference gives rise to the question of whether $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ can be still profitably used to estimate graph $\mathcal{G}_{\mathcal{P}}(N)$.

Sample Granger Estimator. We introduce the *sample* Granger estimator:

$$\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N) = [\boldsymbol{R}_1(T, N)]_{\mathcal{P}}([\boldsymbol{R}_0(T, N)]_{\mathcal{P}})^{-1}, \tag{3.26}$$

which replaces the true covariance matrices appearing in (3.24) with the sample covariance matrices $\boldsymbol{R}_0(T, N)$ and $\boldsymbol{R}_1(T, N)$, whose (k, ℓ)-entries are defined as:

$$[\boldsymbol{R}_j(T, N)]_{k\ell} = \frac{1}{T-j} \sum_{i=1+j}^{T} \boldsymbol{y}_{k,i}(N)\boldsymbol{y}_{\ell,i-j}^{\top}(N), \qquad j = 0, 1. \tag{3.27}$$

By ergodicity, the sample Granger estimator converges to the limiting estimator in (3.24) in the sense of Definition 4.

Regularized Granger Estimator. Another version of the Granger estimator, which will be particularly useful in our sample complexity analysis, is the regularized Granger estimator. When dealing with covariance-based estimators, one source of sample complexity comes from how well-conditioned these matrices are. For this reason, it is useful to replace (3.26) with its *regularized* counterpart, namely, a matrix $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)$ constructed as follows. For $k \in \mathcal{P}$, the k-th row of $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)$ is a solution to the constrained optimization

problem (here $x \in \mathbb{R}^{|\mathcal{P}|}$ is a row vector, and, for a matrix M, the notation $[M]_{k\mathcal{P}}$ denotes the k-th row of submatrix $M_{\mathcal{P}}$):

$$\min_{x \in \mathbb{R}^{|\mathcal{P}|}} \|x\,[\boldsymbol{R}_0(T,N)]_{\mathcal{P}} - [\boldsymbol{R}_1(T,N)]_{k\mathcal{P}}\|_\infty \quad \text{s.t. } \|x\|_1 \leq 1. \tag{3.28}$$

We remark that, when the sample covariance matrix is invertible, the non-regularized Granger estimator in (3.26) is the only matrix that yields a zero residual in (3.28). As a result, when the non-regularized Granger estimator fulfills the constraint in (3.28), the two estimators coincide.

3.6 A modified k-means Algorithm

Let us recall the relation in (3.7) arising from Definition 2. We have that, for N large, the off-diagonal entries of the matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N)$ form two clusters, and in particular: *i*) the entries relative to disconnected pairs are scattered around a value $\boldsymbol{\beta}$, and *ii*) the entries relative to connected pairs are scattered around a strictly higher value $\boldsymbol{\beta}+\boldsymbol{\gamma}$. Moreover, the amount of scattering around these two values becomes asymptotically negligible. This behavior is qualitatively illustrated in Figure 3.1 (bottom panel).

There is no doubt that any reasonable clustering algorithm would be able to properly separate the two clusters when the scattering effect is sufficiently small, namely, *for a sufficiently large network size* N and a suitable sample law T_N. For example, in Section 3.3 we showed that an asymptotically correct separation of the two clusters is obtained by simply choosing as separating threshold the midpoint between the maximum and minimum off-diagonal entries of the matrix estimator. In particular, we showed that this rule implements a correct clustering algorithm in the sense specified by Definition 3, which means that the midpoint rule works well for sufficiently small ε. In terms of the matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N)$, a small ε means that the entries of this estimator must be sufficiently concentrated around the values $\boldsymbol{\beta}$ and $\boldsymbol{\beta}+\boldsymbol{\gamma}$. This concentration requires that N is sufficiently large.

However, the asymptotic correctness of an algorithm is not the only characteristic that we need in practical situations. Another desirable characteristic is a high tolerance to the amount of scattering due to finite network size effects. Therefore, between two or more algorithms fulfilling Definition 3, we will always prefer the one which behaves better for finite network sizes.

In this respect, we have observed that the popular k-means algorithm (in our case, with $k = 2$) seems to work better than the aforementioned midpoint rule in a high variety of practical situations. Unfortunately, it is well-known that this algorithm can have problems in presence of unbalanced clusters [43, 101], a phenomenon illustrated in Figure 3.2. To understand why, let us briefly discuss how the k-means algorithm works. Let v be the vector containing the data to cluster. The k-means attempts to minimize

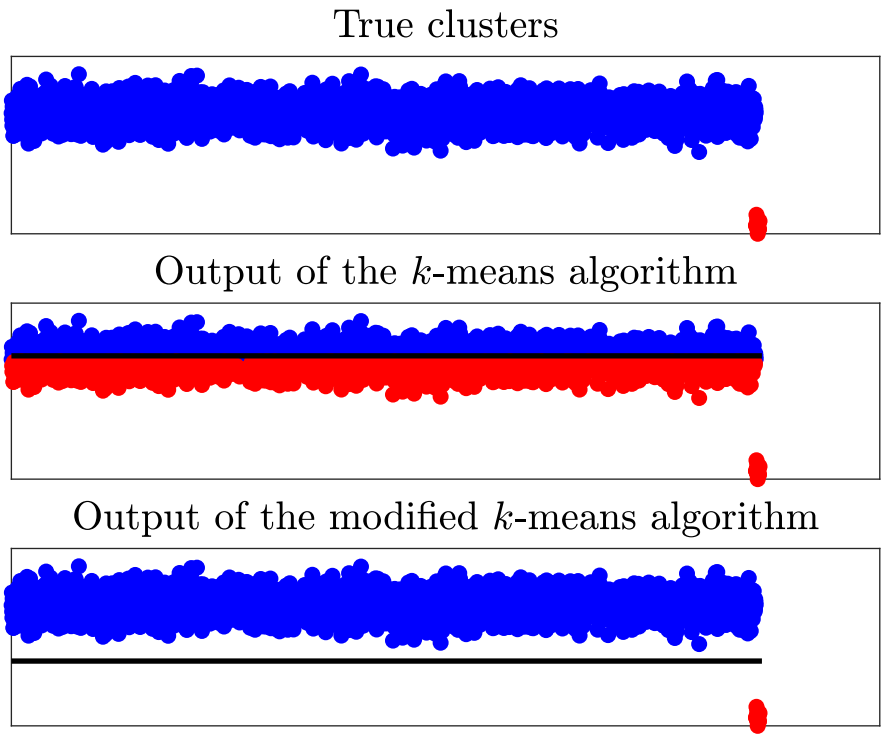

Figure 3.2: An example showing the benefits of the modified k-means algorithm over the classical k-means algorithm with $k = 2$ in presence of unbalanced clusters. In the top panel we show the dataset that we employ for the experiment. Note that the data are organized in two clusters, with very different sizes. In particular, the points in the big cluster are depicted in blue, while the points in the small cluster are depicted in red. In the middle and bottom panels, we show the clusters computed by the k-means algorithm and its modified version reported in Listing 1, respectively. In both panels, the black line represents the midpoint between the centroids computed by the pertinent algorithm. Consequently, all points above the threshold are marked in blue, since they are the estimation of the (true) big cluster shown in the top panel. The remaining points are marked in red. We note that the k-means algorithm returns a wrong solution, which in practice separates the big cluster in two. As described in the main text, this biased behavior arises from the fact that the cost associated to the small cluster in (3.29) is negligible. On the other hand, the modified k-means algorithm does not have the same issue, and in fact it successfully estimates the true clusters.

the following cost function [43, 101]:

$$\sum_{v_j \in \mathcal{C}_0} (v_j - c_0)^2 + \sum_{v_j \in \mathcal{C}_1} (v_j - c_1)^2, \tag{3.29}$$

over all possible clusters $\mathcal{C}_0$ and $\mathcal{C}_1$, where c_0 and c_1 are the clusters' centroids, defined respectively as:

$$c_0 \triangleq \frac{1}{|\mathcal{C}_0|} \sum_{v_j \in \mathcal{C}_0} v_j, \qquad c_1 \triangleq \frac{1}{|\mathcal{C}_1|} \sum_{v_j \in \mathcal{C}_1} v_j. \tag{3.30}$$

It is possible to show that, to minimize the cost function in (3.29), it suffices to consider the cluster pairs $\mathcal{C}_0$ and $\mathcal{C}_1$ whose centroids are such that the midpoint:

$$\frac{c_0 + c_1}{2} \tag{3.31}$$

separates the two clusters, namely,

$$\forall v_j \in \mathcal{C}_0 \quad v_j < \frac{c_0 + c_1}{2}, \qquad \text{and} \qquad \forall v_j \in \mathcal{C}_1 \quad v_j > \frac{c_0 + c_1}{2}. \tag{3.32}$$

We will refer to this set as set of *admissible configurations*, and we will denote it by $\mathcal{A}$. Figure 3.2 illustrates two possible admissible configurations in a case where the true clusters associated to the input vector v have very different sizes (top panel). The first admissible configuration (middle panel) is wrong, since it splits the larger cluster in two. On the other hand, the second configuration (bottom panel) is correct, namely, represents correctly the true data partitioning. In some cases where there is a very large cluster, as the one shown in Figure 3.2, it may happen that the cost function in (3.29) is minimized by the wrong admissible configuration and not by the correct one.

The intuition behind this fact is that the error associated to the small true cluster is negligible, and the cost function in (3.29) creates a bias in favor of the entries of the large cluster. This is of course a undesirable property, which may lead to wrong conclusions even if the true clusters are clearly visible and neatly separated as in the case of Figure 3.2.

Moreover, since in our model it is actually permitted that one cluster dominates the other one. For instance, we will see in the next chapters that over some popular random graph models, in the sparse regime we can have a connected graphs (where any pair of nodes is connected through some path) while the fraction of connected node pairs can even vanish as $N \to \infty$. Owing to the aforementioned problems, we are not guaranteed that the k-means algorithm is asymptotically correct in the sense of Definition 3.

To overcome these limitations, it is possible to devise a straightforward modification of the k-means algorithm [75]. By examining again the two solutions in Figure 3.2, we see that the correct admissible configuration is the one with higher distance $c_1 - c_0$ between the cluster centroids. Therefore, the flaw of the k-means algorithm can be remediated by selecting, among all the admissible solutions in the set $\mathcal{A}$, the one with largest centroids distance. A pseudo-code of this algorithm is reported in Listing 1.

Assume that the input vector v contains L elements, and assume that the elements have been preliminarily arranged in ascending order. The first step of the modified k-means algorithm consists in computing the set $\mathcal{A}$. Since the elements of v are ordered, this set can be computed by performing an exhaustive search over the cluster configurations:

$$\mathcal{C}_0(j) = \{1, 2, \ldots, j\}, \qquad \text{and} \qquad \mathcal{C}_1(j) = \{j+1, j+2, \ldots, L\}, \tag{3.33}$$

Listing 1: Modified k-means algorithm, $j^\star = \mathsf{clu}(v)$

% v is an $L \times 1$ vector with entries sorted in ascending order
% the algorithm computes the set of admissible configurations for the standard k-means
$\mathcal{A} = \emptyset$;
for $j = 1 : L-1$ **do**
  % set 2 tentative clusters
  $\mathcal{C}_0(j) = \{v_1, v_2, \ldots, v_j\}$;
  $\mathcal{C}_1(j) = \{v_{j+1}, v_{j+2}, \ldots, v_L\}$;
  % compute the centroids of the 2 clusters
  $c_0(j) = \frac{1}{j}\sum_{i=1}^{j} v_i, \quad c_1(j) = \frac{1}{L-j}\sum_{i=j+1}^{L} v_i$;
  % check if the midpoint between the centroids separates the clusters
  if $v_j < \dfrac{c_0(j) + c_1(j)}{2} < v_{j+1}$ **then**
    $\mathcal{A} = \mathcal{A} \cup \{j\}$;
  end
end
% select the admissible configuration with largest centroid distance
$j^\star = \underset{j \in \mathcal{A}}{\operatorname{argmax}}[c_1(j) - c_0(j)]$;

for $j \in \{1, 2, \ldots, L-1\}$. Accordingly, we see that any possible partition is identified by an index j. Consequently, the algorithm scans all possible cluster pairs (3.33) spanning the set $\{1, 2, \ldots, L-1\}$, and retains the set of indices fulfilling the required condition in the list $\mathcal{A} = \{j_1, j_2, \ldots\}$.

For the traditional k-means algorithm (with $k = 2$), a criterion to select an admissible configuration is the minimization of (3.29). Contrariwise, the modified algorithm returns the configuration $j^\star$ in $\mathcal{A}$ with maximum distance between the centroids:

$$j^\star = \underset{j \in \mathcal{A}}{\operatorname{argmax}}[c_1(j) - c_0(j)], \tag{3.34}$$

where $c_0(j)$ and $c_1(j)$ are computed from $\mathcal{C}_0(j)$ and $\mathcal{C}_1(j)$ according to (3.30). We remark that in principle we could have multiple maximizers, and in this case the choice for $j^\star$ should be further refined with some additional criterion. However, in our case this refinement is immaterial since we will see later in Theorem 3 that for our assumptions the maximizer $j^\star$ is unique with high probability as $N \to \infty$.

Finally, in order to build a graph estimator of the form (3.8), we need to build a $\mathsf{grapclu}(\cdot)$ operator of the form (3.10) starting from the procedure described in Listing 1. This is an easy task, which can be formally summarized as follows:

1. Given an input matrix M, its off-diagonal entries are vectorized and sorted in ascending order. Let v be the resulting vector.

2. The clustering algorithm in Listing 1 is applied to vector v, and the optimal configuration $j^\star$ is constructed.

3. The zero-diagonal adjacency matrix G, of the same size as M, is constructed such that the off-diagonal entry $g_{k\ell}$ is set to 0 if the corresponding element v_j belongs to[2] $\mathcal{C}_0(j^\star) = \{v_1, v_2, \ldots, v_{j^\star}\}$, and to 1 otherwise, namely, if element v_j belongs to $\mathcal{C}_1(j^\star) = \{v_{j^\star+1}, v_{j^\star+2}, \ldots, v_L\}$.

A pseudo-code of this algorithm is reported in Listing 2. Now that we have a working $\mathsf{graphclu}(\cdot)$ operation of the form (3.10), which benefits from the flexibility of the k-means algorithm while concurrently ruling out the biased behavior reported in Figure 3.2, we show that it satisfies Definition 3.

Theorem 3 (Correctness of the modified k-means algorithm). *The clustering algorithm reported in Listing 2 satisfies Definition 3 with* $\varepsilon \leq \frac{1}{6}$.

Proof: See Appendix A.

In summary, the modified k-means algorithm produces two benefits. First, it is asymptotically correct for $N \to \infty$. Second, it improves the performance of the k-means for finite size networks, since it also works in situations like the one illustrated in Figure 3.2.

3.7 General Scheme for Consistent Graph Estimators

Building upon the analysis developed so far, we can identify a general scheme to construct consistent graph estimators over a wide range of situations adhering to the problem introduced in Chapter 1. We will see in later chapters how this scheme can be successfully exploited to build consistent graph estimators in context where the graph sequence $\mathcal{G}(N)$ is drawn according to some popular random graph models, namely, the Erdős-Rényi and the Bollobás-Riordan models.

Corollary 1 (Sufficient Conditions for Achievability). *Consider the graph learning problem under partial observability over vector autoregressive systems like (1.2). Let the quadruple:*

$$\Big\{\mathcal{G}(N),\ \mathcal{S}_N,\ \boldsymbol{A}(N),\ \widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)\Big\} \tag{3.35}$$

satisfy the following properties:

i) the probed subsets $\mathcal{S}_N$ is such that the probability that $\mathcal{G}_{\mathcal{S}}(N)$ is either fully connected or fully disconnected vanishes as $N \to \infty$,

ii) the combination matrix $\boldsymbol{A}(N)$ asymptotically concentrates according to Definition 1,

iii) the matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ converges to the limiting matrix estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ according to Definition 4, and this limit is regular in the sense of Definition 5.

[2]Note that the vectorization and the sorting operations that we used at step 1 to compute v from M define a one-to-one mapping between the matrix indices (k,ℓ) and the vector indices j.

Listing 2: Graph Estimator based on the modified k-means algorithm, $G = \mathsf{graphclu}(M)$

% M is an $S \times S$ matrix
% the algorithm computes an $S \times S$ zero-diagonal matrix with off-diagonal entries equal to 0 or 1
% mapping the output of the modified k-means algorithm over the entries of M

% vectorize and sort the off-diagonal entries of M,
% also record the index map $\mathsf{idx} : (k, \ell) \mapsto j$ such that if $j = \mathsf{idx}(k, \ell)$ then $m_{k\ell} = v_j$
$v, \mathsf{idx}(\cdot, \cdot) = \mathsf{sort}\big(\mathsf{vec}\big(\mathsf{off\text{-}diag}(M)\big)\big)$;

% cluster the sorted entries and compute the cluster $\mathcal{C}_1(j^\star)$
$j^\star = \mathsf{clu}(v)$; $\quad \mathcal{C}_1(j^\star) = \{v_{j^\star+1}, v_{j^\star+2}, \ldots, v_L\}$;

% initialize the estimated adjacency matrix with all zeros, then set to 1 the off-diagonal entries $g_{k\ell}$ for which $m_{k\ell}$ is in $\mathcal{C}_1(j^\star)$
$G = \mathsf{zeros}(S \times S)$;
for $k, \ell = 1 : S$ **with** $k \neq \ell$ **do**
  $j = \mathsf{idx}(k, \ell)$;
  if $v_j \in \mathcal{C}_1(j^\star)$ **then**
    $g_{k\ell} = 1$;
  end
end

Then, the graph estimator:

$$\widehat{\mathfrak{G}}_{\mathcal{P}}(T, N) = \mathrm{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)\Big), \tag{3.36}$$

where $\mathrm{graphclu}(\cdot)$ *is the clustering procedure reported in Listing 2, satisfies the consistency property (1.6), namely,*

$$\lim_{N \to \infty} \mathbb{P}\left[\widehat{\mathfrak{G}}_{\mathbb{S}}(T_N, N) = \mathfrak{G}_{\mathbb{S}}(N)\right] = 1. \tag{3.37}$$

for some sample law T_N (whose characterization is the aim of the sample complexity analysis).

Proof: This result is a direct consequence of Theorems 1, 2 and 3. ■

Chapter 4

Learning Erdős-Rényi Graphs

In this chapter we present some literature results [73–75, 77, 78, 100] on graph learning under partial observability (as formulated in Chapter 1) when the underlying graph sequence $\mathcal{G}(\boldsymbol{N})$ is drawn according to the Erdős-Rényi model.

4.1 Erdős-Renyi Model

The most popular model to build a random graph is the Erdős-Rényi model. This model generates an *undirected* graph, which means that edges exist always in both directions or, equivalently, that there is no need to talk of directed edges, but simply of edges. Accordingly, the adjacency matrix $\boldsymbol{G}$ is symmetric, and the random process giving rise to it generates only the upper (or lower) triangular part of the matrix. Over an Erdős-Rényi random graph, the edges are determined, one independently from the other, by running a sequence of Bernoulli experiments with identical success (i.e., connection) probability [7, 35]. Accordingly, the variables $\boldsymbol{g}_{k\ell}(N)$, for $k, \ell = 1, 2, \ldots, N$ and $k < \ell$, are independent Bernoulli random variables with connection probability:

$$p_N \triangleq \mathbb{P}[\boldsymbol{g}_{k\ell}(N) = 1]. \tag{4.1}$$

One useful graph descriptor is the node degree, which counts the number of neighbors of a node. The degree of node k is:

$$\boldsymbol{d}_{\mathcal{G},k}(N) \triangleq \sum_{\ell=1}^{N} \boldsymbol{g}_{k\ell}(N). \tag{4.2}$$

According to (4.1) and (4.2), the expected degree (i.e., the expected number of neighbors of each node) is given by:

$$\mathbb{E}[\boldsymbol{d}_{\mathcal{G},k}(N)] = (N-1)p_N, \qquad \text{for each node } k = 1, 2, \ldots, N. \tag{4.3}$$

In particular, in the following analysis we shall use the minimal and maximal degrees, which are defined respectively as:

$$\boldsymbol{\nu}_{\mathcal{G}}(N) \triangleq \min_{k\in[1,N]} \boldsymbol{d}_{\mathcal{G},k}(N), \qquad \boldsymbol{\mu}_{\mathcal{G}}(N) \triangleq \max_{k\in[1,N]} \boldsymbol{d}_{\mathcal{G},k}(N). \tag{4.4}$$

The explicit dependence of the connection probability upon N is a critical feature of the Erdős-Rényi model, since it allows characterizing different types of asymptotic graph behavior. First of all, in order to guarantee that the graph is *connected* with high probability as $N \to \infty$, the connection probability must satisfy [7,35]:

$$p_N = \frac{\log N + c_N}{N}, \qquad c_N \overset{N\to\infty}{\longrightarrow} \infty. \tag{4.5}$$

When (4.5) is verified we say that we are in the *connected regime*. A relevant class of connected graphs is the class where the expected degree in (4.3) goes to infinity faster than $\log N$, which means that, for $0 \leq p < 1$:

$$p_N = \omega_N \frac{\log N}{N} \overset{N\to\infty}{\longrightarrow} p, \qquad \omega_N \overset{N\to\infty}{\longrightarrow} \infty. \tag{4.6}$$

It is known that, for this class of graphs, the minimal and maximal degrees both *concentrate*[1] asymptotically around Np_N, in the following sense.

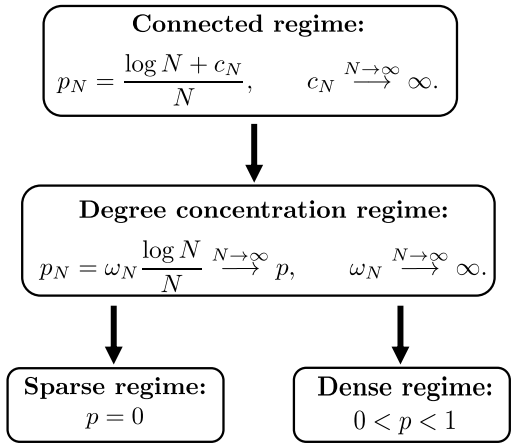

Figure 4.1: Taxonomy of the connected regimes of the Erdős-Rényi model considered in this work. The arrows indicate that we are moving from a more general to a more specific condition.

[1] We remark that here the term "concentration" is borrowed from a common terminology in statistics, which is used to describe situations when some random quantities asymptotically converge around a common value.

Theorem 4 (Degree Concentration in Erdős-Rényi Graphs). *Let $\mathcal{G}(N)$ be a graph sequence generated according to the Erdős-Rényi graph model with connection probability satisfying (4.6). Then, the minimal and the maximal degree sequences $\boldsymbol{\nu}_{\mathcal{G}}(N)$ and $\boldsymbol{\mu}_{\mathcal{G}}(N)$ arising from the graph sequence satisfy:*

$$\frac{\boldsymbol{\nu}_{\mathcal{G}}(N)}{Np_N} \xrightarrow{\text{p}} 1, \qquad \frac{\boldsymbol{\mu}_{\mathcal{G}}(N)}{Np_N} \xrightarrow{\text{p}} 1. \tag{4.7}$$

Proof: See Appendix A in [75]. ■

Recalling (4.3), we see that the physical meaning of (4.7) is that both the minimal and the maximal degrees scale as the expected degree asymptotically as $N \to \infty$. The connection regime described by (4.6) will be referred to as the *degree concentration regime*. Moreover, when the limit connection probability p in (4.6) is zero we talk of *sparse (connected) regime*. Otherwise, if the limit connection probability p is nonzero we talk of *dense regime*. Figure 4.1 illustrates a diagram that summarizes the concentration/sparsity taxonomy of the aforementioned connected regimes.

4.2 Assumptions on the Probed Subset $\mathcal{S}_N$

In order to study the graph learning problem under partial observability, it is necessary to specify how the sequence of probed subsets $\mathcal{S}_N$ scales with N. In particular, we are interested in ruling out the pathological situations where the subgraph of probed nodes end up being fully disconnected or fully connected as $N \to \infty$. The next lemma characterizes how the probed subset must scale with N to avoid these cases.

Lemma 1 (Nontrivial Erdős-Rényi Subgraphs). *Let $\mathcal{S}_N$ be a sequence of subsets satisfying (1.5) and:*

$$|\mathcal{S}_N| \overset{N\to\infty}{\longrightarrow} \infty. \tag{4.8}$$

Let $\mathcal{G}(N)$ be a random graph sequence drawn according to the Erdős-Rényi model with connection probability $p_N \overset{N\to\infty}{\longrightarrow} p$, where $0 \leq p < 1$. Then:

$$\lim_{N\to\infty} \mathbb{P}\big[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully connected}\big] = 0. \tag{4.9}$$

Moreover, if either $0 < p < 1$ or

$$p = 0 \qquad \textit{and} \qquad \lim_{N\to\infty} |\mathcal{S}_N|^2 \, p_N = \infty, \tag{4.10}$$

then:

$$\lim_{N\to\infty} \mathbb{P}\big[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully disconnected}\big] = 0. \tag{4.11}$$

Proof: See Appendix D. ■

An assumption of particular interest for the probed subsets is that their cardinalities scale linearly according to:

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \text{for some } 0 < \xi < 1. \tag{4.12}$$

It is easy to verify that with this choice the sequence $\mathcal{S}_N$ satisfies the hypotheses of Lemma 1. In fact, combining (4.6) and (4.12) we obtain the condition:

$$|\mathcal{S}_N|^2\, p_N \overset{N\to\infty}{\longrightarrow} \infty. \tag{4.13}$$

4.3 Assumptions on the Combination Policy $A(N)$

Once a network graph $\mathcal{G}(N)$ is constructed, it is necessary to define a policy to assign the combination matrix $\boldsymbol{A}(N)$. Combination matrices arise across several domains, including distributed optimization, adaptation and learning over networks, social learning, network stochastic control. In all these domains, the system designer is called to run an algorithm (e.g., for optimization, learning, control) over a certain network topology. To this end, he/she devises a distributed procedure where the network nodes exchange locally information according to a certain combination matrix. Some popular choices in these contexts are the Laplacian and the Metropolis policies [14, 30, 33, 55, 102, 103, 120], which are illustrated below.

4.3.1 Laplacian Matrix

The graph Laplacian $\boldsymbol{L}(N)$ is a matrix whose off-diagonal (k,ℓ)-entry is -1 for connected pairs (k,ℓ) and 0 otherwise, and whose k-th main diagonal entry is the degree of node k. Starting from $\boldsymbol{L}(N)$, the Laplacian combination matrix is defined as:

$$\boldsymbol{A}(N) = \rho \times (I - \boldsymbol{c}\,\boldsymbol{L}(N)), \quad \boldsymbol{c} \triangleq \frac{\lambda}{1+\boldsymbol{\mu}_{\mathcal{G}}(N)}, \tag{4.14}$$

where $\boldsymbol{\mu}_{\mathcal{G}}(N)$ is the maximal degree of $\mathcal{G}(N)$, $\lambda \leq 1$ is a positive parameter that tunes the relative importance of the self-weights, and $\rho < 1$ is a positive parameter that grants stability of the dynamical system in (1.1) — see [103]. The Laplacian combination rule can be conveniently described in terms of the individual entries as follows, for $k \neq \ell$:

$$\begin{cases} \boldsymbol{a}_{k\ell}(N) = 0, & (k,\ell) \text{ disconnected}, \\ \boldsymbol{a}_{k\ell}(N) = \dfrac{\rho\lambda}{1+\boldsymbol{\mu}_{\mathcal{G}}(N)}, & (k,\ell) \text{ connected}, \\ \boldsymbol{a}_{kk}(N) = \rho - \displaystyle\sum_{\substack{\ell=1 \\ \ell\neq k}}^{N} \boldsymbol{a}_{k\ell}(N). \end{cases} \tag{4.15}$$

4.3.2 Metropolis Matrix

The Metropolis combination matrix is defined as follows, for $k \neq \ell$:

$$\begin{cases} \boldsymbol{a}_{k\ell}(N) = 0, & (k,\ell) \text{ disconnected}, \\ \boldsymbol{a}_{k\ell}(N) = \dfrac{\rho}{1 + \max\{\boldsymbol{d}_{\mathcal{G},k}(N), \boldsymbol{d}_{\mathcal{G},\ell}(N)\}}, & (k,\ell) \text{ connected}, \\ \boldsymbol{a}_{kk}(N) = \rho - \displaystyle\sum_{\substack{\ell=1 \\ \ell \neq k}}^{N} \boldsymbol{a}_{k\ell}(N). \end{cases} \tag{4.16}$$

where $\boldsymbol{d}_{\mathcal{G},k}(N)$ is the degree of of node k in $\mathcal{G}(N)$, while $0 < \rho < 1$ has the same meaning as for the Laplacian matrix. The Metropolis combination rule is a special case of the *Hastings rule,* an optimal combination policy that boosts performance in distributed consensus/diffusion networks [103].

4.3.3 Regular Diffusion Matrices

The matrices arising from the Laplacian and Metropolis rules belong to a broader family of policies for which the combination matrices are symmetric and satisfy, for some parameters κ and ρ such that $0 < \kappa \leq \rho < 1$:

$$\begin{cases} \boldsymbol{a}_{k\ell}(N) = 0, & (k,\ell) \text{ disconnected}, \\ \dfrac{\kappa}{1 + \boldsymbol{\mu}_{\mathcal{G}}(N)} \leqslant \boldsymbol{a}_{k\ell}(N) \leqslant \dfrac{\kappa}{1 + \boldsymbol{\nu}_{\mathcal{G}}(N)}, & (k,\ell) \text{ connected}, \\ \boldsymbol{a}_{kk}(N) = \rho - \displaystyle\sum_{\substack{\ell=1 \\ \ell \neq k}}^{N} \boldsymbol{a}_{k\ell}(N). \end{cases} \tag{4.17}$$

We will refer to these kind of matrices as *regular diffusion matrices.* These matrices are *scaled* left-stochastic matrices (i.e., their rows sum up to a constant value ρ) with support graph $\mathcal{G}(N)$. We also remark that the Laplacian matrix and the Metropolis matrix match (4.17) with $\kappa = \rho\lambda$ and $\kappa = \rho$, respectively.

4.4 Achievability for Erdős-Rényi Graphs

The following lemma shows that the regular diffusion matrices in (4.17) asymptotically concentrate when their support graph $\mathcal{G}(N)$ is an Erdős-Rényi graph under the degree concentration regime.

Lemma 2 (Asymptotic Concentration of Regular Diffusion Matrices over Erdős-Rényi graphs). *Consider the case when the support graph of a regular diffusion*

matrix like (4.17) *is the simple graph* $\mathcal{G}(N)$ *drawn according to the Erdős-Rényi model with connection probability* p_N *scaling according to the degree concentration regime* (4.5), (4.6). *Then the combination matrix satisfies Definition 1 with scaling sequence* $c_N \triangleq Np_N$ *and deterministic identifiability gap* $\gamma \triangleq \kappa$.

Proof: See Appendix D. ■

The next result establishes the regularity of the limiting Granger estimator in the considered scenario.

Lemma 3 (Regularity of the Limiting Granger Estimator for Erdős-Rényi graphs). *Let* $\boldsymbol{A}(N)$ *be a regular diffusion matrix* (4.17) *with support graph* $\mathcal{G}(N)$ *drawn according to the Erdős-Rényi model under degree concentration regime* (4.5), (4.6). *Then for any sequence of probed subsets* $\mathcal{S}_N$ *such that:*

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \text{for some } 0 < \xi < 1, \tag{4.18}$$

the limiting Granger estimator satisfies:

$$\left\| Np_N \left(\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N) \right) - \beta \right\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{4.19}$$

with:

$$\beta \triangleq \kappa^2 p \frac{(2\rho - \kappa)(1 - \xi)}{1 - (\rho^2 - 2\rho\kappa\xi + \kappa^2\xi)}, \tag{4.20}$$

namely, it fulfills Definition 5 for the considered sequence $\mathcal{S}_N$, *with (deterministic) bias* β.

Proof: See Appendix D. ■

In view of Lemmas 2 and 3, we can now state the main achievability result for Erdős-Rényi random graphs.

Theorem 5 (Achievability for Erdős-Rényi Graphs). *Let us consider the dynamical system* (1.1), *with a regular diffusion matrix as in* (4.17), *and with network graph* $\mathcal{G}(N)$ *being a simple graph arising from the Erdős-Rényi model under degree concentration regime* (4.5), (4.6). *Then, for any probed subset sequence* $\mathcal{S}_N$ *such that:*

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \text{for some } 0 < \xi < 1, \tag{4.21}$$

the graph estimator:

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T, N) = \text{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N) \Big), \tag{4.22}$$

where graphclu$(\cdot)$ *is the clustering procedure in Listing 2, and* $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)$ *is the sample Granger estimator in* (3.26), *satisfies the consistency property* (1.6):

$$\lim_{N\to\infty} \mathbb{P}\left[\widehat{\mathcal{G}}_{\mathcal{S}}(T_N, N) = \mathcal{G}_{\mathcal{S}}(N) \right] = 1. \tag{4.23}$$

for some scaling law T_N.

Proof: This result is a direct consequence of Lemmas 1, 2 and 3, applied to Corollary 1, along with the fact that the sample Granger estimator in (3.26) converges to the limiting Granger estimator (3.24) by ergodicity. ■

4.5 Results on Sample Complexity

We now report the available results on the sample complexity of the regularized Granger estimator operating over Erdős-Rényi graphs. The analysis, conducted in [75], works under the following classical assumption on system (1.1) [6, 49, 61, 95].

Assumption 1 (Stationary Gaussian Vector Autoregressive System). *For the sample complexity analysis, we assume that the input signals $\boldsymbol{x}_{k,i}(N)$ are standard Gaussian variables (independent w.r.t. to node index k, time index i, and network size N). Under these conditions, given a realization $\boldsymbol{A}(N) = A$ of the combination matrix, the vector autoregressive process in (1.1) admits a Gaussian stationary distribution (which is a function of A) [6, 63]. We assume that, given $\boldsymbol{A}(N) = A$, the initial vector $\boldsymbol{y}_0(N)$ is distributed according to the stationary distribution.*

Theorem 6 (Sample Complexity of the Granger Estimator for Erdős-Rényi Graphs). *Let us consider the dynamical system (1.1) operating under Assumption 1 with a regular diffusion matrix as in (4.17), and with network graph $\mathcal{G}(N)$ arising from the Erdős-Rényi model under degree concentration regime (4.6). Then, for any probed subset sequence $\mathcal{S}_N$ such that:*

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \text{for some } 0 < \xi < 1, \tag{4.24}$$

the graph estimator:

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T, N) = \text{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)\Big), \tag{4.25}$$

where graphclu$(\cdot)$ *is the clustering procedure in Listing 2, and $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)$ is the sample Granger estimator in (3.26), is consistent with sample complexity law:*

$$T_N = C(Np_N)^2 \log |\mathcal{S}_N|, \tag{4.26}$$

for some constant $C > 0$.

Proof: See Appendix I in [75]. ■

We see from (4.26) that under the dense regime (where p_N converges to some nonzero probability p) the sample complexity is essentially quadratic in N. On the other hand, under the sparse regime, recalling from (4.6) that we have $Np_N = \omega_N \log N$, we see that:

$$T_N \sim (\omega_N \log N)^2 \log |\mathcal{S}_N| \tag{4.27}$$

revealing that the specific sample complexity under the sparse regime depends on the specific speed of growth of the sequence ω_N, which regulates the sparsity of the problem.

Chapter 5

Learning Bollobás-Riordan Graphs

5.1 Bollobás-Riordan Model

Preferential attachment graphs are typically obtained through an iterative process that goes as follows. Starting from a graph with a certain structure, at each subsequent iteration one node is added, along with some edges connecting this node to the graph constructed until that iteration. The term "preferential attachment" is used because the probability that the new node is connected to an existing node is proportional to the degree of the latter. Therefore, nodes that have already experienced a large amount of connections are favored, giving rise to a dichotomy in the network, where some nodes emerge as hubs with most of the connections, whereas the remaining nodes become peripheral and feature few connections.

The way to build a preferential attachment model is not unique. Since the pioneering work [4], several preferential attachment models have been proposed. One of the most popular variants is the Bollobás-Riordan random graph, which is the model examined in this work [8, 10]. The Bollobás-Riordan model provides an elegant mathematical formulation that allows to capture many features of real-world networks and to obtain clean analytical results for useful graph descriptors (e.g., node degrees, minimum and maximal degrees, centrality measures). Let us delve into the mathematical description of the Bollobás-Riordan model [8, 10].

First of all, a Bollobás-Riordan graph is a *multi*graph, which means that multiple self-loops and multiple edges are permitted. A random multigraph of size n will be denoted by $\mathfrak{M}(n)$ and its *adjacency matrix* by $\boldsymbol{M}(n)$. Matrix $\boldsymbol{M}(n)$ is the symmetric (since Bollobás-Riordan graphs are undirected) $n \times n$ matrix whose off-diagonal (k, ℓ)-entry $\boldsymbol{m}_{k\ell}(n)$ is the number of edges between nodes k and ℓ, and whose diagonal entry $\boldsymbol{m}_{kk}(n)$ is the number of self-loops of node k.

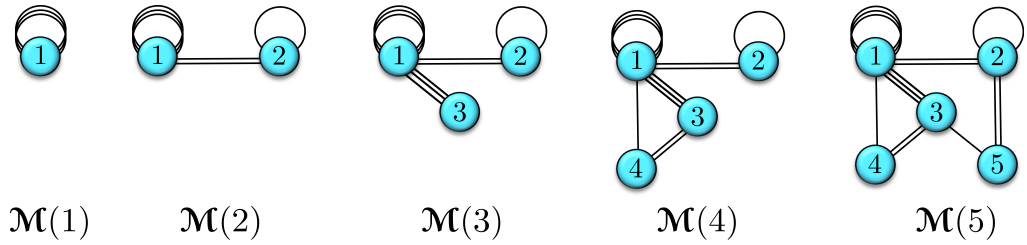

Figure 5.1: One example of iterative construction of a Bollobás-Riordan multigraph with parameter $\eta = 3$.

Then, the Bollobás-Riordan preferential-attachment model with parameter $\eta \in \mathbb{N}$ generates iteratively a random sequence of multigraphs $\mathfrak{M}(n)$, for $n = 1, 2, \ldots$, according to the following procedure — see Figure 5.1 for a graphical illustration. The initial multigraph $\mathfrak{M}(1)$ is a deterministic multigraph with one node and η self-loops. Multigraph $\mathfrak{M}(n)$ is constructed starting from $\mathfrak{M}(n-1)$ by adding a new node n and η new connections (edges or self-loops). Specifically, η steps are performed, and at each step node n is connected to a node randomly chosen from the set $\{1, 2, \ldots, n\}$. The intermediate multigraph obtained at steps $s = 1, 2, \ldots, \eta$, is denoted by $\mathfrak{M}(n; s)$. Accordingly, since after η steps we obtain the updated multigraph $\mathfrak{M}(n)$, we have the identity $\mathfrak{M}(n; \eta) = \mathfrak{M}(n)$. Likewise, we have $\mathfrak{M}(n; 0) = \mathfrak{M}(n-1)$.

Exploiting the procedure shown in Figure 5.1 we can argue that the adjacency matrices possess the following structure:

$$\boldsymbol{M}(n) = \left(\begin{array}{ccc|c} & & & \boldsymbol{m}_{1,n}(n) \\ & \boldsymbol{M}(n-1) & & \vdots \\ & & & \boldsymbol{m}_{n-1,n}(n) \\ \hline \boldsymbol{m}_{n,1}(n) & \cdots & \boldsymbol{m}_{n,n-1}(n) & \boldsymbol{m}_{n,n}(n) \end{array}\right), \tag{5.1}$$

with $\boldsymbol{M}(1) = \eta$. In fact, when passing from $\mathfrak{M}(n-1)$ to $\mathfrak{M}(n)$ we simply attach η new edges to the new node n, so that the number of edges $\boldsymbol{m}_{k\ell}(n-1)$ between any pair of nodes $k, \ell < n$ remains unaltered. In comparison, the adjacency matrix entries relative to the fresh node n evolve according to the following rule:

$$\boldsymbol{m}_{nk}(n) = \boldsymbol{m}_{kn}(n) = \sum_{s=1}^{\eta} \mathbb{I}(\boldsymbol{v}(n; s) = k), \quad k \in \{1, \ldots, n\} \tag{5.2}$$

where $\mathbb{I}(\cdot)$ is the indicator function (which is equal to 1 if its argument is true and is zero otherwise) and we denote by $\boldsymbol{v}(n; s)$ the particular node that becomes connected to n through the edge introduced at step s. For this reason, for any $k \leq \ell \leq n$, the (k, ℓ)-entry

of the adjacency matrix is actually determined only at iteration ℓ, and, hence, it makes sense to drop the dependence on n and write:

$$\boldsymbol{m}_{k\ell}(n) = \boldsymbol{m}_{k\ell}(\ell) \triangleq \boldsymbol{m}_{k\ell} = \boldsymbol{m}_{\ell k}. \tag{5.3}$$

Node Degrees and Preferential Attachment Rule

We adopt the standard convention that the degree of node k, denoted by $\boldsymbol{d}_{\mathcal{M},k}(n)$, is the number of edges connected to k plus twice[1] the number of self-loops [32]:

$$\boldsymbol{d}_{\mathcal{M},k}(n) = \sum_{\substack{\ell=1 \\ \ell \neq k}}^{n} \boldsymbol{m}_{k\ell} + 2\,\boldsymbol{m}_{kk}. \tag{5.4}$$

Likewise, we denote by $\boldsymbol{d}_{\mathcal{M},k}(n;s)$ the degree of node k in the *intermediate* multigraph $\boldsymbol{\mathcal{M}}(n;s)$.

At each step s, the degree of a node $k \neq n$ in the intermediate multigraph $\boldsymbol{\mathcal{M}}(n;s)$ increases by 1 if the node picked at step s is equal to k, namely,

$$\boldsymbol{d}_{\mathcal{M},k}(n;s) = \boldsymbol{d}_{\mathcal{M},k}(n;s-1) + \mathbb{I}(\boldsymbol{v}(n;s) = k). \tag{5.5}$$

In comparison, the degree of the new node n increases by 1 if the node picked at step s is equal to $k < n$, while it increases by 2 if node n itself is picked, since each self-loop is counted twice in the degree, with the initialization $\boldsymbol{d}_{\mathcal{M},n}(n;0) = 0$.

$$\boldsymbol{d}_{\mathcal{M},n}(n;s) = \boldsymbol{d}_{\mathcal{M},n}(n;s-1) + 1 + \mathbb{I}(\boldsymbol{v}(n;s) = n). \tag{5.6}$$

The description of the multigraph construction is now completed by assigning the probability that a particular node is picked. Consider first the probability that the new node n is attached to an existing node $k < n$, namely,

$$\mathbb{P}\left[\boldsymbol{v}(n;s) = k | \boldsymbol{\mathcal{M}}(n;s-1)\right] = \frac{\boldsymbol{d}_{\mathcal{M},k}(n;s-1)}{1 + \sum_{\ell=1}^{n} \boldsymbol{d}_{\mathcal{M},\ell}(n;s-1)}. \tag{5.7}$$

Let us ignore for now the term 1 appearing in the denominator. We see that the probability mass function in (5.7) matches well the preferential attachment paradigm, since we see that nodes with higher degrees in $\boldsymbol{\mathcal{M}}(n;s-1)$ are more likely to be connected to the incoming node n, and so their degrees are more likely to increase further as the multigraph construction proceeds, according to "*the rich get richer*" philosophy.

We switch to the probability that a self-loop is created on the new node n:

$$\mathbb{P}\left[\boldsymbol{v}(n;s) = n | \boldsymbol{\mathcal{M}}(n;s-1)\right] = \frac{1 + \boldsymbol{d}_{\mathcal{M},n}(n;s-1)}{1 + \sum_{\ell=1}^{n} \boldsymbol{d}_{\mathcal{M},\ell}(n;s-1)}. \tag{5.8}$$

[1] If we sum all degrees over index k in (5.4), each edge is counted twice (because $\boldsymbol{m}_{k\ell} = \boldsymbol{m}_{\ell k}$), and each self-loop is counted twice (because of the factor 2). As a result, with the adopted convention the *half*-sum of the degrees in the multigraph is exactly equal to the total number of edges and self-loops.

The term 1 in the numerator corresponds to first attaching one end of a new edge to the new node and updating the degree of that node before attaching the other end of the edge [8, 10]. Note that this term grants a nonzero probability of self-loops (we recall that $\boldsymbol{d}_{\mathcal{M},n}(n;0) = 0$) when the new node enters the system. The term 1 in the denominator is necessary to get an admissible probability mass function, i.e., to let the sum of the probabilities in (5.7) and (5.8) be equal to 1.

It is useful to provide a more explicit representation for the denominator in (5.7) and (5.8). Since we know (see footnote 1) that the half-sum of all degrees is equal to the total number of edges and self-loops, and since at each step the Bollobás-Riordan construction adds exactly η new connections, we get the following equality:

$$1 + \sum_{k=1}^{n} \boldsymbol{d}_{\mathcal{M},k}(n; s-1) = 1 + 2\eta(n-1) + 2(s-1), \tag{5.9}$$

which reveals that the denominator of the preferential attachment probability is a deterministic quantity. Finally, by merging (5.7) and (5.8) in a single equation, and using (5.9) to represent the denominator, we get, for all $k = 1, 2, \ldots, n$, the compact representation:

$$\mathbb{P}\left[\boldsymbol{v}(n;s) = k | \boldsymbol{\mathcal{M}}(n;s-1)\right] = \frac{\delta_{kn} + \boldsymbol{d}_{\mathcal{M},k}(n;s-1)}{1 + 2\eta(n-1) + 2(s-1)}, \tag{5.10}$$

where δ_{kn} is the Kronecker delta.

Maximal Degree

One fundamental graph descriptor that will play a critical role in our analysis is the *maximal degree* $\boldsymbol{\mu}_{\mathcal{M}}(N)$. In particular, we will rely on the asymptotic growth of the maximal degree with the network size N, which, as formally stated in Appendix 5.2, was found to be on the order of $\sqrt{N}$, in the following sense (see Theorem 8.14 at [112, p. 280]):

$$\frac{\boldsymbol{\mu}_{\mathcal{M}}(N)}{\sqrt{N}} \xrightarrow{\text{a.s.}} \boldsymbol{\mu}, \tag{5.11}$$

where $\xrightarrow{\text{a.s.}}$ denotes almost-sure convergence as $N \to \infty$, and $\boldsymbol{\mu}$ is a certain *positive* random variable.

The square-root growth of the maximal degree in a Bollobás-Riordan graph can be related to the well known *power-law* or *scale-free* behavior of these graphs. The power-law decay refers to the average number of nodes with degree equal to d, which was shown to scale as an inverse power of d, precisely as d^{-3} for Bollobás-Riordan graphs. It was shown in [10] that such heavy-tailed behavior, as opposed, for instance, to the exponential tail corresponding to an Erdős-Rényi graph, reflects into a faster growth of the maximal degree, namely, the $\sqrt{N}$ growth prescribed by (5.11).

From Multigraph $\mathcal{M}(N)$ to Graph $\mathcal{G}(N)$

The *multigraph* structure was chosen by Bollobás and Riordan because it was instrumental to prove a number of theoretical results [8,10]. The final goal of their model, however, was to construct a standard (i.e., simple) graph, with single edges and no self-loops. Actually, the multigraphs generated according to the Bollobás-Riordan model are approximately similar to simple graphs, since it is possible to show that the fraction of edges that are either repetitions or self-loops vanishes as N grows, as formally stated in the following lemma. Before stating the lemma, it is useful to notice that, by construction, the number of edges in $\mathfrak{M}(n)$ is equal to ηn since we start with η loops in $\mathfrak{M}(1)$ and add η new edges at a time.

Lemma 4 (Equivalence Between Bollobás-Riordan Multigraphs and Simple Graphs). *Let $\mathfrak{M}(N)$ be a Bollobás-Riordan multigraph of size N, and let $\mathring{\boldsymbol{m}}(N)$ and $\ddot{\boldsymbol{m}}(N)$ be the number of self-loops and redundant edges, respectively. Then, the number of self-loops and redundant edges are asymptotically negligible w.r.t. the total number of connections ηN, namely,*

$$\lim_{N\to\infty} \frac{\mathbb{E}[\mathring{\boldsymbol{m}}(N) + \ddot{\boldsymbol{m}}(N)]}{\eta N} = 0. \tag{5.12}$$

Proof: We start with the analysis of the number of self-loops in $\mathfrak{M}(N)$, which is equal to:

$$\mathring{\boldsymbol{m}}(N) \triangleq \sum_{k=1}^{N} \boldsymbol{m}_{kk}. \tag{5.13}$$

Given a multigraph $\mathfrak{M}(k-1)$, let us consider the η steps, $s = 1, 2, \ldots, \eta$, necessary to build the multigraph $\mathfrak{M}(k)$. Consider a sequence of nodes $v_1, v_2, \ldots, v_\eta$ selected during the η steps, with the prescription that *exactly* m out of the η nodes are equal to k, i.e., we have m self-loops attached to the new node k. We denote by $\mathcal{V}_m$ the ensemble of configurations $v_1, v_2, \ldots, v_\eta$ that match such prescription. We note in passing that the cardinality of $\mathcal{V}_m$ is equal to $\binom{\eta}{m}$. According to the adopted notation, the probability of having exactly m self-loops attached to k admits the following representation:

$$\begin{aligned}
&\mathbb{P}[\boldsymbol{m}_{kk} = m \,|\, \mathfrak{M}(k-1)] \\
&\overset{\text{(a)}}{=} \sum_{\mathcal{V}_m} \prod_{s=1}^{\eta} \mathbb{P}[\boldsymbol{v}(k;s) = v_s \,|\, \{\boldsymbol{v}(k;\tau) = v_\tau\}_{\tau=1}^{s-1}, \mathfrak{M}(k-1)] \\
&\leq \sum_{\mathcal{V}_m} \prod_{s:v_s=k} \mathbb{P}[\boldsymbol{v}(k;s) = k \,|\, \{\boldsymbol{v}(k;\tau) = v_\tau\}_{\tau=1}^{s-1}, \mathfrak{M}(k-1)] \\
&\overset{\text{(b)}}{\leq} \binom{\eta}{m} \frac{1}{(k-1)^m},
\end{aligned} \tag{5.14}$$

where (a) follows by the chain rule and (b) follows from (5.42), once noticing that $\boldsymbol{d}_{\mathcal{M},k}(k-1) = 0$. Now, by exploiting the definition of $\mathfrak{M}(1)$, the multigraph made by a single node with η self-loops, we can write:

$$\mathring{\boldsymbol{m}}(N) = \eta + \sum_{k=2}^{N} \boldsymbol{m}_{kk}, \tag{5.15}$$

which, in view of (5.14) yields:

$$\begin{aligned}
&\mathbb{E}[\mathring{\boldsymbol{m}}(N) \,|\, \mathfrak{M}(k-1)] = \\
&\eta + \sum_{k=2}^{N} \sum_{m=1}^{\eta} m \, \mathbb{P}[\boldsymbol{m}_{kk} = m \,|\, \mathfrak{M}(k-1)] \\
&\leq \eta + \sum_{k=2}^{N} \sum_{m=1}^{\eta} m \binom{\eta}{m} \left(\frac{2\eta}{k-1}\right)^{m} \\
&\leq \eta + C(\eta) \sum_{k=2}^{N} \frac{1}{k-1},
\end{aligned} \tag{5.16}$$

where the finite constant $C(\eta)$ is implicitly defined in the last step of (5.16). We immediately see from (5.16) that:

$$\lim_{N\to\infty} \frac{\mathbb{E}[\mathring{\boldsymbol{m}}(N)]}{N} = 0, \tag{5.17}$$

since the last summation in (5.16) is the harmonic number, which grows logarithmically with N.

We continue by examining the expected number of redundant edges in $\mathfrak{M}(N)$:

$$\mathbb{E}[\ddot{\boldsymbol{m}}(N)] = \sum_{\ell=1}^{N} \sum_{k=1}^{\ell-1} \mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell}], \tag{5.18}$$

where the number of redundant edges between two distinct nodes k and ℓ in the multigraph $\mathfrak{M}(\ell)$ can be conveniently represented as:

$$\ddot{\boldsymbol{m}}_{k\ell} \triangleq \max\{\boldsymbol{m}_{k\ell} - 1, 0\}, \tag{5.19}$$

We want to show that:

$$\lim_{N\to\infty} \frac{\mathbb{E}[\ddot{\boldsymbol{m}}(N)]}{N} = 0. \tag{5.20}$$

In order to prove (5.20), we can call upon the Stolz-Cesàro theorem, and apply (5.50) with the choices $f_\ell = \sum_{k=1}^{\ell-1} \mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell}]$ and $g_\ell = 1$ (which corresponds to apply the Cesàro-mean theorem), implying that it would suffice to show that:

$$\lim_{\ell\to\infty} \sum_{k=1}^{\ell-1} \mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell}] = 0. \tag{5.21}$$

Reasoning as done to obtain (5.14), we have the following representation:

$$\begin{aligned}
&\mathbb{P}[\boldsymbol{m}_{k\ell} = m \,|\, \mathfrak{M}(\ell-1)] \\
&\leq \sum_{\mathcal{V}_m} \prod_{s:v_s=k} \mathbb{P}[\boldsymbol{v}(\ell;s) = k \,|\, \{\boldsymbol{v}(\ell;\tau) = v_\tau\}_{\tau=1}^{s-1}, \mathfrak{M}(\ell-1)] \\
&\leq \binom{\eta}{m} \left(\frac{\boldsymbol{d}_{\mathcal{M},k}(\ell-1) + 2\eta}{2\eta\,(\ell-1)} \right)^m,
\end{aligned} \tag{5.22}$$

where in the last step we applied (5.42). Exploiting (5.19), from (5.22) we can compute the conditional expected value of $\ddot{\boldsymbol{m}}_{k\ell}$, obtaining:

$$\begin{aligned}
&\mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell} \,|\, \mathfrak{M}(\ell-1)] = \sum_{m=1}^{\eta-1} m\, \mathbb{P}[\ddot{\boldsymbol{m}}_{k\ell} = m \,|\, \mathfrak{M}(\ell-1)] \\
&= \sum_{m=2}^{\eta} (m-1)\, \mathbb{P}[\boldsymbol{m}_{k\ell} = m \,|\, \mathfrak{M}(\ell-1)] \\
&\leq \sum_{m=2}^{\eta} (m-1) \binom{\eta}{m} \left(\frac{\boldsymbol{d}_{\mathcal{M},k}(\ell-1) + 2\eta}{2\eta\,(\ell-1)} \right)^m \qquad (5.23) \\
&\leq \left(\frac{\boldsymbol{d}_{\mathcal{M},k}(\ell-1) + 2\eta}{2\eta\,(\ell-1)} \right) \\
&\times \sum_{m=2}^{\eta} (m-1) \binom{\eta}{m} \left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(\ell-1) + 2\eta}{2\eta\,(\ell-1)} \right)^{m-1}, \qquad (5.24)
\end{aligned}$$

where, in the last step the degree of node k has been upper bounded $m-1$ times by the maximal degree. Summing over index k, from (5.24) we get:

$$\begin{aligned}
&\sum_{k=1}^{\ell-1} \mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell} \,|\, \mathfrak{M}(\ell-1)] \leq \frac{\sum_{k=1}^{\ell-1} \left(\boldsymbol{d}_{\mathcal{M},k}(\ell-1) + 2\eta\right)}{2\eta\,(\ell-1)} \\
&\times \sum_{m=2}^{\eta} (m-1) \binom{\eta}{m} \left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(\ell-1) + 2\eta}{2\eta\,(\ell-1)} \right)^{m-1}.
\end{aligned} \tag{5.25}$$

Using (5.9) and recalling that $\boldsymbol{d}_{\mathcal{M},k}(\ell-1) = \boldsymbol{d}_{\mathcal{M},k}(\ell-1;\eta)$, we have that:

$$\sum_{k=1}^{\ell-1} \boldsymbol{d}_{\mathcal{M},k}(\ell-1) = 2\eta(\ell-1), \tag{5.26}$$

which, taking expectation w.r.t. to $\mathfrak{M}(\ell-1)$ in (5.25), yields:

$$\begin{aligned}&\sum_{k=1}^{\ell-1}\mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell}]\\&\leq 2\sum_{m=2}^{\eta}(m-1)\binom{\eta}{m}\mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(\ell-1)+\eta}{2\eta\,(\ell-1)}\right)^{m-1}\right],\end{aligned}\tag{5.27}$$

and the claim in (5.21) follows by (5.39), which, further applying (5.17), completes the proof of the lemma. ■

According to Lemma 4, it makes sense to introduce the *simple*[2] graph $\mathfrak{G}(N)$ associated to a multigraph $\mathfrak{M}(N)$, obtained by uprooting all self-loops and redundant edges from $\mathfrak{M}(N)$. The entries of the adjacency matrix $\boldsymbol{G}(N)$ of graph $\mathfrak{G}(N)$ are:

$$\boldsymbol{g}_{kk}=0,\qquad \boldsymbol{g}_{k\ell}=\min\{\boldsymbol{m}_{k\ell},1\}\quad\text{for } k\neq\ell.\tag{5.28}$$

Likewise (and coherently with (4.2)) the degree of node k in $\mathfrak{G}(n)$ and the corresponding maximal degree are, respectively:

$$\boldsymbol{d}_{\mathcal{G},k}(N)=\sum_{\ell=1}^{N}\boldsymbol{g}_{k\ell},\qquad \boldsymbol{\mu}_{\mathcal{G}}(N)=\max_{k\in[1,N]}\boldsymbol{d}_{\mathcal{G},k}(N).\tag{5.29}$$

The equivalence between $\mathfrak{M}(N)$ and $\mathfrak{G}(N)$ holds also in terms of maximal degrees, as stated in the following lemma.

Lemma 5. *Let $\boldsymbol{\mu}_{\mathcal{M}}(N)$ and $\boldsymbol{\mu}_{\mathcal{G}}(N)$ be the maximal degrees of the multigraph $\mathfrak{M}(N)$ and of the associated simple graph $\mathfrak{G}(N)$, respectively. We have that:*

$$\frac{\boldsymbol{\mu}_{\mathcal{M}}(N)-\boldsymbol{\mu}_{\mathcal{G}}(N)}{\sqrt{N}}\xrightarrow{p}0,\tag{5.30}$$

which further implies:

$$\frac{\boldsymbol{\mu}_{\mathcal{G}}(N)}{\sqrt{N}}\xrightarrow{p}\boldsymbol{\mu},\tag{5.31}$$

where $\boldsymbol{\mu}$ is the same limiting variable introduced in (5.11).

Proof: Considering the definition of the multigraph degree in (5.4), and separating the contribution of the redundant edges in (5.19) from the contribution of the simple graph

[2] In graph theory, the qualification "simple" is used to stress that the graph has no self-loops and no multiple edges.

term in (5.28), we can write:

$$\begin{aligned}\boldsymbol{d}_{\mathcal{M},k}(N) &= 2\,\boldsymbol{m}_{kk} + \sum_{\substack{\ell=1\\ \ell\neq k}}^{N} \boldsymbol{m}_{k\ell} \\ &\overset{\text{(a)}}{=} 2\,\boldsymbol{m}_{kk} + \sum_{\substack{\ell=1\\ \ell\neq k}}^{N} \boldsymbol{g}_{k\ell} + \sum_{\ell=1}^{k-1} \ddot{\boldsymbol{m}}_{k\ell} + \sum_{\ell=k+1}^{N} \ddot{\boldsymbol{m}}_{k\ell} \\ &\overset{\text{(b)}}{<} 3\,\eta + \boldsymbol{d}_{\mathcal{G},k}(N) + \sum_{\ell=k+1}^{N} \ddot{\boldsymbol{m}}_{k\ell}, \end{aligned} \tag{5.32}$$

where in (a) we adopt the convention that the second summation is equal to zero when $k = 1$, and that the third summation is equal to zero when $k = N$; and in (b) the inequality follows because the number of self-loops attached to node k at cycle k, as well as the number of edges connecting node k to a node $\ell < k$, are at most equal to the number of steps η, namely, $\boldsymbol{m}_{kk} \leq \eta$ and $\sum_{\ell=1}^{k-1} \ddot{\boldsymbol{m}}_{k\ell} < \eta$. Taking the maximum over $k \in [1, N]$ in (5.32) we can write:

$$\begin{aligned} 0 \leq \boldsymbol{\mu}_{\mathcal{M}}(N) - \boldsymbol{\mu}_{\mathcal{G}}(N) &< 3\,\eta + \max_{k\in[1,N]} \sum_{\ell=k+1}^{N} \ddot{\boldsymbol{m}}_{k\ell} \\ &= 3\,\eta + \max_{k\in[1,N-1]} \sum_{\ell=k+1}^{N} \ddot{\boldsymbol{m}}_{k\ell}, \end{aligned} \tag{5.33}$$

where the equality follows because of the summation is zero when $k = N$. For any $k \geq 1$, let us introduce the sequence:

$$\boldsymbol{u}_k(\ell) \triangleq \begin{cases} \ddot{\boldsymbol{m}}_{k\ell}, & k < \ell, \\ 0, & \text{otherwise.} \end{cases} \tag{5.34}$$

It is readily verified that the family of sequences $\{\boldsymbol{u}_k(\ell)\}_{\ell\geq 1}$ indexed by the parameter k, matches the hypotheses of Lemma 21, with the choices $\Theta = \mathbb{N} \setminus \{0\}$, $b = \eta$, and with the filtration $\{\mathcal{F}(\ell)\}_{\ell\geq 1}$ generated by the random sequence $\{\boldsymbol{\mathcal{M}}(\ell)\}_{\ell\geq 1}$. In particular, for any $\varepsilon > 0$, and for any $N \geq 1$, by choosing $\mathcal{T} = [1, N]$ and $u = \varepsilon\sqrt{N}$, we can apply Lemma 21 and write:

$$\begin{aligned} &\mathbb{P}\left[\max_{k\in[1,N-1]} \sum_{\ell=k+1}^{N} \ddot{\boldsymbol{m}}_{k\ell} > \varepsilon\sqrt{N}\right] \\ &\leq N e^{-\frac{3\varepsilon}{16\eta}\sqrt{N}} + \mathbb{P}\left[\max_{k\in[1,N-1]} \boldsymbol{C}_k(N) > \frac{\varepsilon}{2}\sqrt{N}\right], \end{aligned} \tag{5.35}$$

where

$$\boldsymbol{C}_k(N) = \sum_{\ell=k+1}^{N} \mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell}|\mathfrak{M}(\ell-1)]. \tag{5.36}$$

On the other hand, for $1 \leq k \leq N-1$ we have that:

$$\begin{aligned}
\boldsymbol{C}_k(N) &\leq \sum_{\ell=2}^{N} \mathbb{E}[\ddot{\boldsymbol{m}}_{k\ell}|\mathfrak{M}(\ell-1)] \\
&\leq \sum_{\ell=2}^{N} \sum_{m=2}^{\eta} (m-1) \binom{\eta}{m} \left(\frac{\boldsymbol{d}_{\mathfrak{M},k}(\ell-1)+2\eta}{2\eta\,(\ell-1)} \right)^m \\
&\leq \sum_{\ell=2}^{N} \sum_{m=2}^{\eta} (m-1) \binom{\eta}{m} \left(\frac{\boldsymbol{\mu}_{\mathfrak{M}}(\ell-1)+2\eta}{2\eta\,(\ell-1)} \right)^m,
\end{aligned} \tag{5.37}$$

where in the second inequality we applied (5.23), whereas the third inequality follows from the definition of maximal degree. Applying Markov's inequality and exploiting (5.37) we obtain:

$$\begin{aligned}
&\mathbb{P}\left[\max_{k\in[1,N-1]} \boldsymbol{C}_k(N) \geq \frac{\varepsilon}{2}\sqrt{N} \right] \\
&\leq \frac{2}{\varepsilon} \sum_{m=2}^{\eta} (m-1) \binom{\eta}{m} \frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathfrak{M}}(\ell)+2\eta}{\ell} \right)^m \right].
\end{aligned} \tag{5.38}$$

Substituting (5.38) into (5.35) and applying Lemma 7, from (5.33) we obtain the convergence in (5.30). Then, the convergence in (5.31) comes directly from (5.11), and the proof is complete. ■

In the following, we will refer to graph $\mathcal{G}(N)$ as Bollobás-Riordan simple graph, or simply as Bollobás-Riordan graph.

5.2 Useful Results on Bollobás-Riordan Multigraphs

We start by enunciating two properties of the maximal degree $\boldsymbol{\mu}_{\mathfrak{M}}(N)$ that will be critical in our development.

Theorem 8.14 in [112, p. 280]. *For $N = 1, 2, \ldots$, let $\boldsymbol{\mu}_{\mathfrak{M}}(N)$ be the maximal degree sequence defined over the multigraph sequence $\mathfrak{M}(N)$. For any $m \in \mathbb{N}$ we have that:*

$$\limsup_{N\to\infty} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathfrak{M}}(N)}{\sqrt{N}} \right)^m \right] < \infty. \tag{5.39}$$

Moreover, there exists a strictly positive random variable $\boldsymbol{\mu}$ such that:

$$\frac{\boldsymbol{\mu}_{\mathfrak{M}}(N)}{\sqrt{N}} \xrightarrow{\text{a.s.}} \boldsymbol{\mu}, \tag{5.40}$$

where $\xrightarrow{\text{a.s.}}$ *denotes almost-sure convergence.* □

We continue by proving a lemma that provides a uniform upper bound on the *preferential attachment probability.*

Lemma 6 (Bounds on the Preferential Attachment Probability). *The preferential attachment probability defined in* (5.10) *obeys the following bound:*

$$\mathbb{P}\left[\boldsymbol{v}(n;s)=k|\boldsymbol{\mathcal{M}}(n;s-1)\right] < \frac{\boldsymbol{d}_{\mathcal{M},k}(n-1)+2\eta}{2\eta\,(n-1)}. \tag{5.41}$$

Moreover, for any set $\mathcal{T} \subseteq \{1,2,\ldots,s-1\}$*, we have that:*

$$\mathbb{P}[\boldsymbol{v}(n;s)=k|\{\boldsymbol{v}(n;\tau)\}_{\tau\in\mathcal{T}},\boldsymbol{\mathcal{M}}(n-1)] < \frac{\boldsymbol{d}_{\mathcal{M},k}(n-1)+2\eta}{2\eta\,(n-1)}. \tag{5.42}$$

Proof: First we focus on the numerator in (5.10). Joining (5.5) and (5.6), we can write the degree of node k in multigraph $\boldsymbol{\mathcal{M}}(n,s-1)$ as:

$$\boldsymbol{d}_{\mathcal{M},k}(n;s-1) = \boldsymbol{d}_{\mathcal{M},k}(n;s-2) + \mathbb{I}(\boldsymbol{v}(n;s-1)=k) + \delta_{kn}, \tag{5.43}$$

where δ_{kn} is the Kronecker delta. Developing the recursion in (5.43) over index s, we get:

$$\begin{aligned}\boldsymbol{d}_{\mathcal{M},k}(n;s-1) &= \boldsymbol{d}_{\mathcal{M},k}(n-1) + \sum_{t=1}^{s-1}\mathbb{I}(\boldsymbol{v}(n;t)=k)\\ &+ (s-1)\delta_{kn} \leq \boldsymbol{d}_{\mathcal{M},k}(n-1) + 2(s-1),\end{aligned} \tag{5.44}$$

which implies that the numerator in (5.10) is upper bounded as:

$$\begin{aligned}\delta_{kn} + \boldsymbol{d}_{\mathcal{M},k}(n;s-1) &\leq \boldsymbol{d}_{\mathcal{M},k}(n-1) + 2s - 1\\ &< \boldsymbol{d}_{\mathcal{M},k}(n-1) + 2\eta,\end{aligned} \tag{5.45}$$

where the last inequality follows by observing that $s \leq \eta$.

We switch to the analysis of the denominator in (5.10), which is lower bounded as:

$$1 + 2\eta(n-1) + 2(s-1) > 2\eta\,(n-1). \tag{5.46}$$

Using (5.45) and (5.46) in (5.10), we get (5.41). It remains to prove (5.42). By applying the law of total probability, we can write:

$$\begin{aligned}&\mathbb{P}\left[\boldsymbol{v}(n;s)=k|\{\boldsymbol{v}(n;\tau)\}_{\tau\in\mathcal{T}},\boldsymbol{\mathcal{M}}(n-1)\right]\\ &\overset{\text{(a)}}{=} \sum_{\mathcal{M}}\mathbb{P}\left[\boldsymbol{v}(n;s)=k|\boldsymbol{\mathcal{M}}(n,s-1)=\mathcal{M}\right]\\ &\times\mathbb{P}\left[\boldsymbol{\mathcal{M}}(n,s-1)=\mathcal{M}|\{\boldsymbol{v}(n;\tau)\}_{\tau\in\mathcal{T}},\boldsymbol{\mathcal{M}}(n-1)\right]\\ &\overset{\text{(b)}}{<} \frac{\boldsymbol{d}_{\mathcal{M},k}(n-1)+2\eta}{2\eta\,(n-1)}.\end{aligned} \tag{5.47}$$

In the summation, $\mathcal{M}$ spans the space of possible multigraphs $\mathfrak{M}(n, s-1)$ compatible with the multigraph $\mathfrak{M}(n-1)$ and the collection of selected nodes $\{\boldsymbol{v}(n;\tau)\}_{\tau\in\mathcal{T}}$, and where: (a) comes from the fact that the preferential attachment rule is Markovian, namely, the selection at step s depends only on the previous multigraph $\mathfrak{M}(n; s-1)$ (regardless of any details about the previous multigraph evolution); while (b) comes from (5.41). ■

Lemma 7 (Sum of Maximal Degree Powers). *For all $m \geq 2$ we have that:*

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(n)+2\eta}{n}\right)^{m}\right] = 0. \tag{5.48}$$

Proof: First we observe that:

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(n)+2\eta}{n}\right)^{m}\right] \\
&= \frac{\sum_{n=1}^{N} n^{-m/2}}{\sqrt{N}} \quad \frac{\displaystyle\sum_{n=1}^{N} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(n)+2\eta}{n}\right)^{m}\right]}{\sum_{n=1}^{N} n^{-m/2}}.
\end{aligned} \tag{5.49}$$

From the Stolz-Cesàro theorem, for any two positive sequences f_N and g_N with $g_N \to \infty$ as $N \to \infty$, we have that [17, 109]:

$$\limsup_{N\to\infty} \frac{\sum_{n=1}^{N} f_n}{\sum_{n=1}^{N} g_n} \leq \limsup_{N\to\infty} \frac{f_N}{g_N}, \tag{5.50}$$

which applied to the last fraction in (5.49) yields:

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{\displaystyle\sum_{n=1}^{N} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(n)+2\eta}{n}\right)^{m}\right]}{\sum_{n=1}^{N} n^{-m/2}} \\
&\leq \limsup_{N\to\infty} \frac{\mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(N)+2\eta}{N}\right)^{m}\right]}{N^{-m/2}} \\
&= \limsup_{N\to\infty} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(N)+2\eta}{\sqrt{N}}\right)^{m}\right] < \infty,
\end{aligned} \tag{5.51}$$

where the last inequality follows by (5.39).

Focusing on the first fraction in (5.49) we have that, for $m > 2$:

$$\sum_{n=1}^{\infty} n^{-m/2} = \zeta(m/2), \tag{5.52}$$

where $\zeta(\cdot)$ is the Riemann zeta function (which is finite), while for $m = 2$ the summation in (5.52) is the harmonic number, which diverges logarithmically as $N \to \infty$. Accordingly, we conclude that for all $m \geq 2$:

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} n^{-m/2}}{\sqrt{N}} = 0. \tag{5.53}$$

The claim of the lemma now follows by applying (5.51) and (5.53). ■

Lemma 8 (Correlation Between Degrees). *For any $1 \leq k < \ell \leq N$, we have that:*

$$\mathbb{E}[\,\boldsymbol{m}_{kN}\boldsymbol{m}_{\ell N}\,|\,\boldsymbol{\mathcal{M}}(N-1)] < \left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(N-1)+2\eta}{N-1}\right)^2. \tag{5.54}$$

Proof: It is convenient to introduce the following binary random variables:

$$\boldsymbol{\beta}_{ks}(N) \triangleq \mathbb{I}(\boldsymbol{v}(N;s)=k), \tag{5.55}$$

for $1 \leq k \leq N$ and $1 \leq s \leq \eta$. Using (5.2), we can write:

$$\begin{aligned}
&\mathbb{E}\,[\boldsymbol{m}_{kN}\boldsymbol{m}_{\ell N}|\boldsymbol{\mathcal{M}}(N-1)] \\
&= \mathbb{E}\left[\sum_{s=1}^{\eta}\boldsymbol{\beta}_{ks}(N)\sum_{s=1}^{\eta}\boldsymbol{\beta}_{\ell s}(N)\middle|\boldsymbol{\mathcal{M}}(N-1)\right] \\
&= \sum_{\substack{1\leq s,t\leq\eta\\ s\neq t}} \mathbb{E}\,[\boldsymbol{\beta}_{ks}(N)\boldsymbol{\beta}_{\ell t}(N)|\boldsymbol{\mathcal{M}}(N-1)]\,,
\end{aligned} \tag{5.56}$$

where the last step holds true since $k \neq \ell$ by assumption. Let us examine the behavior of the individual term in (5.56) and focus, without loss of generality, on the case $s > t$. Since $\boldsymbol{\beta}_{ks}(N)$ and $\boldsymbol{\beta}_{\ell t}(N)$ are indicator variables, we have that:

$$\begin{aligned}
&\mathbb{E}\,[\boldsymbol{\beta}_{ks}(N)\boldsymbol{\beta}_{\ell t}(N)|\boldsymbol{\mathcal{M}}(N-1)] \\
&= \mathbb{P}[\boldsymbol{v}(N;s)=k,\boldsymbol{v}(N;t)=\ell|\boldsymbol{\mathcal{M}}(N-1)] \\
&= \mathbb{P}[\boldsymbol{v}(N;s)=k|\boldsymbol{v}(N;t)=\ell,\boldsymbol{\mathcal{M}}(N-1)] \\
&\times \mathbb{P}[\boldsymbol{v}(N;t)=\ell|\boldsymbol{\mathcal{M}}(N-1)] \\
&< \frac{\eta(\eta-1)}{2}\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(N-1)+2\eta}{2\eta\,(N-1)}\right)^2,
\end{aligned} \tag{5.57}$$

where the last inequality follows from (5.42), and the claim follows by observing that $\eta(\eta-1)/(8\eta^2) < 1$. ■

5.3 Asymptotic Concentration of the Laplacian

In this section we show that the Laplacian matrix, when defined over the Bollobás-Riordan graph presented in Section 5.1, asymptotically concentrates according to Definition 1.

Lemma 9 (Asymptotic Concentration of the Laplacian Combination Matrix over Bollobás-Riordan Graphs). *Consider the case when the support graph of the Laplacian combination matrix (4.15) is the simple graph $\mathcal{G}(N)$ obtained from a Bollobás-Riordan multigraph $\mathcal{M}(N)$ with step parameter η. Then the Laplacian combination matrix satisfies Definition 1 with scaling sequence $c_N \triangleq \sqrt{N}$ and identifiability gap:*

$$\boldsymbol{\gamma} \triangleq \frac{\rho\lambda}{\boldsymbol{\mu}}, \tag{5.58}$$

where $\boldsymbol{\mu}$ is given by (5.11).

Proof: From definition (4.15) of the Laplacian matrix we have that:

$$\boldsymbol{A}(N) - \operatorname{diag}(\boldsymbol{A}(N)) = \rho\lambda \frac{\boldsymbol{G}(N)}{1 + \boldsymbol{\mu}_{\mathcal{G}}(N)}. \tag{5.59}$$

Therefore, the term:

$$\|\sqrt{N}\boldsymbol{A}(N) - \boldsymbol{\gamma}\boldsymbol{G}(N)\|_{\text{max-off}}, \tag{5.60}$$

where $\boldsymbol{\gamma}$ is the identifiability gap introduced in Definition 1, is upper bounded by:

$$\left| \frac{\rho\lambda\sqrt{N}}{1 + \boldsymbol{\mu}_{\mathcal{G}}(N)} - \boldsymbol{\gamma} \right| \xrightarrow{\text{p}} 0, \tag{5.61}$$

with the convergence following from (5.40) and (5.30) since $\boldsymbol{\mu}$ is strictly positive. Thus, we have:

$$\|\sqrt{N}\boldsymbol{A}(N) - \boldsymbol{\gamma}\boldsymbol{G}(N)\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{5.62}$$

which concludes the proof. ■

5.4 Assumptions on the Probed Subset $\mathcal{S}_N$

As done for Erdős-Rényi graphs, we focus on the case where the fraction of probed nodes converges to some value ξ. Notably, this value is arbitrary, i.e., is allowed to be arbitrarily small. The next lemma shows that under this assumption for the probed subsets, the subgraph $\mathcal{G}_{\mathcal{S}}(N)$ is nontrivial w.h.p. as $N \to \infty$.

Lemma 10 (Nontrivial Bollobás-Riordan Subgraphs). *Let $\mathcal{S}_N$ be a sequence of subsets satisfying (1.5) with:*

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \textit{for some } 0 < \xi < 1, \tag{5.63}$$

and let $\mathcal{G}(N)$ be a random graph sequence associated with a Bollobás-Riordan multigraph $\mathcal{M}(N)$. Then: i) for sufficiently large N, subgraph $\mathcal{G}_{\mathcal{S}}(N)$ is not fully connected; and ii) the probability that $\mathcal{G}_{\mathcal{S}}(N)$ is fully disconnected vanishes as $N \to \infty$.

Proof: The sum of all degrees in multigraph $\mathcal{M}(N)$ grows linearly with N. In order to have a fully connected subgraph $\mathcal{G}_{\mathcal{S}}(N)$, the sum of all degrees of this subgraph should be equal to $|\mathcal{S}_N|^2$, which scales as N^2 in view of (5.63). Therefore, subgraph $\mathcal{G}_{\mathcal{S}}(N)$ cannot be fully connected as N grows.

We move on to examine the probability that $\mathcal{G}_{\mathcal{S}}(N)$ is fully disconnected. Let

$$\mathcal{S}_N = \{n_1, n_2, \ldots, n_{|\mathcal{S}_N|}\}, \qquad \mathcal{V}_\ell \triangleq \{n_1, n_2, \ldots, n_\ell\}, \tag{5.64}$$

with:

$$n_1 < n_2 < \ldots < n_{|\mathcal{S}_N|}. \tag{5.65}$$

We can write:

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully disconnected}] \\
&\leq \mathbb{P}\left[\boldsymbol{v}(n_2;1) \notin \mathcal{V}_1, \boldsymbol{v}(n_3;1) \notin \mathcal{V}_2, \ldots, \boldsymbol{v}(n_{|\mathcal{S}_N|};1) \notin \mathcal{V}_{|\mathcal{S}_N|-1}\right] \\
&= \prod_{\ell=2}^{|\mathcal{S}_N|} \mathbb{P}\left[\boldsymbol{v}(n_\ell;1) \notin \mathcal{V}_{\ell-1} | \{\boldsymbol{v}(n_k;1) \notin \mathcal{V}_{k-1}\}_{k=2}^{\ell-1}\right],
\end{aligned} \tag{5.66}$$

where the inequality follows by considering only the connections at step $s = 1$ of the preferential attachment construction, whereas the equality follows by the chain rule. Now, from (5.10), for any $k < n$ we obtain the following lower bound:

$$\mathbb{P}\left[\boldsymbol{v}(n;1) = k | \mathcal{M}(n-1)\right] \geq \frac{\eta}{1 + 2\eta(n-1)} \geq \frac{1}{2n}. \tag{5.67}$$

Accordingly, the individual term in (5.66) can be upper bounded as follows:

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{v}(n_\ell;1) \notin \mathcal{V}_{\ell-1} | \{\boldsymbol{v}(n_k;1) \notin \mathcal{V}_{k-1}\}_{k=2}^{\ell-1}\right] \\
&= 1 - \sum_{k=1}^{\ell-1} \mathbb{P}\left[\boldsymbol{v}(n_\ell;1) = n_k | \{\boldsymbol{v}(n_k;1) \notin \mathcal{V}_{k-1}\}_{k=2}^{\ell-1}\right] \\
&\leq 1 - \frac{\ell-1}{2n_\ell} \leq 1 - \frac{\ell-1}{2N}.
\end{aligned} \tag{5.68}$$

Using (5.68) in (5.66), we finally get:[3]

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully disconnected}] \\
&\leq \prod_{\ell=2}^{|\mathcal{S}_N|} \left(1 - \frac{\ell-1}{2N}\right) \leq \prod_{\ell=|\mathcal{S}_N|/2+1}^{|\mathcal{S}_N|} \left(1 - \frac{\ell-1}{2N}\right) \\
&\leq \left(1 - \frac{|\mathcal{S}_N|}{4N}\right)^{|\mathcal{S}_N|/2} \overset{N\to\infty}{\longrightarrow} 0,
\end{aligned} \tag{5.70}$$

where the second inequality holds because all terms in the product are smaller than 1, while convergence holds in view of (5.63). ■

5.5 Regularity of the Limiting Granger Estimator

Lemma 11 (Regularity of the Limiting Granger Estimator over Bollobás-Riordan graphs). *Consider the case when the support graph of the Laplacian combination matrix (4.15) is the simple graph $\mathcal{G}(N)$ obtained from a Bollobás-Riordan multigraph $\mathfrak{M}(N)$ with step parameter η. Then for any sequence of probed subsets $\mathcal{S}_N$ the limiting Granger estimator satisfies:*

$$\sqrt{N}\|\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N)\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{5.71}$$

namely, it fulfills Definition 5 for any sequence $\mathcal{S}_N$, with bias $\beta = 0$.

Proof: Since Lemma 16 holds for any symmetric matrix A fulfilling (B.1) and any subset $\mathcal{S}_N \subseteq \{1, 2, \ldots, N\}$, in view of (B.46) we can write:

$$\|\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N)\|_{\text{max-off}} \leq \kappa\, \mathfrak{M}\big(\boldsymbol{A}(N)\big), \tag{5.72}$$

where $\mathfrak{M}(A)$ is defined in (B.2). Applying Lemma 17 to (5.72), we conclude that:

$$\sqrt{N}\|\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N)\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{5.73}$$

for *any sequence of probed subsets $\mathcal{S}_N$.* ■

[3] For each multigraph $\mathfrak{M}(n_\ell - 1)$ such that $\{\boldsymbol{v}(n_k; 1) \notin \mathcal{V}_{k-1}\}_{k=2}^{\ell-1}$, we can write:

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{v}(n_\ell; 1) = n_k | \{\boldsymbol{v}(n_k; 1) \notin \mathcal{V}_{k-1}\}_{k=2}^{\ell-1}, \mathfrak{M}(n_\ell - 1)\right] \\
&= \mathbb{P}\left[\boldsymbol{v}(n_\ell; 1) = n_k | \mathfrak{M}(n_\ell - 1)\right].
\end{aligned} \tag{5.69}$$

Since the bound in (5.67) does not depend on $\mathfrak{M}(n-1)$, by applying the law of total probability, we can use (5.67) to bound also the probability $\mathbb{P}\left[\boldsymbol{v}(n_\ell; 1) = n_k | \{\boldsymbol{v}(n_k; 1) \notin \mathcal{V}_{k-1}\}_{k=2}^{\ell-1}\right]$.

5.6 Achievability for Bollobás-Riordan Graphs

We are now ready to state the main achievability result involving Bollobás-Riordan graphs.

Theorem 7 (Achievability for Bollobás-Riordan Graphs). *Let us consider the dynamical system (1.1), with Laplacian combination matrix as in (4.15), and with network graph $\mathcal{G}(N)$ being a simple graph obtained from a Bollobás-Riordan multigraph $\mathfrak{M}(N)$ with step parameter η. Then, for any probed subset sequence $\mathcal{S}_N$ such that:*

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \textit{for some } 0 < \xi < 1, \tag{5.74}$$

the graph estimator:

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T, N) = \text{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)\Big), \tag{5.75}$$

where graphclu$(\cdot)$ *is the clustering procedure in Listing 2, and $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T, N)$ is the sample Granger estimator (3.28), satisfies the consistency property (1.6):*

$$\lim_{N\to\infty} \mathbb{P}\left[\widehat{\mathcal{G}}_{\mathcal{S}}(T_N, N) = \mathcal{G}_{\mathcal{S}}(N)\right] = 1. \tag{5.76}$$

for some scaling law T_N.

Proof: This result is a direct consequence of Lemmas 9, 10 and 11, applied to Corollary 1, along with the fact that the sample Granger estimator (3.28) converges to the limiting Granger estimator (3.24) by ergodicity. ■

Equation (5.11) reveals that the *limiting* (scaled) maximal degree $\boldsymbol{\mu}$ is an *intrinsic* property of the specific Bollobás-Riordan multigraph instance. In other words, as the Bollobás-Riordan multigraph construction progresses, the maximal degree, scaled by $\sqrt{N}$, tends to become stable, and converges to a certain value $\boldsymbol{\mu}$. However, this value is *random*, implying that if we repeat the Bollobás-Riordan multigraph construction with the same parameters, we obtain a *different* value for $\boldsymbol{\mu}$. In view of (5.58), this implies that the value of the identifiability gap that is critical for graph learning purposes is random as well, i.e., it depends on the particular graph sequence. This is a fundamental difference that distinguishes the behavior of Bollobás-Riordan graphs from the behavior of Erdős-Rényi graphs, where the identifiability gap is instead deterministic and independent of the particular graph realization. However, and remarkably, we already know from Theorem 7 that randomness of the gap does not impair the possibility of consistent graph recovery.

We observe that the coupling between Bollobás-Riordan graphs (which are undirected) and the Laplacian rule gives rise to a symmetric combination matrix. Under this assumption, the series for the covariance matrix in (3.21) can be computed as (I denotes the $N \times N$ identity matrix):

$$\boldsymbol{R}_0(N) = \sum_{i=0}^{\infty} \boldsymbol{A}^{2i}(N) = \big(I - \boldsymbol{A}^2(N)\big)^{-1}, \tag{5.77}$$

whose structure is exploited in the proofs of our results. While symmetry is useful to develop these technical arguments, we remark that the Granger estimator is based upon relation (3.22), which does *not* rely on symmetry at all. This observation, along with the series structure in (3.21) that is similar to the structure exploited in Appendix B for the symmetric case, suggests that the Granger estimator can work also with non-symmetric matrices, as we will show in Section 7.2 by examining a *directed* version of Bollobás-Riordan graphs.

Chapter 6

Sample Complexity

We are now ready to illustrate our analysis of the sample complexity of the regularized Granger estimator operating over Bollobás-Riordan graphs. Following [6,49,61,75,78,95], the analysis is performed under Assumption 1, which is classical for systems like (1.1).

6.1 Preliminary Results

The following lemmas characterize the rate of convergence of the sample covariance estimators. Preliminarily, it is convenient to introduce the following auxiliary function:

$$f_T(x) \triangleq |\mathcal{S}_N|^2 e^{-T/2} + |\mathcal{S}_N|^2 e^{-\left[\sqrt{T}x - \sqrt{2}\right]^2}, \tag{6.1}$$

and the error matrices, for $j \in \{0, 1\}$:

$$\boldsymbol{E}_j \triangleq \left[\widehat{\boldsymbol{R}}_j(T, N) - [\boldsymbol{R}_j(N)\right]_{\mathcal{S}_N}. \tag{6.2}$$

Lemma 12 (Sample Covariance Errors). *Let us consider the dynamical system (1.1) operating under Assumption 1, with Laplacian combination matrix as in (4.15), and with network graph $\mathcal{G}(N)$ being a simple graph obtained from a Bollobás-Riordan multigraph $\mathfrak{M}(N)$ with step parameter η. Then there exists a constant C such that:*

$$\mathbb{P}[\|\boldsymbol{E}_0\|_{\max} > \epsilon] \leq 3\, f_T\big(\epsilon\, C\big) \qquad \text{for } T > \frac{2}{(\epsilon\, C)^2}, \tag{6.3}$$

$$\mathbb{P}[\|\boldsymbol{E}_1\|_{\max} > \epsilon] \leq 4\, f_{T-1}\big(\epsilon\, C\big) \quad \text{for } T > 1 + \frac{2}{(\epsilon\, C)^2}. \tag{6.4}$$

Proof: In this proof we will often consider conditional probabilities given $\boldsymbol{A}(N) = A$. This is tantamount to assuming that the dynamical system (1.2) is run with a *deterministic* matrix A. In such a scenario, matrix $\boldsymbol{R}_0(N)$ becomes deterministic, since according (3.21) it is a deterministic function of $\boldsymbol{A}(N)$. Accordingly, we will conveniently denote

by the normal-font symbol R_0 the realization of $\boldsymbol{R}_0(N)$ corresponding to A. In contrast, the quantity $\boldsymbol{R}_0(T,N)$ in (3.27) will remain random, since by definition it depends also on the source of randomness given by $\{\boldsymbol{x}_t(N)\}_{t=1}^T$ and $\boldsymbol{y}_0(N)$.

The proof of (6.3) and (6.4) is a slight variation of the bounding technique used in Lemma 1 of [49]. In particular, let us define for $t,\tau = 0,1,\ldots,$ with $\tau \leq t$, the conditional cross-covariance between $\boldsymbol{y}_t(N)$ and $\boldsymbol{y}_\tau(N)$, namely,

$$\Sigma_{t,\tau} \triangleq \mathbb{E}\left[\boldsymbol{y}_t(N)\boldsymbol{y}_\tau^\top(N)|\boldsymbol{A}(N)=A\right] = A^{t-\tau}\Sigma_{\tau,\tau} = A^{t-\tau}R_0, \tag{6.5}$$

where the intermediate equality is a classical result on vector autoregressive models [63], while the last equality comes from the enforced stationarity assumption. Starting from (6.5), in Lemma 1 of [49] the following bound is used:

$$\|\Sigma_{t,\tau}\|_{\max} \leq \|\Sigma_{t,\tau}\|_2 \leq \|A^{t-\tau}\|_2\|R_0\|_2. \tag{6.6}$$

In our case we can exploit additional constraints on A to replace (6.6) by:

$$\|\Sigma_{t,\tau}\|_{\max} \leq \|A^{t-\tau}\|_\infty\|R_0\|_{\max} = \rho^{t-\tau}\max_{k=1,2,\ldots,N}[R_0]_{kk}, \tag{6.7}$$

where the inequality comes from the fact that, for any two matrices M_1, M_2 of compatible dimensions, we have $\|M_1M_2\|_{\max} \leq \|M_1\|_\infty\|M_2\|_{\max}$. The equality in (6.7) follows by using (B.8) and by applying Cauchy-Schwarz inequality to obtain $|[R_0]_{k\ell}| \leq \sqrt{[R_0]_{kk}[R_0]_{\ell\ell}}$.

Using (6.7) in place of (6.6), and leaving other arguments in the proof of Lemma 1 of [49] unaltered, we get, for any T such that:[1]

$$T > \frac{2}{\left(\epsilon\,\varphi(R_0)\right)^2}, \tag{6.8}$$

the following bound:

$$\mathbb{P}[|[\boldsymbol{E}_0]_{k\ell}|>\epsilon|\boldsymbol{A}(N)=A] \leq 3\left(e^{-T/2}+e^{-\left[\sqrt{T}\epsilon\,\varphi(R_0)-\sqrt{2}\right]^2}\right), \tag{6.9}$$

where:

$$\varphi(R_0) \triangleq \frac{1-\rho}{16\sqrt{2}}\frac{\min\limits_{k=1,2,\ldots,N}[R_0]_{kk}}{\max\limits_{k=1,2,\ldots,N}[R_0]_{kk}^2}. \tag{6.10}$$

From (5.77), (B.3), and (B.29) we have the inequalities:

$$\min_{k=1,2,\ldots,N}[R_0]_{kk} \geq 1, \qquad \max_{k=1,2,\ldots,N}[R_0]_{kk} \leq \bar{\alpha}, \tag{6.11}$$

which can be used to bound the quantity $\varphi(R_0)$ in (6.10) as:

$$\varphi(R_0) \geq \frac{1}{\bar{\alpha}^2}\frac{1-\rho}{16\sqrt{2}} \triangleq C. \tag{6.12}$$

[1] Condition (6.8) is explicitly stated in Lemma 3 of [49], and basically requires that the function $e^{-(\sqrt{T}x-\sqrt{2})^2}$ appearing in (6.9) is evaluated in the region where it is decreasing, i.e., for $x > \sqrt{2/T}$.

Since the function $e^{-(\sqrt{T}x-\sqrt{2})^2}$ is decreasing for any $x > \sqrt{2/T}$, we conclude from (6.9) and (6.12) that, under the condition on T in (6.3), we have:

$$\mathbb{P}[|[\boldsymbol{E}_0]_{k\ell}| > \epsilon | \boldsymbol{A}(N) = A] \leq 3\left(e^{-T/2} + e^{-\left[\sqrt{T}\epsilon\, C - \sqrt{2}\right]^2}\right), \tag{6.13}$$

which is a bound independent of the current realization A. Therefore, by applying the law of total probability in (6.13) we get:

$$\mathbb{P}\left[|[\boldsymbol{E}_0]_{k\ell}| > \epsilon\right] \leq 3\left(e^{-T/2} + e^{-\left[\sqrt{T}\epsilon\, C - \sqrt{2}\right]^2}\right). \tag{6.14}$$

Now, using the union bound over the set of probed nodes $\mathcal{S}_N$ we can write:

$$\begin{aligned} \mathbb{P}[\|\boldsymbol{E}_0\|_{\max} > \epsilon] &\leq \sum_{k,\ell \in [1,N]} \mathbb{P}\left[|[\boldsymbol{E}_0]_{k\ell}| > \epsilon\right] \\ &\leq 3|\mathcal{S}_N|^2 \left(e^{-T/2} + e^{-\left[\sqrt{T}\epsilon\, C - \sqrt{2}\right]^2}\right), \end{aligned} \tag{6.15}$$

and the claim in (6.3) follows from the definition of f_T in (6.1). In order to obtain (6.4), we must apply the same steps shown above to the proof of Lemma 2 of [49]. ■

Lemma 13 (Scaling Law Useful for Sample Complexity). *Assume the same conditions used in Lemma 12, and consider the following scaling law for the number of samples:*

$$T_N = \omega_N N \log N, \tag{6.16}$$

for some positive sequence ω_N diverging in an arbitrarily slow fashion as $N \to \infty$. Then, for any sequence $\mathcal{S}_N$ such that:

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \textit{for some } 0 < \xi < 1, \tag{6.17}$$

and for $j \in \{0, 1\}$, we have that:

$$\sqrt{N}\|[\boldsymbol{R}_j(T_N, N)]_{\mathcal{S}} - [\boldsymbol{R}_j(N)]_{\mathcal{S}}\|_{\max} \xrightarrow{\text{p}} 0, \tag{6.18}$$

Proof: We need to show that, for any $\delta > 0$:

$$\lim_{N\to\infty} \mathbb{P}\left[\sqrt{N}\|[\boldsymbol{R}_j(T_N, N)]_{\mathcal{S}} - [\boldsymbol{R}_j(N)]_{\mathcal{S}}\|_{\max} > \delta\right] = 0. \tag{6.19}$$

We will prove the claim with reference to the case $j = 0$, with the proof being identical for $j = 1$. Let us consider Lemma 12 with the choice $\epsilon = \delta/\sqrt{N}$. Since Eq. (6.16) implies that:

$$T_N(\epsilon\, C)^2 = T_N \frac{(\delta\, C)^2}{N} \overset{N\to\infty}{\longrightarrow} \infty, \tag{6.20}$$

we see that condition on T in (6.3) is met for N sufficiently large. We conclude that to prove (6.19) it suffices to show that, for any $\delta > 0$:

$$\lim_{N\to\infty} f_{T_N}\left(\frac{\delta\, C}{\sqrt{N}}\right) = 0. \tag{6.21}$$

Now, the first term on the RHS of (6.1) converges to zero since T_N in (6.16) tends to $+\infty$ faster than $\log|\mathcal{S}_N|$. On the other hand, the second term on the RHS of (6.1) can be written as:

$$\begin{aligned}&\exp\left\{-\left(\sqrt{T_N}\,\frac{\delta\, C}{\sqrt{N}}-\sqrt{2}\right)^2+\log|\mathcal{S}_N|^2\right\}\\ &=\exp\left\{-\left[\left(\sqrt{\delta^2C^2\frac{\omega_N\log N}{\log|\mathcal{S}_N|}}-\frac{1}{\sqrt{\log|\mathcal{S}_N|}}\right)^2-1\right]\log|\mathcal{S}_N|^2\right\},\end{aligned} \tag{6.22}$$

and vanishes as $N\to\infty$ in view of (6.17) and the fact that $\omega_N\to\infty$ by assumption. ■

6.2 Sample Complexity of the Regularized Granger Estimator

Theorem 8 (Sample Complexity of the Regularized Granger Estimator for Bollobás-Riordan Graphs). *Let us consider the dynamical system (1.1) operating under Assumption 1, with Laplacian combination matrix as in (4.15), and with network graph $\mathcal{G}(N)$ being a simple graph obtained from a Bollobás-Riordan multigraph $\mathfrak{M}(N)$ with step parameter η. Then, for any probed subset sequence $\mathcal{S}_N$ the graph estimator:*

$$\widehat{\mathcal{G}}_{\mathcal{P}}(T,N) = \text{graphclu}\Big(\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)\Big), \tag{6.23}$$

where graphclu$(\cdot)$ *is the clustering procedure in Listing 2, and $\widehat{\boldsymbol{A}}_{\mathcal{P}}(T,N)$ is the regularized Granger estimator in (3.28), is consistent with sample complexity law:*

$$T_N = \omega_N N\log N, \tag{6.24}$$

where ω_N can be chosen as a positive sequence diverging in an arbitrarily slow fashion.

Proof: Calling upon Lemma 11 in [75], we have the following bound:

$$\begin{aligned}&\sqrt{N}\,\|\widehat{\boldsymbol{A}}_{\mathcal{S}}(T,N)-\widehat{\boldsymbol{A}}_{\mathcal{S}}(N)\|_{\max}\\ &\leq 2\|[\boldsymbol{R}_0]_{\mathcal{S}}^{-1}\|_1\sqrt{N}\,(\|[\boldsymbol{R}_0(T,N)]_{\mathcal{S}}-[\boldsymbol{R}_0(N)]_{\mathcal{S}}\|_{\max}\\ &+\|[\boldsymbol{R}_1(T,N)]_{\mathcal{S}}-[\boldsymbol{R}_1(N)]_{\mathcal{S}}\|_{\max})\,.\end{aligned} \tag{6.25}$$

Moreover, in [75, Eq. (321)], it was shown that:

$$\|[\boldsymbol{R}_0]_{\mathcal{S}}^{-1}\|_1 \leq 1 + \rho^2. \tag{6.26}$$

If we now use (6.26) in (6.25), from Lemma 13 we conclude that:

$$\sqrt{N}\|\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N) - \widehat{\boldsymbol{A}}_{\mathcal{S}}(N)\|_{\max} \xrightarrow{\text{p}} 0. \tag{6.27}$$

From Theorems 9 and 11 we have, by application of the triangle inequality, that:

$$\|\sqrt{N}\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N) - \boldsymbol{\gamma}\boldsymbol{G}_{\mathcal{S}}(N)\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{6.28}$$

which in turn implies (see footnote 1) that (A.1) holds true with high probability as $N \to \infty$ with the sample scaling law in (6.16). This means that the clustering algorithm graphclu is able to reconstruct correctly the subgraph of probed nodes, provided that the latter is neither fully connected nor fully disconnected. However, the probability that $\mathcal{G}_{\mathcal{S}}(N)$ is fully connected or fully disconnected vanishes $N \to \infty$ in view of Lemma 10, and the proof is complete. ■

Let us comment on the main ramifications of Theorem 8. First of all, since the sequence ω_N can grow in an arbitrarily slow fashion, *any* sample complexity that scales slightly faster than $N \log N$ achieves consistency. Therefore, the bottom line of Theorem 8 is that the sample complexity of the proposed estimator is *essentially linear*. Let us now see where this linear law originates from.

According to the Laplacian matrix structure in (4.15), the growth of the maximal degree determines the way the nonzero entries of the combination matrix shrink down as $N \to \infty$. The smaller they are, the higher is the precision required by the sample estimators to distinguish the nonzero entries from the zero entries. For this reason, faster scaling laws of the maximal degree become more demanding in terms of number of samples. This argument can be made rigorous, and is in fact exploited in the proof of Theorem 8 to show that the sample complexity goes essentially (i.e., up to a $\log N$ factor) as $\boldsymbol{\mu}_{\mathcal{G}}^2(N)$. As a result, the $\sqrt{N}$-growth of the maximal degree over Bollobás-Riordan graphs reflects into a final sample complexity that is essentially linear in N.

In summary, from a technical viewpoint we conclude that the main factor influencing sample complexity is the maximal degree of the graph. However, it is useful to relate this behavior to more "physical" attributes of the system, to capture the factors that play a domineering role on sample complexity. One important attribute of Bollobás-Riordan graphs is *sparsity*. Bollobás-Riordan graphs are *very sparsely* connected, since, over a total number of possible $N(N-1)/2$ edges, only ηN edges are drawn, which results into a sparsity ratio (no. of connected edges over total no. of possible edges) scaling as $1/N$. The sporadic presence of connections might suggest that the nodes have small degree, which, in the light of the previous discussion, would suggest a slow growth of the maximal degree. However, this conclusion is not precise. To understand why, it is useful to contrast Bollobás-Riordan graphs against Erdős-Rényi graphs. We consider in particular Erdős-Rényi graphs under the degree concentration regime because under

this regime results on sample complexity are available [75, 78]. Erdős-Rényi graphs are built homogeneously (i.e., presence/absence of edges is established in an i.i.d. manner). This homogeneity implies in particular that the maximal and average degree of an Erdős-Rényi graph scale comparably, i.e., $\boldsymbol{\mu}_{\mathcal{G}}(N) \sim Np$, where p is the connection probability. Accordingly, over *sparse* Erdős-Rényi graphs where $p \approx (\log N)/N$, the sample complexity is *polylogarithmic* in N, whereas over *dense* graphs with constant p it is *almost-quadratic* in N.

Let us see what happens over Bollobás-Riordan graphs. Notably, the latter graphs are *sparser* than the sparsest connected Erdős-Rényi graphs! In fact, we observed that the sparsity ratio of Bollobás-Riordan graphs is $1/N$, whereas for sparse connected Erdős-Rényi graphs we have a sparsity ratio given by the connection probability $\approx (\log N)/N$. However, despite such increased sparsity, the maximal degree of Bollobás-Riordan graphs grows as $\sqrt{N}$, namely, faster than the logarithmic law characterizing sparse Erdős-Rényi graphs. This difference must be ascribed to the fact that Bollobás-Riordan graphs are highly *inhomogeneous* and, hence, even with a small number of overall connections, there are nodes with a very large number of neighbors, inducing a faster growth of the *maximal* degree.

Chapter 7

Simulations and Experiments

7.1 Synthetic Data

According to Theorem 2, and in view of Lemmas 9 and 11, we have that the sample Granger estimator in (3.26) achieves universal local structural consistency according to Definition 2. As discussed in Chapter 3, this condition implies that the entries of the sample matrix estimator exhibit the dichotomy illustrated in Figure 3.1 (bottom panel). Therefore, we start by reproducing this behavior on synthetic data.

The two panels in Figure 7.1 display the pattern exhibited by the entries of the sample Granger estimator, $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$, for two realizations of the random graph $\mathcal{G}(N)$ with a probed subset $\mathcal{P}$ containing half the nodes of the entire network. For both realizations, the entries of $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ are scaled by $\sqrt{N}$, and, for clarity of visualization, they are vectorized and rearranged so that the entries corresponding to disconnected nodes come first. The vertical arrow displays the gap $\boldsymbol{\gamma}$, which was estimated using (5.11), (5.30) and (5.58), with reference to the pertinent graph topology shown in the figure. The following notable effects are observed. First, in perfect accordance with Definition 2, we see the emergence of an *identifiability gap* that separates clearly the entries corresponding to disconnected node pairs from the entries corresponding to connected node pairs. We also recall that in this case the bias is zero. Second, *clustering* is definitely visible: the entries pertaining to disconnected nodes cluster around zero, whereas the entries corresponding to connected nodes around $\boldsymbol{\gamma}$, the limiting value displayed by the vertical arrow. Last but not least, by comparing side-by-side the panels in Figure 7.1, we see that the two different realizations correspond to different values of the gap $\boldsymbol{\gamma}$, which confirms that this gap is in fact random.

We see from (5.58) that the limiting random variable $\boldsymbol{\mu}$ is one fundamental ingredient of the identifiability gap. It is therefore useful to examine the statistical distribution of $\boldsymbol{\mu}$, and in particular its behavior in comparison to the finite-size (scaled) maximal degree $\boldsymbol{\mu}_{\mathcal{G}}(N)/\sqrt{N}$ — see (5.31). To this end, in Figure 7.2 we display: *i*) the empirical histograms of the scaled maximal degree, for three values of N (first three panels from the

left); and *ii*) the empirical histogram corresponding to the limiting variable $\boldsymbol{\mu}$ (rightmost panel). To obtain the latter histogram, we exploit the following result.

Theorem 17 in [8]. *Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ be the points of a Poisson process with rate η, i.e., equal to the number of new edges added at each iteration of the Bollobás-Riordan procedure (see Section 5.1). The limiting random variable in (5.40) is equal to:*

$$\boldsymbol{\mu} = \max_{n=1,2,\ldots} \boldsymbol{z}_n - \boldsymbol{z}_{n-1}. \tag{7.1}$$

where, for $n = 0, 1, \ldots,$ we defined:

$$\boldsymbol{z}_0 \triangleq 0, \qquad \boldsymbol{z}_n \triangleq 2\eta\sqrt{\boldsymbol{p}_{\eta n}}. \tag{7.2}$$

□

According to this theorem, we simulate a Poisson process with rate η and use the expression of $\boldsymbol{\mu}$ provided in (7.1) and (7.2). By comparing the different panels in Figure 7.2, we see that the distribution of the scaled maximal degree approaches the distribution of $\boldsymbol{\mu}$ as N increases, and that the result is stable yet for the values $N = 100$ and $N = 250$. These are interesting values since, in the range $[100, 250]$, the probability of correct graph learning is close to 1, as we can appreciate from the quantitative performance analysis reported in Figure 7.3.

More specifically, in Figure 7.3 we show the probability of correct graph learning evaluated empirically over 10^3 Monte Carlo runs, as a function of the network size N. Specifically, the dynamical evolution in (1.1) is simulated over a network of increasing size N ranging from 50 to 250, and we consider a subset of probed nodes having cardinality $\lfloor \xi N \rfloor$, with $\xi = 0.15$. The curves displayed with continuous line refer to the *limiting* Granger estimator in (3.24), which is obtained by using the *true* covariance matrices. Markers refer to the regularized Granger estimator in (3.28), which is instead computed over the samples. The take-away messages from Figure 7.3 are that: *i*) for sufficiently large number of samples, the learning curve of the empirical Granger estimator reaches the curve of the limiting Granger estimator; and *ii*) consistent learning is progressively achieved as N grows.

In Figure 7.3, a relatively large number of samples is considered, and kept constant across all values of the network size N. Another useful analysis pertains to the effective number of samples necessary to achieve a target learning probability. In Figure 7.4 we evaluate empirically the number of samples needed to get a probability equal to 90% of the probability of correct learning achieved by the limiting Granger estimator. The blue curve corresponds to Bollobás-Riordan graphs, and shows a growth that matches well the almost-linear growth prescribed by Theorem 8.

It is useful to compare the observed behavior against the behavior of Erdős-Rényi graphs. The sample complexity laws relative to Erdős-Rényi graphs mentioned in the previous section are confirmed by the curves in Figure 7.4, revealing in particular that: *i*) the intermediate growth rate is given by Bollobás-Riordan graphs (blue curve), with *almost-linear* sample complexity; *ii*) the highest sample complexity is *quadratic*, and is

required by *dense* Erdős-Rényi graphs (green curve); and *iii*) the lowest sample complexity is achieved by *sparse* Erdős-Rényi graphs (red curve), and depends on the specific law chosen for the vanishing connection probability p.

The bottom line is that: *i*) *sparsity* of preferential attachment graphs makes them easier to learn than *dense* graphs; whereas *ii*) *heterogeneity* of preferential attachment graphs implies a power-law behavior that reflects into a $\sqrt{N}$-growth of the maximal degree, making them harder to learn than *sparse homogeneous* graphs.

Finally, we provide some quantitative data as regards the computational complexity of the graph learning strategy in the considered examples. To this end, we now report the run times relative to the XPS 7390 laptop of Dell Inc.®, equipped with an i7 Intel® processor and a 16GB RAM. The graph learning algorithm can be decoupled in two steps: *i*) computing the Granger estimator; and *ii*) performing the clustering algorithm on its entries. The cost associated to the clustering algorithm is negligible. In the first step, if we use the regularized Granger estimator, we need an optimization algorithm to solve numerically (3.28). In our simulations, we employed the MATLAB® package CVX [46, 47], which exhibited a run time ranging from $\approx 3\,\mathrm{s}$ to $\approx 8\,\mathrm{s}$ when N ranges from 100 to 250, with $\xi = 0.15$. The run time reduces to less than 1 ms if we use instead the non-regularized Granger estimator[1] in (3.26), which in the considered examples was found to coincide with its regularized counterpart — see the discussion following (3.28).

7.2 Real Networks and Directed Graphs

So far, we tested our results over *synthetic* network topologies generated according to the Bollobás-Riordan procedure described in Section 5.1. Since the main motivation behind the challenging study of these graphs is their similarity to real-world graphs, in this section we examine some topologies of existing networks. The examples that we are going to illustrate should be intended in the following way. We are given the topology of a real-world network, such as, e.g., a power-grid network, a network of routers, or a social network, which can support the implementation of distributed learning algorithms for different useful purposes. We therefore use the assigned network topology to build/run on top of it a distributed algorithm, for example, an adaptive distributed detection algorithm, or a social learning algorithm, which are examples matching well the considered model — see Section V-A in [78], and [11, 108]. Then, the focus of topology inference is to solve the reverse learning problem of retrieving the network graph from partial observation of the nodes' output.

We are now ready to illustrate the tests conducted over two real-world networks provided by a popular web-repository [97], corresponding to the topologies shown in Figure 7.5. As a preliminary comment, we can see that these topologies exhibit a dichotomous structure with "hubs" featuring many connections as opposed to "peripheral" nodes with few connections. Similarly shaped structures match well the heterogeneity

[1] We remark that no theoretical proof on the sample complexity of the non-regularized Granger estimator is available.

guaranteed by Bollobás-Riordan models, while they are impossible to mimic through the independent/homogeneous Erdős-Rényi generation.

The example in the top panel of Figure 7.5 refers to a power-grid network composed of $N = 4941$ nodes, connected according to the displayed topology. Over this topology, we let the autoregressive system (1.1) run with a Laplacian combination policy, and then applied our inference algorithm under the case that only one third of the nodes are probed, with $T = 2 \cdot 10^5$ samples. The results of the test are shown in the top plot in Figure 7.5. We see that the clustering algorithm, when applied over the entries of the regularized Granger estimator in (3.28), is able to separate correctly the disconnected/connected nodes, therefore providing faithful graph learning.

The second example, illustrated in the bottom panel of Figure 7.5, refers to a network of 100 routers connected according to the shown topology, which was extracted from a bigger network reported in the web-repository [97]. In this case, 50% of the nodes are probed, and we have $T = 10^6$ available samples. The results of the test are shown in the bottom plot in Figure. 7.5, where we can appreciate that the graph learning algorithm successfully classifies the node connections within the probed subnetwork. In comparison to the top panel, we see that in the bottom panel the spread of the sample estimators is reduced, which is a consequence of the fact that we have a larger number of samples and a smaller network size.

It is also useful to test whether the Granger estimator can achieve faithful graph learning over *directed* graphs. While Bollobás-Riordan graphs are naturally undirected, there are of course several straightforward ways to devise directed variants thereof, see, e.g., [8]. Perhaps the simplest way is to perform, at each step of the preferential attachment construction: *i*) the insertion of a *directed* edge *from the new node n to an existing node*, based on an attachment probability ruled by the *in-degree* of the existing nodes; and *ii*) the insertion of a *directed* edge *from an existing node to the new node n*, based on an attachment probability ruled by the *out-degree* of the existing nodes. Such construction is used in Figure 7.6, where we report two realizations relative to the parameters described in the caption. Regarding the Laplacian matrix, in the directed case we use definition (4.15) with the maximal *in*-degree. We see that, even in this non-symmetric case, the regularized Granger estimator is still able to separate well connected from disconnected pairs.

7.3 Dynamic Graphs

In this section we consider the *dynamic* graph setting, where the underlying graph is allowed to grow over time, while the node signals needed to perform topology inference are concurrently collected. We denote the graph at time t by $\mathcal{G}_t$, and its size by N_t. When the graph size increases, passing from $n-1$ nodes to n nodes, the topology of the previous subgraph relative to the nodes $\{1, 2, \ldots, n-1\}$ remains unaltered. In other words, the graph is dynamic in the sense that during time new nodes are attached to the previous structure by some new edges, but: *i)* if an edge was added between two nodes it will never

disappear; *ii)* if an edge was not added between two nodes it will never be added later

The incremental construction of such graphs can be done using both the Bollobás-Riordan procedure and the Erdős-Rényi model (with fixed connection probability $p_N = p$) in a natural way. In fact, in the former case we have just to follow the construction rules as in Figure 5.1, by adding new nodes at each time instant when N_t increases by 1. In the latter case, when the n-th node is added, we need to draw $n-1$ i.i.d. Bernoulli experiments with success probability p, in order to determine which new edge must be added between the new node and the previous $n-1$ nodes.

Moreover, we assume that the probed subset $\mathcal{P}$ is fixed (and such that $|\mathcal{P}| = N_0$); accordingly, it is the graph involving the latent nodes, including connections between latent and probed nodes, that grows over time.

The dynamic graph setting is illustrated in Figure 7.7, where we also show the difference with the static setting. In the static setting (bottom diagram), a fixed graph underlies the diffusion process for the entire observation interval during which topology inference is performed. In contrast, in the dynamic setting (top diagram) the graph grows incrementally over time with the probed set kept fixed.

While our technical analysis relies on the static case, in this section we present some preliminary experiments showing that graph learning in the dynamic setting is still possible, and that some new features arise, especially in terms of sample complexity.

Under the aforementioned dynamic graph setting, we need to modify (1.2) into:[2]

$$\boldsymbol{y}_t = \boldsymbol{A}_t \boldsymbol{y}_{t-1} + \boldsymbol{x}_t, \tag{7.3}$$

where the combination matrix $\boldsymbol{A}_t$ is obtained by using (4.15) over the dynamic graph $\mathcal{G}_t$. Moreover, motivated by the fact that for any t we have [63]:

$$\boldsymbol{R}_1(t) = \boldsymbol{A}_t \boldsymbol{R}_0(t-1) \quad \Longrightarrow \quad \boldsymbol{A}_t = \boldsymbol{R}_1(t)[\boldsymbol{R}_0(t-1)]^{-1}, \tag{7.4}$$

where:

$$\boldsymbol{R}_0(t-1) \triangleq \mathbb{E}\left[\boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}^{\top} \middle| \boldsymbol{A}_t\right], \qquad \boldsymbol{R}_1(t) \triangleq \mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^{\top} \middle| \boldsymbol{A}_t\right], \tag{7.5}$$

by following the same reasoning as in Section 3.5, we consider the following estimator:

$$\widehat{\boldsymbol{A}}_{\mathcal{P}}(t) = [\boldsymbol{R}_1(t)]_{\mathcal{P}}[\boldsymbol{R}_0(t-1)]_{\mathcal{P}}^{-1}, \tag{7.6}$$

which provides the best linear prediction of the future samples from the past one-lag samples collected *over the probed subset*. Actually, after having collected a certain number T of samples:

$$\left\{ [\boldsymbol{y}_t]_{\mathcal{P}} \,:\, t \in [1,T] \right\}, \tag{7.7}$$

we approximate (7.6) using the available dataset (7.7), namely, by substituting $[\boldsymbol{R}_j(T)]_{\mathcal{P}}$, for $j \in \{0,1\}$, with:

$$\widehat{\boldsymbol{R}}_{j,\mathcal{P}}(T) \triangleq \left[\frac{1}{T-j}\sum_{t=1+j}^{T} \boldsymbol{y}_t\boldsymbol{y}_{t-j}^{\top}\right]_{\mathcal{P}} = \frac{1}{T-j}\sum_{t=1+j}^{T} [\boldsymbol{y}_t]_{\mathcal{P}}[\boldsymbol{y}_t]_{\mathcal{P}}^{\top}, \tag{7.8}$$

[2]Note that, since now the network size is a function of time, we can drop the functional dependence on N_t and simplify the notation, for example, using $\boldsymbol{y}_t$ instead of $\boldsymbol{y}_t(N_t)$.

yielding the sample Granger estimator:

$$\widehat{\boldsymbol{R}}_{1,\mathcal{P}}(T)\big(\widehat{\boldsymbol{R}}_{0,\mathcal{P}}(T)\big)^{-1}. \tag{7.9}$$

We consider graph growths of the form:

$$N_t = N_0 + \alpha t^{\beta}, \tag{7.10}$$

where N_0 is the initial graph size. Note that there is a connection between the graph growth N_t and the sample complexity. Indeed, saying that the network size at time t is $N_t \sim t^{\beta}$ means that the number of samples employed to estimate a graph of size N is $T_N \sim N^{1/\beta}$. Therefore, the following remarkable coupling between sample complexity and graph growth emerges in the dynamic setting. If a certain *minimum* sample complexity is necessary to learn faithfully, this means that a *maximum* graph growth is permitted. In other words, *sample complexity places a limit on the maximum allowable velocity at which the dynamic graph can grow over time*. Building on the results available from Theorems 6 and 8, we would expect that, in order to successfully learn the graph topology in the dynamic case, the graph growth should be slower than:

$$N_t \sim t^{1/2} \qquad \text{[Erdős-Rényi graphs]}, \tag{7.11}$$
$$N_t \sim t \qquad \text{[Bollobás-Riordan graphs]}. \tag{7.12}$$

However, our experiments show that successful results can be obtained also with faster growth rates, and in particular by considering:

$$N_t \sim t^{4/5} \qquad \text{[Erdős-Rényi graphs]}, \tag{7.13}$$
$$N_t \sim t^{3/2} \qquad \text{[Bollobás-Riordan graphs]}, \tag{7.14}$$

we get the performance shown in Figure 7.8.

The exponents 4/5 and 3/2 are based on numerical experiments, as there are currently no counterparts of Theorems 6 and 8 available for the dynamic case. As was mentioned, in (7.13) and (7.14) we increase the velocity at which the graph grows, for both the Erdős-Rényi and the Bollobás-Riordan models, which corresponds to reducing the number of samples. In particular, by inverting the relations in (7.13) and (7.14), the new sample scaling laws are:

$$T_N \sim N^{5/4} < N^2 \qquad \text{[Erdős-Rényi graphs]}, \tag{7.15}$$
$$T_N \sim N^{2/3} < N \qquad \text{[Bollobás-Riordan graphs]}. \tag{7.16}$$

Remarkably, the plots in Figure 7.8 reveal that, despite the increased velocity (i.e., the reduced number of samples), the graph learning problem remains feasible. In contrast, in the static case (where the graph has constant size $N = 200$, which is the size corresponding to the end of the observation window for the dynamic case), the performance is not good since we are violating the prescriptions of Theorems 6 and 8.

We have also tested the directed counterparts of Erdős-Rényi graphs and of Bollobás-Riordan graphs. In particular, for directed Erdős-Rényi graphs, a.k.a. binomial graphs, directed edges corresponding to pairs (k, ℓ) and (ℓ, k) are drawn independently, while for directed Bollobás-Riordan graphs we consider preferential-attachment probabilities based on in-degrees and out-degrees to build directed edges as described in Section 7.2. We obtained results similar to those shown in Figure 7.8.

One conclusion arising from these results is that, under partial observability, application of the Granger estimator over dynamic graphs can deliver superior performance as compared to the static case. This is a remarkable and perhaps unexpected behavior. It is possible to provide an interpretation of this behavior based on Theorems 6 and 8. Even though these theorems characterize only the static case, their proofs reveal that the main factor determining the sample complexity is the magnitude of the nonzero entries in the combination matrix: the smaller they are, the higher the sample complexity will be. On the other hand, the nonzero entries are inversely proportional to the maximum degree of the graph, which increases with the network size, leading to an increase in sample complexity. Under a static model, the system works during the entire observation interval with the largest graph. In contrast, under the dynamic model the system works with growing graphs and, hence, on average the network size is smaller (i.e., more favorable) than the size considered in the static case (see Figure 7.7). This is one reason why the dynamic case looks less demanding in terms of samples, ultimately implying that a faster growth is permitted for the sequence of dynamic graphs.

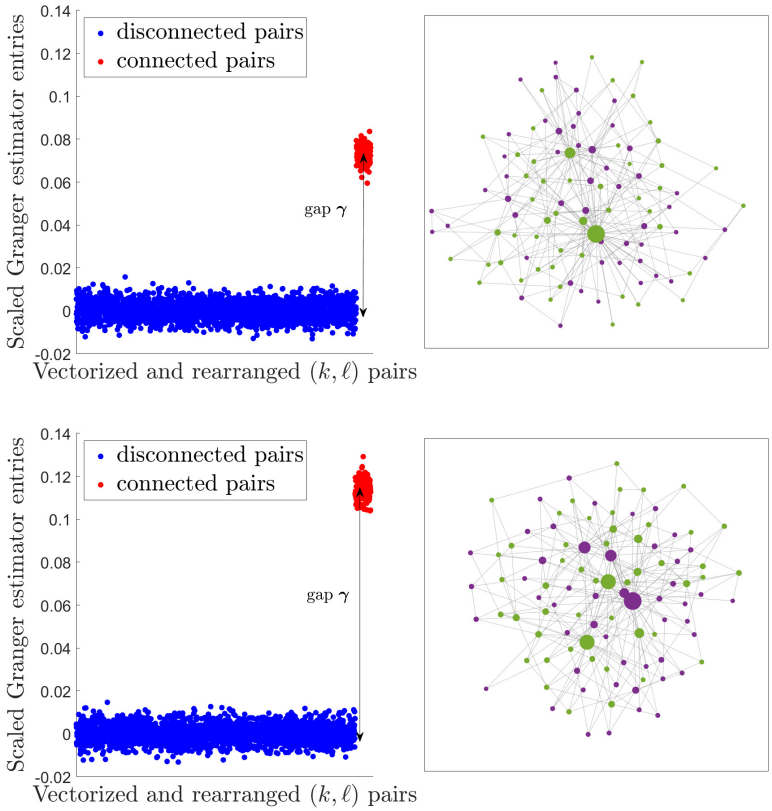

Figure 7.1: Illustration of universal local structural consistency of the sample Granger estimator in (3.26) for two realizations of a Bollobás-Riordan graph with number of nodes $N = 100$ and parameter $\eta = 3$. The plots show the entries of the sample Granger estimator, scaled by $\sqrt{N}$, vectorized and rearranged so that entries corresponding to disconnected nodes come first. The vertical arrow displays the gap $\boldsymbol{\gamma}$. In the shown network topologies, probed nodes are displayed in green, while latent nodes in purple, with the circle radius being proportional to the node degree. The probed subset has cardinality $N/2 = 50$, and its nodes are randomly picked from $\{1, 2, \ldots, N\}$ without replacement. The parameters of the Laplacian matrix are $\rho = 0.5$ and $\lambda = 0.75$. In practice, the connections of the graph can be estimated by applying a clustering procedure over the entries of the sample Granger estimator shown in the plot, which accurately reproduce the dichotomous pattern of the true (scaled) combination matrix revealed by Lemma 9. According to the analysis presented in Chapter 3, the possibility of retrieving the graph by means of a clustering procedure is a consequence of the fact that the sample Granger estimator in (3.26) achieves universal local structural consistency — see Definition 2.

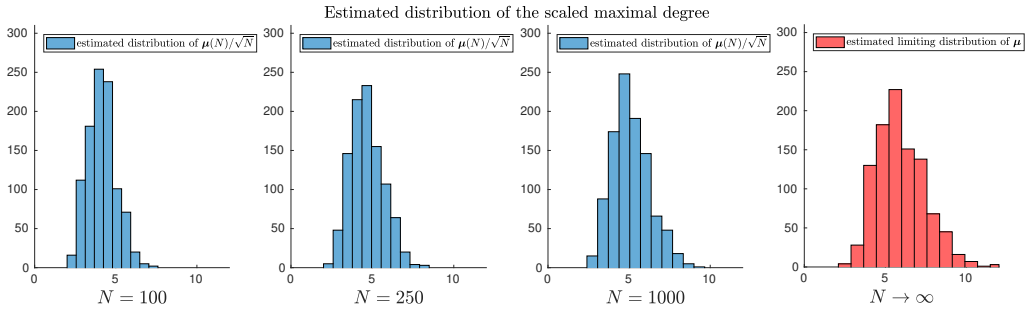

Figure 7.2: *First three panels.* Empirical histograms, obtained over 10^3 Monte Carlo runs, relative to the scaled maximal degree $\boldsymbol{\mu}_{\mathcal{G}}(N)/\sqrt{N}$. *Rightmost panel.* Empirical histogram relative to the limiting random variable $\boldsymbol{\mu}$, obtained by simulating, over 10^3 Monte Carlo runs, a Poisson process of rate η, and by exploiting relations (7.1) and (7.2).

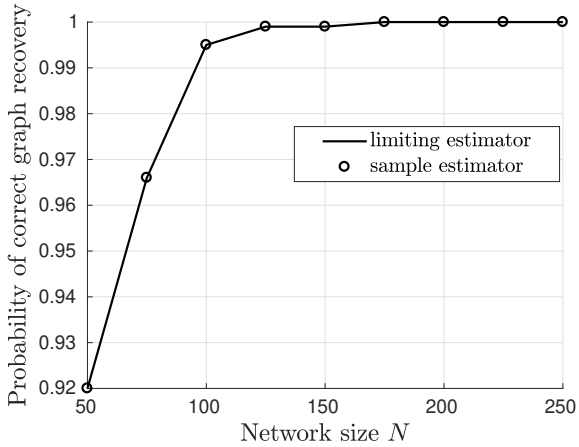

Figure 7.3: Probability of correct graph recovery for different values of the network size N. For each network size, the probability of correct graph recovery is estimated empirically over 10^3 Monte Carlo runs. We remark that *correct* graph recovery here means that graphs with even a single wrong edge are counted as an erroneous experiment. The graphs are generated according to a Bollobás-Riordan model with parameter $\eta = 3$. The sequence of probed subsets fulfills (5.63) with $\xi = 0.15$. We consider: the *limiting* estimator (3.24) obtained by using the true covariances (solid line); and the *empirical* estimator (3.28) obtained by using the sample covariances (markers) evaluated over $T = 3 \cdot 10^6$ samples. The clustering algorithm applied to the Granger estimator is the modified k-means algorithm proposed in [75]. The parameters of the Laplacian matrix are $\rho = 0.5$ and $\lambda = 0.75$.

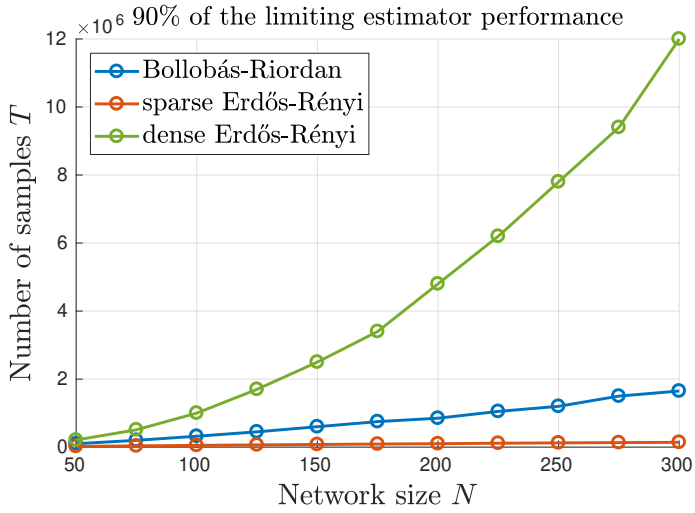

Figure 7.4: Sample complexity of preferential attachment graphs, compared against sparse and dense Erdős-Rényi graphs. The curves depict the number of samples needed by the *empirical* estimator to attain 90% of the performance (i.e., probability of correct graph recovery) of the *limiting* estimator, for different values of N. The preferential attachment graphs are generated as Bollobás-Riordan graphs with parameter $\eta = 3$. The sparse Erdős-Rényi graphs are generated with a connection probability $p = \frac{\log N}{N} \cdot \operatorname{loglog} N$, whereas for the dense Erdős-Rényi graphs we have $p = 0.5$. The underlying probability of correct graph recovery is evaluated over 10^3 Monte Carlo runs. The parameters of the Laplacian matrix are $\rho = 0.5$ and $\lambda = 0.75$.

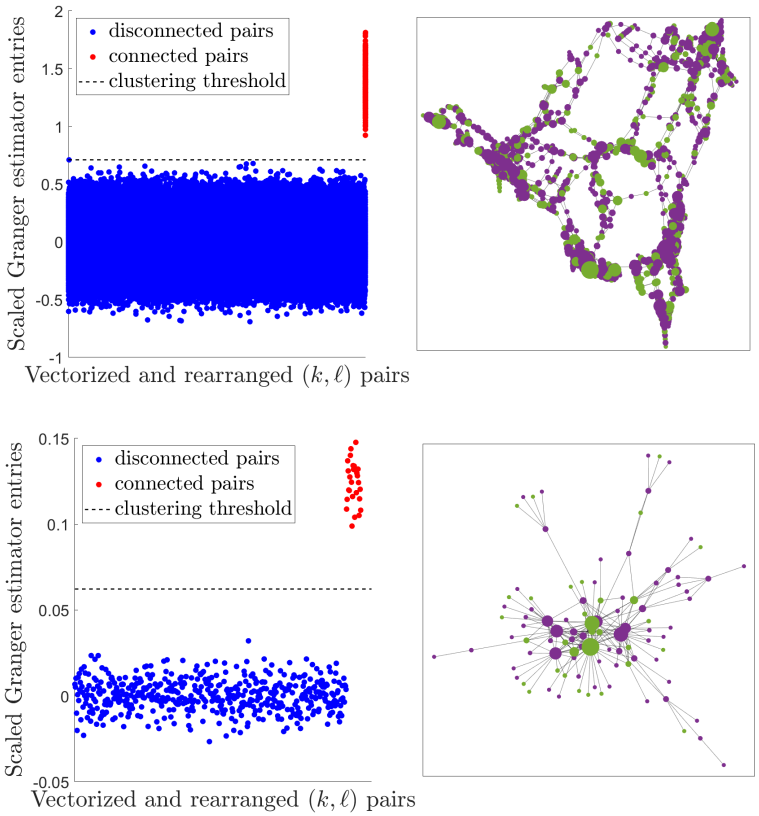

Figure 7.5: Experiments over real-world topologies. The parameters of the Laplacian matrix are $\rho = 0.5$ and $\lambda = 0.75$. *Top.* A simulation run of (1.1) over a power-grid network of $N = 4941$ nodes taken from the web-repository [97]. The plot shows the entries of the regularized Granger estimator in (3.28), scaled by $\sqrt{N}$, vectorized and rearranged so that entries corresponding to disconnected nodes come first. In this run we set $T = 2 \cdot 10^5$, and consider a subset of probed nodes having cardinality $N/3 = 1647$. In the network topology, probed nodes are displayed in green, while latent nodes in purple, with the circle radius being proportional to the node degree. *Bottom.* The same general setting as in the left panel, with reference to a network of $N = 100$ real-world routers, whose connection topology was extracted from a bigger network available in the web-repository [97]. In this run we set $T = 10^6$, and consider a subset of probed nodes having cardinality $N/2 = 50$. In both cases, we run the clustering algorithm in Listing 1 over the entries of the matrix estimator using the procedure defined in Listing 2. We display the resulting clustering threshold (dashed line), which represents the midpoint of the centroids of the two clusters constructed by the algorithm. We see that the threshold correctly separates the entries relative to the connected node pairs from the entries relative to unconnected node pairs, leading to a correct estimation of the graph.

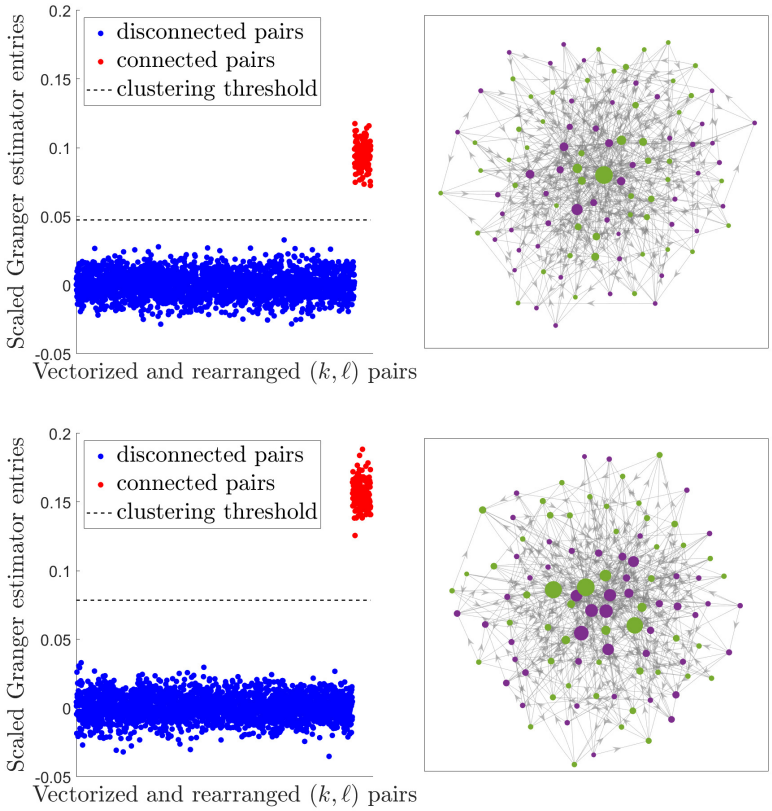

Figure 7.6: Two realizations of a *directed* Bollobás-Riordan graph, whose construction is detailed in the main text. We set $N = 100$ and $\eta = 3$. The plot shows the entries of the regularized Granger estimator in (3.28) computed over $T = 10^6$ samples, scaled by $\sqrt{N}$, vectorized and rearranged so that entries corresponding to disconnected nodes come first. In the shown network topologies, probed nodes are displayed in green, while latent nodes in purple, with the circle radius being proportional to the node *in-degree*. The probed subset has cardinality $N/2 = 50$, and its nodes are randomly picked from $\{1, 2, \ldots, N\}$ without replacement. The parameters of the Laplacian matrix are $\rho = 0.5$ and $\lambda = 0.75$. In both cases, we run the clustering algorithm in Listing 1 over the entries of the matrix estimator using the procedure defined in Listing 2. We display the resulting clustering threshold (dashed line), which represents the midpoint of the centroids of the two clusters constructed by the algorithm. We see that the threshold correctly separates the entries relative to the connected node pairs from the entries relative to unconnected node pairs, leading to a correct estimation of the graph.

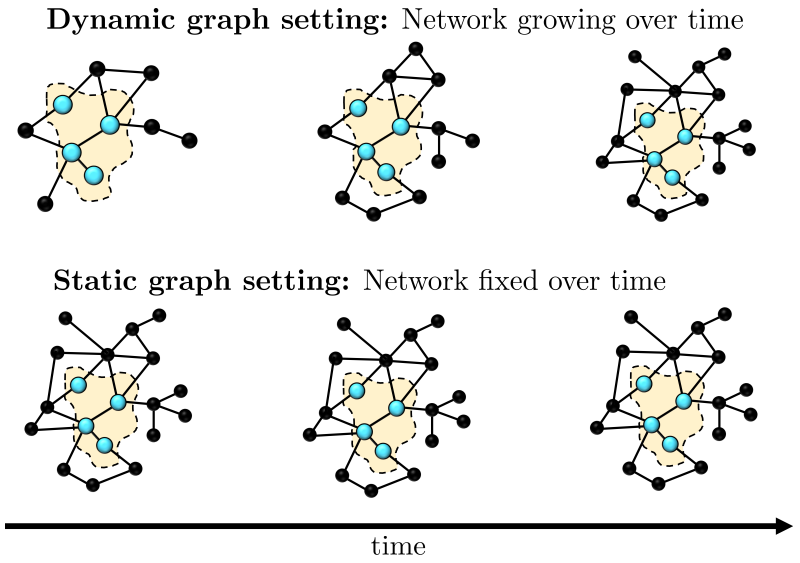

Figure 7.7: Dynamic graphs vs. static graphs. The unobserved nodes are shown in black, while the probed nodes in cyan. The probed set $\mathcal{P}$ is further highlighted by the yellow area.

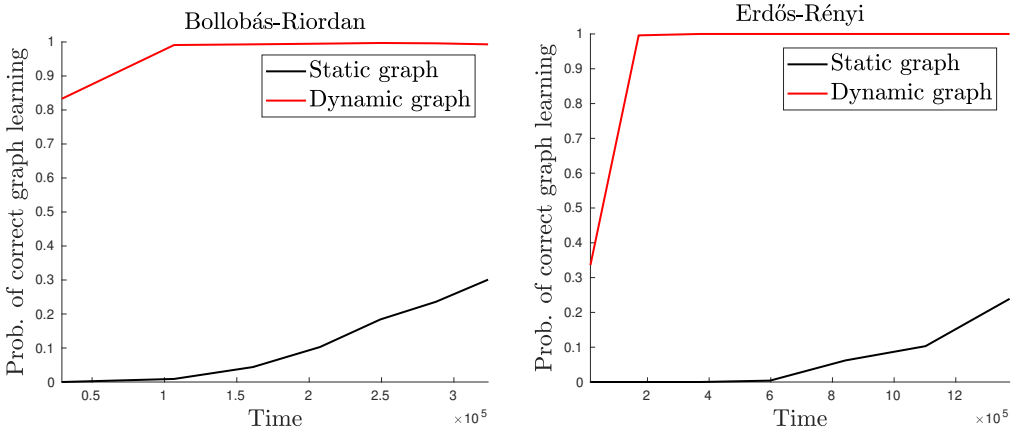

Figure 7.8: *Left plot.* Erdős-Rényi graphs with connection probability $p = 0$. In the dynamic case, the network size scales as $N_t \sim t^{4/5}$. *Right plot.* Bollobás-Riordan graphs with parameter $\eta = 3$. In the dynamic case, the network size scales as $N_t \sim t^{3/2}$. In both experiments we use the sample Granger estimator (7.9), followed by the modified k-means algorithm in Listing 2; the probed subset is $\mathcal{P} = \{1, 2, \ldots, 10\}$, and $N_0 = 15$; the parameters of the Laplacian combination matrix are $\lambda = 0.75$ and $\rho = 0.5$; and we use 10^3 Monte Carlo runs.

Conclusion

In this thesis, we have examined the problem of learning a network graph from the signals diffusing across the network according to the vector autoregressive model in (1.1). The distinguishing features of our work are: *i*) the network topology is modeled as a preferential attachment random graph; and *ii*) only part of the network is monitored (partial observability). We established that the Granger estimator under partial observability achieves faithful graph learning (with high probability as the network grows) for the class of Bollobás-Riordan graphs when the signals of neighboring nodes are combined according to the Laplacian matrix of the graph.

Previous results on consistent graph learning under partial observability, for diffusion models like (1.1), were relative to Erdős-Rényi graphs, which cannot reproduce faithfully the behavior of several useful real-world networks. In contrast, preferential attachment graphs were shown to be powerful in capturing useful real-world effects such as node heterogeneity and statistical dependence across graph edges. Accordingly, moving from Erdős-Rényi to Bollobás-Riordan graphs constitutes a useful research advance, which was rather demanding, especially because the multigraph construction relies on a preferential attachment mechanism, which introduces significant dependence across the edges, thus preventing from application of the simpler i.i.d. models adopted for the former graph models. Exploiting statistical concentration results for dependent processes, we are able to examine in detail the limiting properties of these graphs, and to ascertain that the entries of the Granger matrix estimator computed over the probed subnetwork split into two classes, separated by an *identifiability gap*. As a peculiar feature of the Bollobás-Riordan model, this gap is a *random* variable, which depends on the particular instance of the multigraph generation, as opposed to the deterministic gap observed for Erdős-Rényi graphs. We proved that the emergence of the gap is critical to enable *achievable graph learning,* and characterized the scaling law that relates how the number of samples must grow with the network size (Theorem 8), finding that the sample complexity is slightly larger than $N \log N$.

There are many other open questions that might deserve attention. For example, achievability is expressed in a worst-case perspective where *perfect* graph reconstruction is required. It would be useful to relax the criterion to encompass a possible fraction of misclassified edges, and see how this impacts the performance and the requested number of samples. Moreover, since classification is performed by using an unsupervised clustering

algorithm, it would be useful to see whether one could explore side-information to set a classification threshold, and see how this changes the fraction of misclassified nodes.

We established that the identifiability gap over Bollobás-Riordan graphs is *random*, as opposed to the deterministic nature that was proved over Erdős-Rényi graphs. This difference stimulates an open question that concerns the connections between the nature (deterministic or random) of the identifiability gap and the generative mechanism of the graph. For example, it would be interesting to consider other useful graph models, such as the stochastic block model, Chung-Lu graphs, random dot product or random geometric graphs. In particular, it would be interesting to ascertain whether randomness of the identifiability gap is related to the scale-free property or the preferential attachment mechanism.

The conducted analysis does not depend on a specific selection rule for the subset of probed nodes. Indeed, some of our results hold for any (deterministic) subset of probed nodes. For Theorem 8, we just need to impose that the cardinality of this subset scales linearly with N so as to avoid trivial (i.e., fully connected or fully disconnected) subgraphs as $N \to \infty$. Regarding the choice of the probed subset, it is interesting to consider an adversarial perspective where a malicious entity wants to select it to impair the topology inference algorithm. In the traditional setting, attacks to network graphs have the goal of impairing the connectivity properties the network. One classical attack is the *deletion* attack, where the attacker has the freedom of deleting some nodes [9]. The goal is to reduce the connectivity of the selected subgraph surviving after the deletion process. In particular, when the attacker leverages knowledge of the preferential attachment construction, he/she can delete all the "oldest" nodes to minimize connectivity. In our context, connectivity of probed nodes is not of interest, but one way to impair the clustering algorithm would be to select a subset that is fully connected or fully disconnected. However, when the cardinality of the probed subset grows linearly with N, we know that both these extreme situations occur with vanishing probability as N grows. The picture changes if the attack does not rely only on the graph model, and the adversary has the power of choosing the subset based on the actual realizations of the graph and/or of the nodes' output. In this case, the subset becomes statistically dependent on other random variables, and the statistical properties of the combination matrix and its estimated counterpart change due to the dependence introduced between the subset and the matrix entries. Carrying out the analysis under this scenario requires a different analysis that is not covered by the results of this work.

Devising matrix estimators different from (3.28), exploiting in particular some other structural constraints such as sparsity or smoothness can be useful to reduce sample complexity.

Another important aspect pertains to *online* graph learning algorithms [108, 114]. A preliminary analysis has been presented in Section 7.3, where we have implemented an online variation of the sample Granger estimator that learns from streaming data, encompassing the possibility that the graph topology changes due to the evolutionary mechanism of the Bollobás-Riordan graph. We have shown experimentally that this estimator continues to guarantee correct graph learning, and even with a smaller sample complexity

w.r.t. the case when the graph is static (i.e., does not grow over time). However, a formal analysis concerning these results is still missing.

Useful research advances concern the generalization to higher-order vector autoregressive processes, nonlinear models, and other classes of combination matrices. One particularly interesting extension regards the case where the combination matrix is asymmetric. This case as been considered in the experimental analysis (Section 7.2), where the asymmetry of the combination matrix arises from the fact that the underlying graph is directed. We have shown that the Granger estimator is still useful to learn the true graph, but a formal proof of this result is currently unavailable.

As a final point, we note that a graph estimator that works under the partial observability assumption could be useful even when all the nodes in the network can be probed. For instance, this happens in the situation where, over large networks, one can eventually probe all nodes, but not simultaneously, due to various types of constraints (i.e., computation and accessibility), it might be impractical to measure/collect all signals from the network at once. In this situation, even if the final goal is to reconstruct the whole graph, the only viable solution is to learn separately several patches of the network graph and merge the partial results coming from each patch to eventually estimate the entire network graph. This approach has been first proposed in [78]. An interesting open problem is the determination of a strategy to form and select the patches so as to optimize the efficiency of the learning algorithm.

Appendices

Appendix A

Proofs of Theorems 1, 2 and 3

Proof of Theorem 1: From universal local structural consistency, recalling (3.7), we know that for any $\varepsilon > 0$ and all $k \neq \ell$ we have with high probability as $N \to \infty$:

$$c_N\, \widehat{\boldsymbol{a}}_{k\ell}(T_N, N) \in \begin{cases} \Big[\boldsymbol{\beta} + (1-\varepsilon)\boldsymbol{\gamma},\ \boldsymbol{\beta} + (1+\varepsilon)\boldsymbol{\gamma}\Big], & (k,\ell) \text{ connected}, \\ \Big[\boldsymbol{\beta} - \varepsilon\boldsymbol{\gamma}, \boldsymbol{\beta} + \varepsilon\boldsymbol{\gamma}\Big], & (k,\ell) \text{ disconnected}. \end{cases} \tag{A.1}$$

By assumption, there exists a certain $\varepsilon > 0$ such that the algorithm graphclu$(\cdot)$ achieves successful classification for all configurations fulfilling (A.1), provided that $\mathcal{G}_{\mathbb{S}}(N)$ is neither fully connected nor fully disconnected. Thus, the proof is complete because we have just noted that configurations fulfilling (A.1) occur with high probability as $N \to \infty$, and by assumption the probability that $\mathcal{G}_{\mathbb{S}}(N)$ is fully connected or fully disconnected vanishes as $N \to \infty$. ■

Proof of Theorem 2: By application of the triangle inequality we can write:

$$\begin{aligned} &\|c_N \widehat{\boldsymbol{A}}_{\mathbb{S}}(T,N) - \boldsymbol{\gamma}\boldsymbol{G}_{\mathbb{S}}(N) - \boldsymbol{\beta}\|_{\text{max-off}} \\ &\quad \leq c_N \|\widehat{\boldsymbol{A}}_{\mathbb{S}}(T,N) - \widehat{\boldsymbol{A}}_{\mathbb{S}}(N)\|_{\text{max-off}} \\ &\quad + \|c_N(\widehat{\boldsymbol{A}}_{\mathbb{S}}(N) - \boldsymbol{A}_{\mathbb{S}}(N)) - \boldsymbol{\beta}\|_{\text{max-off}} \\ &\quad + \|c_N \boldsymbol{A}_{\mathbb{S}}(N) - \boldsymbol{\gamma}\boldsymbol{G}_{\mathbb{S}}(N)\|_{\text{max-off}}. \end{aligned} \tag{A.2}$$

Let us focus on the first term on the RHS of (A.2). From the definition of limiting estimator in (3.17), we have, for any $N \in \mathbb{N}$ and $\varepsilon > 0$:

$$\lim_{T\to\infty} \mathbb{P}\left[c_N \|\widehat{\boldsymbol{A}}_{\mathbb{S}}(T,N) - \widehat{\boldsymbol{A}}_{\mathbb{S}}(N)\|_{\text{max-off}} > \varepsilon\right] = 0. \tag{A.3}$$

By definition of limit, from (A.3) we conclude that, for any $N \in \mathbb{N}$ and any $\delta, \varepsilon > 0$, there exists always a value $T_0(N, \delta, \varepsilon)$ such that for all $T \geq T_0(N, \delta, \varepsilon)$:

$$\mathbb{P}\left[c_N \|\widehat{\boldsymbol{A}}_{\mathcal{S}}(T, N) - \widehat{\boldsymbol{A}}_{\mathcal{S}}(N)\|_{\text{max-off}} > \varepsilon\right] \leq \delta. \tag{A.4}$$

Let now f_N and g_N be two positive sequences vanishing with N with arbitrary laws. Since the reduction of δ and/or ε is a more demanding condition, the function $T_0(N, \delta, \varepsilon)$ can be always chosen to be non-increasing w.r.t. both δ and ε, which implies that for sufficiently large N:

$$T_0(N, f_N, g_N) \geq T_0(N, \delta, \varepsilon), \tag{A.5}$$

further implying, in view of (A.4):

$$\mathbb{P}\left[c_N \|\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_0(N, f_N, g_N), N) - \widehat{\boldsymbol{A}}_{\mathcal{S}}(N)\|_{\text{max-off}} > \varepsilon\right] \leq \delta. \tag{A.6}$$

In other words, if the number of samples scales with N as $T_N = T_0(N, f_N, g_N)$, we can write:

$$c_N \|\widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N) - \widehat{\boldsymbol{A}}_{\mathcal{S}}(N)\|_{\text{max-off}} \xrightarrow{\text{p}} 0. \tag{A.7}$$

Plugging this result into (A.2) and noticing that the second and third terms on the RHS of (A.2) vanish in probability in view of Definitions 5 and 1, respectively, we conclude that:

$$\|c_N \widehat{\boldsymbol{A}}_{\mathcal{S}}(T_N, N) - \boldsymbol{\gamma} \boldsymbol{G}_{\mathcal{S}}(N) - \boldsymbol{\beta}\|_{\text{max-off}} \xrightarrow{\text{p}} 0, \tag{A.8}$$

■

Proof of Theorem 3: Let

$$v_1 \leq v_2 \leq \ldots \leq v_L \tag{A.9}$$

be the vectorized and reordered set of the off-diagonal entries of the input matrix that feeds the graphclu$(\cdot)$ procedure. By assumption, we have that for these entries the property in (3.9) holds for some values ε, γ and β. In particular we are interested in the case $\varepsilon \leq 1/6$, for which we can surely write:

$$\beta + \varepsilon\gamma < \beta + (1 - \varepsilon)\gamma, \tag{A.10}$$

and therefore there exists an index $j^\star \in [1, L]$ such that, recalling the notation used in (3.33), the entries relative to connected pairs are in the set:

$$\mathcal{C}_0(j^\star) = \{v_1, v_2, \ldots, v_{j^\star}\}, \tag{A.11}$$

and satisfy:

$$\beta - \varepsilon\gamma \leq v_j \leq \beta + \varepsilon\gamma, \qquad \text{for } j = 1, 2, \ldots, j^\star, \tag{A.12}$$

while the entries relative to disconnected pairs are in the set:

$$\mathcal{C}_1(j^\star) = \{v_{j^\star+1}, v_{j^\star+2}, \ldots, v_L\}. \tag{A.13}$$

and satisfy:

$$\beta + (1-\varepsilon)\gamma \leq v_j \leq \beta + (1+\varepsilon)\gamma, \qquad \text{for } j = j^\star + 1, j^\star + 2, \ldots, L. \tag{A.14}$$

Of course, the clusters $\mathcal{C}_0(j^\star)$ and $\mathcal{C}_1(j^\star)$ are the correct solution to the graph learning problem. Thus, to prove the claim we have to show that the index $j^\star$ actually coincides with the solution provided by Listing 1. We recall that Listing 1 selects its output within the set $\mathcal{A}$ made of the eligible solutions of the k-means algorithm. Accordingly, we start by showing that $j^\star$ is added to the set $\mathcal{A}$, proving that the centroids midpoint fulfills the condition:

$$v_{j^\star} < \frac{c_0(j^\star) + c_1(j^\star)}{2} < v_{j^\star+1}, \tag{A.15}$$

where we recall that $c_0(j^\star)$ and $c_1(j^\star)$ are the centroids of clusters $\mathcal{C}_0(j^\star)$ and $\mathcal{C}_1(j^\star)$, respectively. From (A.12) and (A.14) we have:

$$\beta - \varepsilon\gamma \leq c_0(j^\star) \leq \beta + \varepsilon\gamma, \qquad \beta + (1-\varepsilon)\gamma \leq c_1(j^\star) \leq \beta + (1+\varepsilon)\gamma, \tag{A.16}$$

and, hence,

$$\beta + \left(\frac{1}{2} - \varepsilon\right)\gamma \leq \frac{c_0(j^\star) + c_1(j^\star)}{2} \leq \beta + \left(\frac{1}{2} + \varepsilon\right)\gamma. \tag{A.17}$$

Since $\varepsilon \leq 1/6$ (and in particular $\varepsilon < 1/4$) we have:

$$\frac{1}{2} - \varepsilon > \varepsilon, \qquad \text{and} \qquad \frac{1}{2} + \varepsilon < 1 - \varepsilon, \tag{A.18}$$

which, used into (A.17), yield:

$$\beta + \varepsilon\gamma < \frac{c_0(j^\star) + c_1(j^\star)}{2} < \beta + (1-\varepsilon)\gamma, \tag{A.19}$$

and therefore (A.15) is proved by recalling (A.12) and (A.14).

In principle, other eligible solutions could exist, and it remains to prove that if another value $j \in \mathcal{A} \backslash \{j^\star\}$ exists, then this configuration must necessarily exhibit a smaller distance between the clusters, namely,

$$c_1(j) - c_0(j) < c_1(j^\star) - c_0(j^\star) \qquad \forall j \in \mathcal{A} \setminus \{j^\star\}. \tag{A.20}$$

First, we note that, in view of (A.16), the distance between the centroids relative to $j^\star$ has lower bound:

$$c_1(j^\star) - c_0(j^\star) \geq (1 - 2\varepsilon)\gamma. \tag{A.21}$$

Let now $j \in \mathcal{A}$ with $j > j^\star$. By definition we have:

$$\beta + (1-\varepsilon)\gamma \leq v_j < \frac{c_0(j) + c_1(j)}{2}, \tag{A.22}$$

where the upper bound is due to the fact that $j \in \mathcal{A}$, whereas the lower bound is due to the assumption $j > j^\star$, which implies $v_j \in \mathcal{C}_1(j^\star)$. Multiplying by -2 the leftmost and rightmost terms in (A.22) and adding $2c_1(j)$, we obtain:

$$c_1(j) - c_0(j) < 2c_1(j) - 2\beta - 2(1-\varepsilon)\gamma, \tag{A.23}$$

and since we certainly have $c_1(j) \leq v_L \leq \beta + (1+\varepsilon)\gamma$, we finally get:

$$c_1(j) - c_0(j) < 4\varepsilon\gamma. \tag{A.24}$$

Since the condition $\varepsilon \leq 1/6$ is equivalent to:

$$4\varepsilon \leq 1 - 2\varepsilon, \tag{A.25}$$

by comparing (A.21) and (A.24) we obtain (A.20) for each $j > j^\star$. Since similar arguments can be used for the remaining case $j < j^\star$, the proof is complete. ■

Appendix B

Deterministic Properties of the Limiting Granger Estimator

In this section we obtain an upper bound on the error of the limiting Granger estimator. To this aim, we start by proving two auxiliary lemmas that hold for any $N \times N$ scaled left-stochastic matrix $A = [a_{k\ell}]$, namely, for any matrix whose entries satisfy the conditions:

$$a_{k\ell} \geq 0, \qquad \sum_{\ell=1}^{N} a_{k\ell} = \rho. \tag{B.1}$$

In the following analysis we denote by $a_{k\ell}^{(i)}$ the (k,ℓ)-entry of the matrix power A^i, and we use the following quantity:

$$\mathfrak{M}(A) \triangleq \max_{\substack{k,\ell \in [1,N] \\ k \neq \ell}} \sum_{\substack{j=1 \\ j \neq k,\ell}}^{N} a_{kj} a_{j\ell}. \tag{B.2}$$

Lemma 14 (Bounds on Matrix Powers). *Let A be an $N \times N$ scaled left-stochastic matrix as in (B.1). For $i = 1, 2, \ldots,$ we have that:*

- *The main diagonal entries of A^{2i} satisfy the inequalities:*

$$a_{kk}^{(2i)} \leq \alpha_i, \tag{B.3}$$

  where the sequence α_i is recursively defined as:

$$\alpha_1 = \rho^2, \qquad \alpha_{i+1} = \rho^2 \alpha_i + \rho^{2(i+1)}. \tag{B.4}$$

- *The off-diagonal entries of A^{2i} satisfy the inequalities:*

$$a_{k\ell}^{(2i)} \leq \beta_i \, a_{k\ell} + \gamma_i \tag{B.5}$$

where β_i and γ_i are two sequences recursively defined as:

$$\beta_1 = 2\rho, \qquad \beta_{i+1} = 2\rho\,\alpha_i + \rho^2\,\beta_i, \tag{B.6}$$

$$\gamma_1 = \mathfrak{M}(A), \qquad \gamma_{i+1} = \mathfrak{M}(A)\alpha_i + 3\rho\mathfrak{M}(A)\beta_i + \rho^2\,\gamma_i. \tag{B.7}$$

Proof: Preliminarily, it is useful to observe that:[1]

$$\sum_{\ell=1}^{N} a_{k\ell}^{(i)} = \rho^i. \tag{B.9}$$

We start by proving (B.3) by induction. For $i = 1$, the claim follows directly from (B.9). We shall therefore prove that (B.3) holds for $i+1$, assuming that it holds for i. To this aim, let us write the diagonal terms of matrix $A^{2(i+1)}$ as:

$$a_{kk}^{(2i+2)} = \sum_{\ell=1}^{N} a_{k\ell}^{(2i)} a_{\ell k}^{(2)} = a_{kk}^{(2i)} a_{kk}^{(2)} + \sum_{\substack{\ell=1 \\ \ell\neq k}}^{N} a_{k\ell}^{(2i)} a_{\ell k}^{(2)}. \tag{B.10}$$

We observe that (B.9) implies in particular the following inequalities:

$$a_{kk}^{(2)} \leq \rho^2, \qquad a_{\ell k}^{(2)} \leq \rho^2, \qquad \sum_{\substack{\ell=1 \\ \ell\neq k}}^{N} a_{k\ell}^{(2i)} \leq \rho^{2i}, \tag{B.11}$$

which, applied in (B.10), yield:

$$a_{kk}^{(2i+2)} \;\leq\; \rho^2\, a_{kk}^{(2i)} + \rho^{2(i+1)}. \tag{B.12}$$

Since $a_{kk}^{(2i)} \leq \alpha_i$ by the induction hypothesis, from (B.12) we get:

$$a_{kk}^{(2i+2)} \leq \rho^2\,\alpha_i + \rho^{2(i+1)} = \alpha_{i+1}, \tag{B.13}$$

which corresponds to (B.3) for the case $i+1$, and the claim for the diagonal entries is proved.

We continue by proving (B.5) by induction. For any $k, \ell = 1, 2, \ldots, N$, with $k \neq \ell$, we have:

$$\begin{aligned} a_{k\ell}^{(2)} = \sum_{h=1}^{N} a_{kh} a_{h\ell} &= (a_{kk} + a_{\ell\ell})\, a_{k\ell} + \sum_{\substack{h=1 \\ h\neq k,\ell}}^{N} a_{kh} a_{h\ell} \\ &\leq 2\rho\, a_{k\ell} + \mathfrak{M}(A) \triangleq \beta_1\, a_{k\ell} + \gamma_1, \end{aligned} \tag{B.14}$$

[1] We can prove this property by induction as follows. For $i = 1$ the property is exactly (B.1), while the induction step comes from:

$$\sum_{\ell=1}^{N} a_{k\ell}^{(i+1)} = \sum_{\ell=1}^{N} \sum_{h=1}^{N} a_{kh}^{(i)} a_{h\ell} = \sum_{h=1}^{N} a_{kh}^{(i)} \sum_{\ell=1}^{N} a_{h\ell} = \rho^i \rho. \tag{B.8}$$

where: i) in the inequality we exploited the fact that the diagonal entries of A are upper bounded by ρ in view of (B.1), and we used the definition of $\mathfrak{M}(A)$ in (B.2); and ii) in the last equality we applied the definitions of β_1 and γ_1 appearing in (B.6) and (B.7), respectively. We conclude from (B.14) that the claim in (B.5) holds for $i = 1$. Let us now show that, if the claim holds for a generic i, then it holds for $i + 1$. To this aim, we observe that:

$$\begin{aligned} a_{k\ell}^{(2i+2)} &= \sum_{h=1}^{N} a_{kh}^{(2)} a_{h\ell}^{(2i)} = a_{k\ell}^{(2)} a_{\ell\ell}^{(2i)} + \sum_{\substack{h=1 \\ h\neq\ell}}^{N} a_{kh}^{(2)} a_{h\ell}^{(2i)} \\ &\leq a_{k\ell}^{(2)} a_{\ell\ell}^{(2i)} \; + \sum_{\substack{h=1 \\ h\neq\ell}}^{N} a_{kh}^{(2)} \big(\beta_i \, a_{h\ell} + \gamma_i\big) \\ &\leq \alpha_i \, a_{k\ell}^{(2)} + \beta_i \left(a_{kk}^{(2)} a_{k\ell} + \sum_{\substack{h=1 \\ h\neq k,\ell}}^{N} a_{kh}^{(2)} a_{h\ell} \right) + \gamma_i \sum_{\substack{h=1 \\ h\neq\ell}}^{N} a_{kh}^{(2)}, \end{aligned} \tag{B.15}$$

where the first inequality follows by applying the induction hypothesis to term $a_{h\ell}^{(2i)}$, while in the second inequality we used (B.3). Let us now bound the individual terms that multiply the quantities α_i, β_i, and γ_i in (B.15).

- From (B.14) we have:

$$a_{k\ell}^{(2)} \leq 2\rho \, a_{k\ell} + \mathfrak{M}(A). \tag{B.16}$$

- From (B.9) we have:

$$a_{kk}^{(2)} \leq \rho^2, \qquad \sum_{\substack{h=1 \\ h\neq\ell}}^{N} a_{kh}^{(2)} \leq \rho^2. \tag{B.17}$$

- We can write:

$$\begin{aligned} \sum_{\substack{h=1 \\ h\neq k,\ell}}^{N} a_{kh}^{(2)} a_{h\ell} &\leq \sum_{\substack{h=1 \\ h\neq k,\ell}}^{N} \big(2\rho \, a_{kh} + \mathfrak{M}(A)\big) a_{h\ell} \\ &\leq 2\rho\mathfrak{M}(A) + \rho\mathfrak{M}(A) = 3\rho\mathfrak{M}(A), \end{aligned} \tag{B.18}$$

  where in the first inequality we applied (B.14), while in the second inequality we applied (B.1) and (B.2).

Using (B.16), (B.17), and (B.18) in (B.15), we get:

$$\begin{aligned} a_{k\ell}^{(2i+2)} &\leq \alpha_i \big(2\rho \, a_{k\ell} + \mathfrak{M}(A)\big) + \beta_i \big(\rho^2 a_{k\ell} + 3\rho\mathfrak{M}(A)\big) + \gamma_i \, \rho^2 \\ &= \beta_{i+1} a_{k\ell} + \gamma_{i+1}, \end{aligned} \tag{B.19}$$

where the equality comes from (B.6) and (B.7), and the proof is complete. ■

Lemma 15 (Bounds on a Useful Matrix Power Series). *Let A be an $N \times N$ scaled left-stochastic matrix as in (B.1). Let*

$$C \triangleq [A^2]_{\mathcal{P}'}, \qquad H \triangleq (I_{\mathcal{P}'} - C)^{-1} = \sum_{i=0}^{\infty} C^i, \tag{B.20}$$

where we recall that $\mathcal{P}' \triangleq \{1, 2, \ldots, N\} \setminus \mathcal{P}$. Let further

$$\bar{\alpha} \triangleq 1 + \frac{\rho^2}{(1-\rho^2)^2}, \quad \bar{\beta} \triangleq 2\rho\frac{\bar{\alpha}}{1-\rho^2}, \quad \bar{\gamma} \triangleq \frac{\bar{\alpha} + 3\rho\bar{\beta}}{1-\rho^2}. \tag{B.21}$$

Then, for $i = 1, 2, \ldots,$ we have that:

- *The main diagonal entries of matrix H satisfy the inequalities:*

$$0 < h_{kk} \leq \bar{\alpha}. \tag{B.22}$$

- *The off-diagonal entries of matrix H satisfy the inequalities:*

$$0 \leq h_{k\ell} \leq \bar{\beta} a_{k\ell} + \mathfrak{M}(A)\bar{\gamma}. \tag{B.23}$$

Proof: The fact that $h_{k\ell} \geq 0$ for any k and ℓ is an immediate consequence of the definition of matrix H in (B.20), since the entries of matrix powers C^i are nonnegative for any i. So in the next we will focus on the upper bounds in (B.22) and (B.23).

As it can be trivially verified by induction, we first note that for any $k, \ell \in \mathcal{P}'$ and any $i = 1, 2, \ldots$:

$$c_{k\ell}^{(i)} \leq a_{k\ell}^{(2i)}, \tag{B.24}$$

implying that the upper bounds provided in Lemma 14 are also valid for the matrix powers C^i. Therefore, by the definition of H in (B.20), we have:

$$h_{kk} \leq 1 + \sum_{i=1}^{\infty} \alpha_i, \qquad h_{k\ell} \leq \sum_{i=1}^{\infty} \beta_i a_{k\ell} + \sum_{i=1}^{\infty} \gamma_i, \tag{B.25}$$

where the sequences α_i, β_i and γ_i are defined in (B.4), (B.6) and (B.7), respectively. According to (B.25), to establish the upper bounds in (B.22) and (B.23) it suffices to show that:

$$1 + \sum_{i=1}^{\infty} \alpha_i = \bar{\alpha}, \qquad \sum_{i=1}^{\infty} \beta_i = \bar{\beta}, \qquad \sum_{i=1}^{\infty} \gamma_i = \mathfrak{M}(A)\,\bar{\gamma}. \tag{B.26}$$

To this aim, we note that the sequence α_i in (B.4) matches (E.1) in Lemma 20 with the choices:

$$f_1 = a = d = \rho^2, \qquad b = c = 0. \tag{B.27}$$

Therefore, we can apply (E.2) to obtain:

$$\alpha_i = \rho^{2(i-1)} \left(\rho^2 + \rho^2(i-1)\right) = i\,\rho^{2i}, \tag{B.28}$$

which, recalling the series $\sum_{i=1}^{\infty} i\,a^i = \frac{a}{(1-a)^2}$ (for $|a|<1$), allows us to write:

$$1 + \sum_{i=1}^{\infty} \alpha_i = 1 + \frac{\rho^2}{(1-\rho^2)^2} = \bar{\alpha}. \tag{B.29}$$

Thus, we proved (B.22).

Let us move on to prove (B.23). By substituting (B.28) in (B.6), we get:

$$\beta_{i+1} = \rho^2 \beta_i + 2\rho^{2i+1} i, \tag{B.30}$$

and therefore we see that the sequence β_i matches (E.1) in Lemma 20 with the choices:

$$f_1 = c = 2\rho, \qquad a = \rho^2, \qquad b = d = 0. \tag{B.31}$$

In view of (E.2), we conclude that:

$$\beta_i = \rho^{2(i-1)} \left(2\rho + 2\rho \frac{i(i-1)}{2}\right) = \rho^{2i-1}\left(i^2 - i + 2\right), \tag{B.32}$$

which implies that the series $\sum_{i=1}^{\infty} \beta_i$ converges. Since we showed that also the series $\sum_{i=1}^{\infty} \alpha_i$ is convergent, by summing over index i in (B.6) we can write:

$$\sum_{i=1}^{\infty} \beta_{i+1} = 2\rho \sum_{i=1}^{\infty} \alpha_i + \rho^2 \sum_{i=1}^{\infty} \beta_i, \tag{B.33}$$

or

$$\sum_{i=1}^{\infty} \beta_i - \beta_1 = 2\rho \sum_{i=1}^{\infty} \alpha_i + \rho^2 \sum_{i=1}^{\infty} \beta_i, \tag{B.34}$$

which, using $\beta_1 = 2\rho$ and (B.29), yields:

$$\sum_{i=1}^{\infty} \beta_i = 2\rho \frac{1 + \sum_{i=1}^{\infty} \alpha_i}{1-\rho^2} = 2\rho \frac{\bar{\alpha}}{1-\rho^2} = \bar{\beta}. \tag{B.35}$$

It remains to examine the behavior of the summation in (B.25) involving the sequence γ_i in (B.7). Substituting (B.28) and (B.32) in (B.7) we get:

$$\begin{aligned} \gamma_{i+1} &= \rho^2 \gamma_i + \mathfrak{M}(A)\rho^{2i} i + 3\mathfrak{M}(A)\rho^{2i}\left(i^2 - i + 2\right) \\ &= \rho^2 \gamma_i + \mathfrak{M}(A)\rho^{2i}\left(3i^2 - 2i + 6\right), \end{aligned} \tag{B.36}$$

which shows that the sequence γ_i matches (E.1) in Lemma 20 with the choices:

$$f_1 = \mathfrak{M}(A), \quad a = \rho^2, \quad b = 3\mathfrak{M}(A), \quad c = -2\mathfrak{M}(A), \quad d = 6\mathfrak{M}(A). \tag{B.37}$$

We conclude that the series $\sum_{i=1}^{\infty} \gamma_i$ is convergent. Thus, by summing over i in (B.7), we can write:

$$\sum_{i=1}^{\infty} \gamma_i - \gamma_1 = \mathfrak{M}(A) \sum_{i=1}^{\infty} \alpha_i + 3\mathfrak{M}(A) \sum_{i=1}^{\infty} \beta_i + \rho^2 \sum_{i=1}^{\infty} \gamma_i, \tag{B.38}$$

which, using $\gamma_1 = \mathfrak{M}(A)$ along with (B.29) and (B.35), yields:

$$\sum_{i=1}^{\infty} \gamma_i = \mathfrak{M}(A) \frac{\bar{\alpha} + 3\rho\bar{\beta}}{1-\rho^2} = \mathfrak{M}(A)\, \bar{\gamma}, \tag{B.39}$$

and the proof is complete. ■

We are now ready to apply Lemmas 14 and 15 to obtain a bound on the error of the limiting Granger estimator. By definition, the limiting Granger estimator $\widehat{\boldsymbol{A}}_{\mathcal{P}}(N)$ is a *deterministic* function of the combination matrix $\boldsymbol{A}(N)$. In fact, for a realization A of $\boldsymbol{A}(N)$, the limiting Granger estimator is:

$$\widehat{A}_{\mathcal{P}} \triangleq [R_1]_{\mathcal{P}} [R_0]_{\mathcal{P}}^{-1}, \tag{B.40}$$

with:

$$R_0 \triangleq \sum_{i=0}^{\infty} A^i [A^i]^{\top}, \qquad R_1 \triangleq A R_0. \tag{B.41}$$

When A is symmetric, we have (recall that I denotes the $N \times N$ identity matrix):

$$R_0 = (I - A^2)^{-1}, \tag{B.42}$$

and the limiting Granger estimator admits the following expression, first proved in Appendix A of [77] and in particular corresponding to Eq. (66) of [77]:

$$\widehat{A}_{\mathcal{P}} = A_{\mathcal{P}} + A_{\mathcal{P}\mathcal{P}'} (I_{\mathcal{P}'} - [A^2]_{\mathcal{P}'})^{-1} [A^2]_{\mathcal{P}'\mathcal{P}}, \tag{B.43}$$

which is critical to prove the following lemma. Thus, with the notation introduced in (B.20), when A is symmetric the limiting Granger estimator admits the following representation:

$$\widehat{A}_{\mathcal{P}} - A_{\mathcal{P}} = A_{\mathcal{P}\mathcal{P}'} H [A^2]_{\mathcal{P}'\mathcal{P}}, \tag{B.44}$$

or, in terms of the individual (k,ℓ)-entry:

$$\begin{aligned} e_{k\ell} \triangleq \left[A_{\mathcal{P}\mathcal{P}'} H [A^2]_{\mathcal{P}'\mathcal{P}}\right]_{k\ell} &= \sum_{j,m \in \mathcal{P}'} a_{kj} h_{jm} a^{(2)}_{m\ell} \\ &= \sum_{j \in \mathcal{P}'} a_{kj} h_{jj} a^{(2)}_{j\ell} + \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} h_{jm} a^{(2)}_{m\ell}\,. \end{aligned} \tag{B.45}$$

The following result provides a useful bound for such error entry.

Lemma 16 (Bound on the Error of the Limiting Granger Estimator). *Let A be an $N \times N$ scaled left-stochastic matrix as in (B.1). If A is symmetric, then for any $\mathcal{P}$ and any $k, \ell \in \mathcal{P}$ with $k \neq \ell$ we have:*

$$0 \leq e_{k\ell} \leq \kappa\, \mathfrak{M}(A) \qquad \forall \mathcal{P} \subseteq \{1, 2, \ldots, N\}, \tag{B.46}$$

where κ is a positive constant, and $\mathfrak{M}(A)$ is defined in (B.2).

Proof: We note that $e_{k\ell} \geq 0$ since all involved matrices are nonnegative — see (B.22) and (B.23) for what concerns H. Therefore, it suffices to prove that:

$$e_{k\ell} \leq \kappa \mathfrak{M}(A), \tag{B.47}$$

for some positive constant κ. To this aim, let us consider two indices $k, \ell \in \mathcal{P}$ with $k \neq \ell$. Calling upon Lemma 15, we can apply (B.22) and (B.23) in (B.45), yielding:

$$e_{k\ell} \leq \bar{\alpha} \sum_{j \in \mathcal{P}'} a_{kj} a^{(2)}_{j\ell} + \bar{\beta} \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} a_{jm} a^{(2)}_{m\ell} + \mathfrak{M}(A) \bar{\gamma} \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} a^{(2)}_{m\ell}. \tag{B.48}$$

The first summation in (B.48) can be upper bounded as follows:

$$\begin{aligned} \sum_{j \in \mathcal{P}'} a_{kj} a^{(2)}_{j\ell} &\leq \sum_{j \in \mathcal{P}'} a_{kj} \big(2\rho\, a_{j\ell} + \mathfrak{M}(A)\big) \\ &= 2\rho \sum_{j \in \mathcal{P}'} a_{kj} a_{j\ell} + \mathfrak{M}(A) \sum_{j \in \mathcal{P}'} a_{kj} \\ &\leq 2\rho \mathfrak{M}(A) + \mathfrak{M}(A) = 3\mathfrak{M}(A), \end{aligned} \tag{B.49}$$

where in the first inequality we used (B.14), while in the second inequality we used the following bounds:

$$\sum_{j \in \mathcal{P}'} a_{kj} a_{j\ell} \leq \sum_{\substack{j=1 \\ j \neq k,\ell}}^{N} a_{kj} a_{j\ell} \leq \mathfrak{M}(A), \qquad \sum_{j \in \mathcal{P}'} a_{kj} \leq \rho. \tag{B.50}$$

Here, we remark that the inequality on the left exploits the fact that $k, \ell \in \mathcal{P}$ and $j \in \mathcal{P}'$, so we are allowed to extend the sum across $j \in \mathcal{P}'$ to a sum across $j \in \{1, 2, \ldots, N\} \backslash \{k, \ell\}$.

Next we focus on the second summation in (B.48), which can be upper bounded as follows:

$$\begin{aligned} \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} a_{jm} a^{(2)}_{m\ell} &\leq \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} a_{jm} \big(2\rho\, a_{m\ell} + \mathfrak{M}(A)\big) \\ &= 2\rho \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} a_{jm} a_{m\ell} + \mathfrak{M}(A) \sum_{\substack{j,m \in \mathcal{P}' \\ j \neq m}} a_{kj} a_{jm}, \\ &\leq 2\rho^2 \mathfrak{M}(A) + \rho^2 \mathfrak{M}(A) = 3\rho^2 \mathfrak{M}(A), \end{aligned} \tag{B.51}$$

where in the first inequality we used (B.14), while in the second inequality we used (B.1) and (B.2) to get:

$$\sum_{\substack{j,m\in\mathcal{P}'\\ j\neq m}} a_{kj}a_{jm}a_{m\ell} = \sum_{j\in\mathcal{P}'} a_{kj} \sum_{\substack{m\in\mathcal{P}'\\ j\neq m}} a_{jm}a_{m\ell} \leq \sum_{j\in\mathcal{P}'} a_{kj}\mathfrak{M}(A) \leq \rho\mathfrak{M}(A), \tag{B.52}$$

and:

$$\sum_{\substack{j,m\in\mathcal{P}'\\ j\neq m}} a_{kj}a_{jm} = \sum_{j\in\mathcal{P}'} a_{kj} \sum_{\substack{m\in\mathcal{P}'\\ j\neq m}} a_{jm} \leq \sum_{j\in\mathcal{P}'} a_{kj}\rho \leq \rho^2. \tag{B.53}$$

Finally, the third summation in (B.48) can be manipulated as follows:

$$\begin{aligned}\sum_{\substack{j,m\in\mathcal{P}'\\ j\neq m}} a_{kj}a^{(2)}_{m\ell} &= \sum_{\substack{j,m\in\mathcal{P}'\\ j\neq m}} a_{kj} \sum_{h=1}^{N} a_{mh}a_{h\ell} \\ &= \sum_{j\in\mathcal{P}'} a_{kj} \sum_{h=1}^{N} a_{h\ell} \sum_{\substack{m\in\mathcal{P}'\\ m\neq j}} a_{mh} \leq \rho^3,\end{aligned} \tag{B.54}$$

where in the last step we applied repeatedly (B.1), further noticing that in view of the symmetry of A we can write:

$$\sum_{\substack{m\in\mathcal{P}'\\ m\neq j}} a_{mh} = \sum_{\substack{m\in\mathcal{P}'\\ m\neq j}} a_{hm} \leq \rho. \tag{B.55}$$

Using (B.49), (B.51), and (B.54) in (B.48), we get:

$$e_{k\ell} \leq \left(3\rho\bar{\alpha} + 3\rho^2\bar{\beta} + \rho^3\bar{\gamma}\right)\mathfrak{M}(A) \triangleq \kappa\mathfrak{M}(A), \tag{B.56}$$

which proves the claim. ■

Appendix C

Useful Convergence Results

In the previous appendix we obtained some upper and lower bounds on the limiting Granger estimation error, and in particular we have seen that these bounds are functions of the term $\mathfrak{M}(A)$ defined in (B.2). In this appendix we will provide some useful convergence results involving these quantities, which will be exploited later to prove the regularity of the limiting Granger estimator.

Lemma 17 (Useful Convergence Results for Bollobás-Riordan Graphs). *Let $\boldsymbol{A}(N)$ be the Laplacian combination matrix (4.15) with support graph $\mathcal{G}(N)$ obtained from a Bollobás-Riordan multigraph $\mathfrak{M}(N)$ with step parameter η. Then we have that:*

$$\sqrt{N}\,\mathfrak{M}\big(\boldsymbol{A}(N)\big) \xrightarrow{p} 0, \tag{C.1}$$

where from (B.2):

$$\mathfrak{M}\big(\boldsymbol{A}(N)\big) = \max_{\substack{k,\ell\in[1,N]\\ k\neq\ell}} \sum_{\substack{j=1\\ j\neq k,\ell}}^{N} \boldsymbol{a}_{jk}(N)\boldsymbol{a}_{j\ell}(N). \tag{C.2}$$

Proof: Using (4.15) in (C.2) we get:

$$\sqrt{N}\;\mathfrak{M}\big(\boldsymbol{A}(N)\big) = \left(\frac{\rho\lambda\sqrt{N}}{1+\boldsymbol{\mu}_{\mathcal{G}}(N)}\right)^{2} \underbrace{\frac{1}{\sqrt{N}} \max_{\substack{k,\ell\in[1,N]\\ k<\ell}} \sum_{\substack{j=1\\ j\neq k,\ell}}^{N} \boldsymbol{g}_{jk}\boldsymbol{g}_{j\ell}}_{\boldsymbol{t}_N}. \tag{C.3}$$

Note that, by exploiting the fact that $\boldsymbol{A}(N)$ is symmetric by assumption, we have replaced the constraint $k \neq \ell$ in (C.2) with $k < \ell$ in (C.3). Since the term $\frac{\sqrt{N}}{1+\boldsymbol{\mu}_{\mathcal{G}}(N)}$ converges almost surely to $1/\boldsymbol{\mu}$, it is sufficient to show that the term $\boldsymbol{t}_N$ in (C.3) vanishes in probability as

$N \to \infty$. To this aim, it is expedient to work in terms of the original multigraph $\mathfrak{M}(N)$ that originates the simple graph $\mathcal{G}(N)$. We have that:

$$\begin{aligned} \boldsymbol{t}_N &\leq \frac{1}{\sqrt{N}} \max_{\substack{k,\ell \in [1,N] \\ k<\ell}} \sum_{\substack{j=1 \\ j\neq k,\ell}}^{N} \boldsymbol{m}_{jk}\boldsymbol{m}_{j\ell} \\ &= \underbrace{\frac{1}{\sqrt{N}} \max_{\substack{k,\ell \in [1,N] \\ k<\ell}} \sum_{\substack{j=1 \\ j\neq k,\ell}}^{\ell} \boldsymbol{m}_{jk}\boldsymbol{m}_{j\ell}}_{\boldsymbol{t}'_N} \\ &+ \underbrace{\frac{1}{\sqrt{N}} \max_{\substack{k,\ell \in [1,N] \\ k<\ell}} \sum_{j=\ell+1}^{N} \boldsymbol{m}_{jk}\boldsymbol{m}_{j\ell}}_{\boldsymbol{t}''_N}. \end{aligned} \tag{C.4}$$

By construction, $\boldsymbol{m}_{k\ell} \in \{1, \ldots, \eta\}$, and using (5.4), for any $k < \ell$ we have:

$$\sum_{\substack{j=1 \\ j\neq k,\ell}}^{\ell} \boldsymbol{m}_{jk}\boldsymbol{m}_{j\ell} \;\leq\; \eta \sum_{\substack{j=1 \\ j\neq k,\ell}}^{\ell} \boldsymbol{m}_{j\ell} \;\leq\; \eta\, \boldsymbol{d}_{\mathcal{M},\ell}(\ell) \;\leq\; 2\eta^2, \tag{C.5}$$

where the last inequality holds because, in the multigraph $\mathfrak{M}(\ell)$, node ℓ has only η edges, and so its degree $\boldsymbol{d}_{\mathcal{M},\ell}(\ell)$ is upper bounded by 2η. Applying (C.5) to the random sequence $\boldsymbol{t}'_N$ in (C.4), we conclude that $\boldsymbol{t}'_N$ vanishes almost surely as $N \to \infty$. It remains to show that $\boldsymbol{t}''_N$ in (C.4) vanishes in probability. To this end, we call upon Lemma 21, by introducing the following family of sequences, for any $k, \ell \in \mathbb{N}$ with $1 \leq k < \ell$:

$$\boldsymbol{u}_{k\ell}(j) \triangleq \begin{cases} \boldsymbol{m}_{jk}\,\boldsymbol{m}_{j\ell}, & j > \ell, \\ 0, & \text{otherwise.} \end{cases} \tag{C.6}$$

Following the notation adopted in Lemma 21, we have a family of sequences $\{\boldsymbol{u}_{k\ell}(j)\}_{j\geq 1}$ parameterized by the set:

$$\Theta = \{(k,\ell) \in \mathbb{N}^2 : 1 \leq k < \ell\}. \tag{C.7}$$

Moreover, we introduce the aggregate variable:

$$\boldsymbol{U}_{k\ell}(N) \triangleq \sum_{j=1}^{N} \boldsymbol{u}_{k\ell}(j). \tag{C.8}$$

Convergence to zero of the random variable $\boldsymbol{t}''_N$ in (C.4) is equivalent to the following statement:

$$\forall \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\left[\max_{\substack{k,\ell \in [1,N] \\ k<\ell}} \boldsymbol{U}_{k\ell}(N) > \varepsilon\sqrt{N} \right] = 0. \tag{C.9}$$

Now, for any $(k,\ell) \in \Theta$:

$$\boldsymbol{u}_{k\ell}(1) = 0, \qquad \text{and} \qquad \forall j > 1, \quad 0 \leq \boldsymbol{u}_{k\ell}(j) \leq \eta^2, \tag{C.10}$$

where the upper bound follows because, in view of (5.2) and (5.3), both $\boldsymbol{m}_{jk}$ and $\boldsymbol{m}_{j\ell}$ cannot exceed the number of steps η. From (C.10) we see that the family of sequences in (C.6) meets the hypotheses of Lemma 21 with the filtration $\{\mathcal{F}(n)\}_{n\geq 1}$ generated by the random sequence $\{\boldsymbol{\mathcal{M}}(n)\}_{n\geq 1}$. We conclude that the probability in (C.9) is upper bounded by:

$$\frac{N(N-1)}{2}\, e^{-\frac{3}{16\eta^2}\varepsilon\sqrt{N}} + \mathbb{P}\left[\max_{\substack{k,\ell\in[1,N]\\ k<\ell}} \boldsymbol{C}_{k\ell}(N) > \frac{\varepsilon}{2}\sqrt{N}\right], \tag{C.11}$$

where:

$$\boldsymbol{C}_{k\ell}(N) \triangleq \sum_{j=1}^{N} \mathbb{E}[\boldsymbol{u}_{k\ell}(j)|\boldsymbol{\mathcal{M}}(j-1)]. \tag{C.12}$$

On the other hand, from Lemma 8 we have that:

$$\max_{\substack{k,\ell\in[1,N]\\ k<\ell}} \boldsymbol{C}_{k\ell}(N) < \sum_{j=1}^{N} \left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(j-1) + 2\eta}{j-1}\right)^2. \tag{C.13}$$

Therefore, applying Markov's inequality and (C.13) to the second term in (C.11), we conclude that this term is upper bounded by:

$$\frac{2}{\varepsilon\sqrt{N}} \sum_{j=1}^{N} \mathbb{E}\left[\left(\frac{\boldsymbol{\mu}_{\mathcal{M}}(j-1) + 2\eta}{j-1}\right)^2\right], \tag{C.14}$$

which vanishes as $N \to \infty$ in view of Lemma 7, concluding the proof of the theorem. ■

Appendix D

Graph Learning over Erdős-Rényi Graphs

Proof of Lemma 1: Since by assumption $0 \leq p < 1$ and $|S_N| \overset{N\to\infty}{\longrightarrow} \infty$, the claim in (4.9) is immediately verified since we have that:

$$\mathbb{P}\big[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully connected}\big] = p_N^{\frac{|\mathcal{S}_N|(|\mathcal{S}_N|-1)}{2}} \overset{N\to\infty}{\longrightarrow} 0. \tag{D.1}$$

Similarly, when the limiting connection probability p is in $(0,1)$, since $|S_N| \overset{N\to\infty}{\longrightarrow} \infty$, the convergence in (4.11) is immediately verified as follows:

$$\mathbb{P}\big[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully disconnected}\big] = \big(1-p_N\big)^{\frac{|\mathcal{S}_N|(|\mathcal{S}_N|-1)}{2}} \overset{N\to\infty}{\longrightarrow} 0. \tag{D.2}$$

It remain to prove (4.11) when $p = 0$. In this case we first write:

$$\mathbb{P}\big[\mathcal{G}_{\mathcal{S}}(N) \text{ is fully disconnected}\big] = (1-p_N)^{\frac{|\mathcal{S}_N|(|\mathcal{S}_N|-1)}{2}} = \exp\left\{\frac{|\mathcal{S}_N|(|\mathcal{S}_N|-1)}{2}\log(1-p_N)\right\}. \tag{D.3}$$

Now since we assumed:

$$|\mathcal{S}_N|^2 p_N \overset{N\to\infty}{\longrightarrow} \infty, \tag{D.4}$$

and considering the well known limits (for $p_N \to 0$):

$$\frac{\log(1-p_N)}{p_N} \overset{N\to\infty}{\longrightarrow} -1, \qquad \frac{|\mathcal{S}_N|(|\mathcal{S}_N|-1)}{|\mathcal{S}_N|^2} \overset{N\to\infty}{\longrightarrow} 1, \tag{D.5}$$

we have that the probability in (D.3) vanishes as $N \to \infty$, which concludes the proof. ■

Proof of Lemma 2: Let us introduce the quantities:

$$\boldsymbol{\delta}_{\nu,N} \triangleq \frac{\kappa}{1+\boldsymbol{\nu}_{\mathcal{G}}(N)}, \qquad \boldsymbol{\delta}_{\mu,N} \triangleq \frac{\kappa}{1+\boldsymbol{\mu}_{\mathcal{G}}(N)}. \tag{D.6}$$

Since by assumption we have (4.5), which implies:

$$Np_N \overset{N\to\infty}{\longrightarrow} \infty, \tag{D.7}$$

and recalling Lemma 4, we obtain:

$$Np_N\, \boldsymbol{\delta}_{\nu,N} \xrightarrow{\text{p}} \kappa, \qquad \text{and} \qquad Np_N\, \boldsymbol{\delta}_{\mu,N} \xrightarrow{\text{p}} \kappa. \tag{D.8}$$

From definition (4.17), a regular diffusion matrix is such that:

$$\boldsymbol{\delta}_{\mu,N}\, \boldsymbol{G}(N) \leq \boldsymbol{A}(N) - \text{diag}(\boldsymbol{A}(N)) \leq \boldsymbol{\delta}_{\nu,N}\, \boldsymbol{G}(N), \tag{D.9}$$

where diag$(\cdot)$ is a diagonal matrix having on the main diagonal the entries of its matrix argument. Thus, using (D.8) and (D.9) we can write:

$$\|Np_N \boldsymbol{A}(N) - \kappa \boldsymbol{G}(N)\|_{\text{max-off}} \leq \max\left\{ \left|Np_N \boldsymbol{\delta}_{\nu,N} - \kappa\right|, \left|Np_N \boldsymbol{\delta}_{\mu,N} - \kappa\right| \right\} \xrightarrow{\text{p}} 0, \tag{D.10}$$

which concludes the proof. ■

The proof of Lemma 3, originally presented in [75], relies on a bound on the Granger estimator different from the bound found in Lemma 16. This bound is detailed in the next lemma.

Lemma 18 (Other Useful Bounds on the Error of the Limiting Granger Estimator). *Let A be an $N \times N$ scaled left-stochastic matrix as in (B.1). If A is symmetric, then we have for any $\mathcal{P}$ and any $k, \ell \in \mathcal{P}$ with $k \neq \ell$:*

$$\begin{aligned} e_{k\ell} \;\leq\; & \overline{\boldsymbol{\alpha}}(A)\Big[2\,\mathfrak{S}(A)\,\mathfrak{M}(A,\mathcal{P}') + \mathfrak{M}(A)\,\widetilde{\mathfrak{B}}(A,\mathcal{P}')\Big] \\ + \; & \overline{\boldsymbol{\beta}}(A)\Big[2\,\mathfrak{S}(A)\,\mathfrak{B}_3(A,\mathcal{P}') + \mathfrak{M}(A)\,\widetilde{\mathfrak{M}}(A,\mathcal{P}')\Big] \\ + \; & \overline{\boldsymbol{\gamma}}(A)\Big[2\,\mathfrak{S}(A)\,\widetilde{\widetilde{\mathfrak{M}}}(A,\mathcal{P}') + \widetilde{\widetilde{\mathfrak{B}}}(A,\mathcal{P}')\Big]\mathfrak{M}(A) \end{aligned} \tag{D.11}$$

$$\begin{aligned} e_{k\ell} \;\geq\; & \underline{\boldsymbol{\alpha}}(A)\left[2\,\mathfrak{s}(A)\,\mathfrak{m}(A,\mathcal{P}') + \mathfrak{m}(A)\,\widetilde{\mathfrak{b}}(A,\mathcal{P}')\right] \\ + \; & \underline{\boldsymbol{\beta}}(A)\left[2\,\mathfrak{s}(A)\,\mathfrak{b}_3(A,\mathcal{P}') + \mathfrak{m}(A)\,\widetilde{\mathfrak{m}}(A,\mathcal{P}')\right] \\ + \; & \underline{\boldsymbol{\gamma}}(A)\left[2\,\mathfrak{s}(A)\,\widetilde{\widetilde{\mathfrak{m}}}(A,\mathcal{P}') + \widetilde{\widetilde{\mathfrak{b}}}(A,\mathcal{P}')\right]\mathfrak{m}(A) \end{aligned} \tag{D.12}$$

where the new quantities introduced in these bounds are defined in Table D.1.

Proof: See Appendices B, C, D and G of [75]. ■

In order to prove Lemma 3, we then need to characterize the asymptotic behavior of the bounds in Lemma 18. This task can be accomplished thanks to the convergence results stated in the next lemma.

Lemma 19 (Useful Convergence Results for Erdős-Rényi Graphs). *Let $\boldsymbol{A}(N)$ be a regular diffusion matrix (4.17) with support graph $\mathcal{G}(N)$ drawn according to the Erdős-Rényi model under the degree concentration regime (4.6). Let $\mathcal{S}_N$ be any probed subset sequence such that:*

$$\lim_{N\to\infty} \frac{|\mathcal{S}_N|}{N} = \xi, \qquad \textit{for some } 0 < \xi < 1. \tag{D.13}$$

Then, the quantities defined in Table D.1 satisfy:

$$Np_N\,\mathfrak{M}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \kappa^2 p \qquad Np_N\,\mathfrak{m}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \kappa^2 p \tag{D.14}$$

$$\mathfrak{S}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \kappa^2(\rho-\kappa) \qquad \mathfrak{s}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \kappa^2(\rho-\kappa) \tag{D.15}$$

$$Np_N\mathfrak{M}\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^2 p(1-\xi) \qquad Np_N\mathfrak{m}\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^2 p(1-\xi) \tag{D.16}$$

$$\widetilde{\mathfrak{B}}\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa(1-\xi) \qquad \widetilde{\mathfrak{b}}\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa(1-\xi) \tag{D.17}$$

$$\mathfrak{B}_3\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^3\rho(1-\xi)^2 \qquad \mathfrak{b}_3\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^3\rho(1-\xi)^2 \tag{D.18}$$

$$\widetilde{\mathfrak{M}}(A,\mathcal{P}')\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^2(1-\xi)^2 \qquad \widetilde{\mathfrak{m}}(A,\mathcal{P}')\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^2(1-\xi)^2 \tag{D.19}$$

$$\widetilde{\widetilde{\mathfrak{M}}}(A,\mathcal{P}')\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^2(1-\xi)^2 \qquad \widetilde{\widetilde{\mathfrak{m}}}(A,\mathcal{P}')\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^2(1-\xi)^2 \tag{D.20}$$

$$\widetilde{\widetilde{\mathfrak{B}}}(A,\mathcal{P}')\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^3(1-\xi)^2 \qquad \widetilde{\widetilde{\mathfrak{b}}}(A,\mathcal{P}')\big(\boldsymbol{A}(N),\mathcal{S}'_N\big) \xrightarrow{\text{p}} \kappa^3(1-\xi)^2. \tag{D.21}$$

and:

$$\overline{\boldsymbol{\alpha}}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} 1 + \frac{\zeta^2}{1-\zeta^2} \qquad \underline{\boldsymbol{\alpha}}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} 1 + \frac{\zeta^2}{1-\zeta^2} \tag{D.22}$$

$$\overline{\boldsymbol{\beta}}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \frac{2\zeta}{(1-\zeta^2)^2} \qquad \underline{\boldsymbol{\beta}}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \frac{2\zeta}{(1-\zeta^2)^2} \tag{D.23}$$

$$\overline{\boldsymbol{\gamma}}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \varphi \qquad \underline{\boldsymbol{\gamma}}\big(\boldsymbol{A}(N)\big) \xrightarrow{\text{p}} \varphi \tag{D.24}$$

for:

$$\varphi \triangleq \frac{1-\zeta^2+2\,\zeta[2\,\zeta\,(1-\xi)+\kappa(1-\xi)]}{[1-(\rho^2-2\rho\kappa\xi+\kappa^2\xi)][1-\zeta^2]^2} \qquad \textit{and} \qquad \zeta \triangleq \rho-\kappa. \tag{D.25}$$

Proof: See Appendix F in [75]. ■

Now, we are ready to prove Lemma 3.

Proof of Lemma 3: It is well known that for any set of values $\mathcal{X} \subseteq \mathbb{R}$ such that:

$$\forall x \in \mathcal{X} \quad a \leq x \leq b, \tag{D.26}$$

with $a, b \in \mathbb{R}$ some constants, we can write:

$$0 \leq \max_{x\in\mathcal{X}} |x| \leq |a| + |b|. \tag{D.27}$$

Now, let $\mathfrak{z}(A,\mathcal{P})$ and $\mathfrak{Z}(A,\mathcal{P})$ compactly denote the lower bound in (D.12) and the upper bound in (D.11), respectively. If we consider:

$$\mathcal{X} \triangleq \{Np_N[\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N)]_{k\ell} - \beta : k,\ell \in \mathcal{S}, k \neq \ell\}, \tag{D.28}$$

then by exploiting Lemma 18 we can use (D.27) with:

$$a \triangleq Np_N\mathfrak{z}(A,\mathcal{P}) - \beta, \qquad b \triangleq Np_N\mathfrak{Z}(A,\mathcal{P}) - \beta, \tag{D.29}$$

to obtain:

$$\left\|Np_N\left(\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N)\right) - \beta\right\|_{\text{max-off}} \leq \left|Np_N\mathfrak{z}\big(\boldsymbol{A}(N),\mathcal{S}_N\big) - \beta\right| + \left|Np_N\mathfrak{Z}\big(\boldsymbol{A}(N),\mathcal{S}_N\big) - \beta\right|. \tag{D.30}$$

We can use the convergences listed in Lemma 19 and, after some algebraic computations, we obtain for any sequence of probed subsets $\mathcal{S}_N$ satisfying (4.18):

$$Np_N\mathfrak{z}\big(\boldsymbol{A}(N),\mathcal{S}_N\big) \xrightarrow{\text{p}} \beta, \qquad Np_N\mathfrak{Z}\big(\boldsymbol{A}(N),\mathcal{S}_N\big) \xrightarrow{\text{p}} \beta. \tag{D.31}$$

From the squeeze theorem for convergence in probability, using (D.30) and (D.31) we get the claim:

$$\left\|Np_N\left(\widehat{\boldsymbol{A}}_{\mathcal{S}}(N) - \boldsymbol{A}_{\mathcal{S}}(N)\right) - \beta\right\|_{\text{max-off}} \xrightarrow{\text{p}} 0. \tag{D.32}$$

■

New terms in bounds (D.11), (D.12)		**Original notation [75]**
$\mathfrak{S}(A) \triangleq \max_{k\in[1,N]} a_{kk}$	$\mathfrak{s}(A) \triangleq \min_{k\in[1,N]} a_{kk}$	$\mathfrak{M}_{a,\text{self}}$ $\mathfrak{m}_{a,\text{self}}$
$\mathfrak{M}(A,\mathcal{P}') \triangleq \max_{\substack{k,\ell\in[1,N]\\k\neq\ell}} \sum_{\substack{j\in\mathcal{P}'\\j\neq k,\ell}} a_{kj}a_{j\ell}$	$\mathfrak{m}(A,\mathcal{P}') \triangleq \min_{\substack{k,\ell\in[1,N]\\k\neq\ell}} \sum_{\substack{j\in\mathcal{P}'\\j\neq k,\ell}} a_{kj}a_{j\ell}$	$\mathfrak{M}^{(\mathcal{P}')}$ $\mathfrak{m}^{(\mathcal{P}')}$
$\widetilde{\mathfrak{B}}(A,\mathcal{P}') \triangleq \max_{k\in\mathcal{P}} \sum_{\ell\in\mathcal{P}'} a_{k\ell}$	$\widetilde{\mathfrak{b}}(A,\mathcal{P}') \triangleq \min_{k\in\mathcal{P}} \sum_{\ell\in\mathcal{P}'} a_{k\ell}$	$\widetilde{\mathfrak{M}}^{(\mathcal{S}')}_{a,\text{sum}}$ $\widetilde{\mathfrak{m}}^{(\mathcal{S}')}_{a,\text{sum}}$
$\mathfrak{B}_3(A,\mathcal{P}') \triangleq \max_{\substack{k,\ell\in\mathcal{P}\\k\neq\ell}} \sum_{\substack{j,m\in\mathcal{P}'\\j\neq m}} a_{kj}a_{jm}a_{m\ell}$	$\mathfrak{b}_3(A,\mathcal{P}') \triangleq \min_{\substack{k,\ell\in\mathcal{P}\\k\neq\ell}} \sum_{\substack{j,m\in\mathcal{P}'\\j\neq m}} a_{kj}a_{jm}a_{m\ell}$	$\mathfrak{M}^{(\mathcal{S}')}_{a_3,\text{sum}}$ $\mathfrak{m}^{(\mathcal{S}')}_{a_3,\text{sum}}$
$\widetilde{\mathfrak{M}}(A,\mathcal{P}') \triangleq \max_{k\in\mathcal{P}} \sum_{\substack{j,m\in\mathcal{P}'\\j\neq m}} a_{kj}a_{jm}$	$\widetilde{\mathfrak{m}}(A,\mathcal{P}') \triangleq \min_{k\in\mathcal{P}} \sum_{\substack{j,m\in\mathcal{P}'\\j\neq m}} a_{kj}a_{jm}$	$\widetilde{\mathfrak{M}}^{(\mathcal{P}')}$ $\widetilde{\mathfrak{m}}^{(\mathcal{P}')}$
$\widetilde{\widetilde{\mathfrak{M}}}(A,\mathcal{P}') \triangleq \max_{\substack{k,\ell\in\mathcal{P}\\k\neq\ell}} \sum_{\substack{j,m\in\mathcal{P}'\\j\neq m}} a_{kj}a_{m\ell}$	$\widetilde{\widetilde{\mathfrak{m}}}(A,\mathcal{P}') \triangleq \min_{\substack{k,\ell\in\mathcal{P}\\k\neq\ell}} \sum_{\substack{j,m\in\mathcal{P}'\\j\neq m}} a_{kj}a_{m\ell}$	$\widetilde{\widetilde{\mathfrak{M}}}^{(\mathcal{P}')}$ $\widetilde{\widetilde{\mathfrak{m}}}^{(\mathcal{P}')}$
$\widetilde{\widetilde{\mathfrak{B}}}(A,\mathcal{P}') \triangleq \max_{k,\ell\in\mathcal{P}} \sum_{j\in\mathcal{P}'} a_{kj} \sum_{\substack{h=1\\h\neq\ell}}^{N} a_{h\ell} \sum_{\substack{m\in\mathcal{P}'\\m\neq j,h}} a_{mh}$	$\widetilde{\widetilde{\mathfrak{b}}}(A,\mathcal{P}') \triangleq \min_{k,\ell\in\mathcal{P}} \sum_{j\in\mathcal{P}'} a_{kj} \sum_{\substack{h=1\\h\neq\ell}}^{N} a_{h\ell} \sum_{\substack{m\in\mathcal{P}'\\m\neq j,h}} a_{mh}$	$\widetilde{\widetilde{\mathfrak{M}}}^{(\mathcal{P}')}_{a,\text{sum}}$ $\widetilde{\widetilde{\mathfrak{m}}}^{(\mathcal{P}')}_{a,\text{sum}}$
The terms $\overline{\boldsymbol{\alpha}}(A)$, $\overline{\boldsymbol{\beta}}(A)$ and $\overline{\boldsymbol{\gamma}}(A)$ have the same role as $\bar{\alpha}$, $\bar{\beta}$ and $\bar{\gamma}$ in (B.22) and (B.23), but produce sharper upper bounds.	The terms $\underline{\boldsymbol{\alpha}}(A)$, $\underline{\boldsymbol{\beta}}(A)$ and $\underline{\boldsymbol{\gamma}}(A)$ are similar to $\overline{\boldsymbol{\alpha}}(A)$, $\overline{\boldsymbol{\beta}}(A)$ and $\overline{\boldsymbol{\gamma}}(A)$, but correspond to sharper lower bounds.	$\overline{\boldsymbol{\Phi}}_\alpha$ $\underline{\boldsymbol{\Phi}}_\alpha$ $\overline{\boldsymbol{\Phi}}_\beta$ $\underline{\boldsymbol{\Phi}}_\beta$ $\overline{\boldsymbol{\Phi}}_\gamma$ $\underline{\boldsymbol{\Phi}}_\gamma$

Table D.1: Definitions of the new terms introduced in bounds (D.11), (D.12) of Lemma 18. They arise from the analysis conducted in [75] which is analogous but more general than the one presented in this appendix. We changed the original notation of these terms to conform them to the style adopted in this document. In the second column of the table we report the formulation used in [75].

Appendix E

Auxiliary Technical Results

Lemma 20. *Let f_i be the sequence recursively defined as:*

$$f_{i+1} = a f_i + a^i \left(b\, i^2 + c\, i + d\right), \qquad i = 1, 2, \ldots \tag{E.1}$$

with $0 < a < 1$ and $b, c, d \in \mathbb{R}$. Then we have that:

$$f_i = a^{i-1}\left(f_1 + b\frac{i(i-1)(2i-1)}{6} + c\frac{i(i-1)}{2} + d(i-1)\right). \tag{E.2}$$

Proof: Unfolding the recursion in (E.1), we conclude that, for all $i > 1$:

$$f_i = a^{i-1}\left(f_1 + \sum_{j=1}^{i-1}\left(bj^2 + cj + d\right)\right). \tag{E.3}$$

Thus, to obtain (E.2) we use the well-known results:

$$\sum_{j=1}^{i-1} j = \frac{i(i-1)}{2}, \qquad \sum_{j=1}^{i-1} j^2 = \frac{i(i-1)(2i-1)}{6}. \tag{E.4}$$

■

The following lemma is an adaptation of Theorem 2.1 in [38], useful for the proofs of Lemma 5 and Lemma 17.

Lemma 21. *Let us consider a family of random sequences $\{\boldsymbol{u}_\theta(n)\}_{n\geq 1}$ spanned by the parameter $\theta \in \Theta$ and defined on the same probability space. Assume that the following conditions are met for all $\theta \in \Theta$:*

$$\boldsymbol{u}_\theta(1) = 0, \qquad 0 \leq \boldsymbol{u}_\theta(n) \leq b \quad \text{for all } n > 1, \tag{E.5}$$

for a positive constant b. Let us further define the first two conditional moment sequences $\{\boldsymbol{\nu}_\theta(n)\}_{n\geq 1}$ and $\{\boldsymbol{\chi}_\theta(n)\}_{n\geq 1}$ w.r.t. a given filtration $\{\mathcal{F}(n)\}_{n\geq 1}$ of the underlying space:

$$\boldsymbol{\nu}_\theta(1) \triangleq 0, \quad \boldsymbol{\nu}_\theta(n) \triangleq \mathbb{E}[\boldsymbol{u}_\theta(n)|\mathcal{F}(n-1)], \tag{E.6}$$

$$\boldsymbol{\chi}_\theta(1) \triangleq 0, \quad \boldsymbol{\chi}_\theta(n) \triangleq \mathbb{E}[\boldsymbol{u}_\theta^2(n)|\mathcal{F}(n-1)], \tag{E.7}$$

and finally consider:

$$\boldsymbol{U}_\theta(N) \triangleq \sum_{n=1}^{N} \boldsymbol{u}_\theta(n), \qquad \boldsymbol{C}_\theta(N) \triangleq \sum_{n=1}^{N} \boldsymbol{\nu}_\theta(n), \tag{E.8}$$

$$\boldsymbol{Q}_\theta(N) \triangleq \sum_{n=1}^{N} \boldsymbol{\chi}_\theta(n), \qquad \bar{\boldsymbol{U}}_\theta(N) \triangleq \boldsymbol{U}_\theta(N) - \boldsymbol{C}_\theta(N). \tag{E.9}$$

Then, for any subset $\mathcal{T} \subseteq \Theta$ and any $u > 0$ we have:

$$\mathbb{P}\left[\max_{\theta\in\mathcal{T}} \boldsymbol{U}_\theta(N) > u\right] \leq |\mathcal{T}|\, e^{-\frac{3}{16b}u} + \mathbb{P}\left[\max_{\theta\in\mathcal{T}} \boldsymbol{C}_\theta(N) > \frac{u}{2}\right]. \tag{E.10}$$

Proof: For any two events $\mathcal{E}_1$ and $\mathcal{E}_2$, it is true that ($\bar{\mathcal{E}}_2$ is the complement of event $\mathcal{E}_2$):

$$\mathbb{P}[\mathcal{E}_1] = \mathbb{P}[\mathcal{E}_1, \mathcal{E}_2] + \mathbb{P}[\mathcal{E}_1, \bar{\mathcal{E}}_2] \leq \mathbb{P}[\mathcal{E}_1, \mathcal{E}_2] + \mathbb{P}[\bar{\mathcal{E}}_2], \tag{E.11}$$

so that we can write:

$$\begin{aligned}\mathbb{P}\left[\max_{\theta\in\mathcal{T}} \boldsymbol{U}_\theta(N) > u\right] \leq\ & \mathbb{P}\left[\max_{\theta\in\mathcal{T}} \boldsymbol{U}_\theta(N) > u,\ \max_{\theta\in\mathcal{T}} \boldsymbol{C}_\theta(N) \leq \frac{u}{2}\right] \\ & + \mathbb{P}\left[\max_{\theta\in\mathcal{T}} \boldsymbol{C}_\theta(N) > \frac{u}{2}\right].\end{aligned} \tag{E.12}$$

Let us focus on the first term on the RHS of (E.12). We have the following relations:

$$\begin{aligned}
&\left\{\bigcup_{\theta\in\mathcal{T}} \{\boldsymbol{U}_\theta(N) > u\}\right\} \bigcap \left\{\bigcap_{\theta'\in\mathcal{T}} \{\boldsymbol{C}_{\theta'}(N) \leq u/2\}\right\} \\
&\overset{(a)}{=} \left\{\bigcup_{\theta\in\mathcal{T}} \{\bar{\boldsymbol{U}}_\theta(N) > u - \boldsymbol{C}_\theta(N)\}\right\} \bigcap \left\{\bigcap_{\theta'\in\mathcal{T}} \{\boldsymbol{C}_{\theta'}(N) \leq u/2\}\right\} \\
&\overset{(b)}{\subseteq} \bigcup_{\theta\in\mathcal{T}} \{\bar{\boldsymbol{U}}_\theta(N) > u - \boldsymbol{C}_\theta(N)\} \cap \{\boldsymbol{C}_\theta(N) \leq u/2\} \\
&\overset{(c)}{\subseteq} \bigcup_{\theta\in\mathcal{T}} \{\bar{\boldsymbol{U}}_\theta(N) > u - u/2\} \cap \{\boldsymbol{C}_\theta(N) \leq u/2\} \\
&\overset{(d)}{\subseteq} \bigcup_{\theta\in\mathcal{T}} \{\bar{\boldsymbol{U}}_\theta(N) > u - u/2\} \cap \{\boldsymbol{Q}_\theta(N) \leq bu/2\},
\end{aligned} \tag{E.13}$$

where (a) follows from the definition of $\bar{\boldsymbol{U}}_\theta(N)$ in (E.9); (b) is obtained by retaining only the event corresponding to $\theta' = \theta$ in the intersection; (c) holds since in the intersection $-u/2 \leq -\boldsymbol{C}_\theta(N)$; and (d) follows by observing that, in view of (E.5), (E.6) and (E.7) we have $\boldsymbol{\chi}_\theta(n) \leq b\boldsymbol{\nu}_\theta(n)$, which in turn implies, from the definitions in (E.8) and (E.9), that $\boldsymbol{Q}_\theta(N) \leq b\,\boldsymbol{C}_\theta(N)$. Using (E.13) in the first term on the RHS of (E.12), and further applying the union bound, we obtain:

$$\begin{aligned}&\mathbb{P}\left[\max_{\theta\in\mathcal{T}} \boldsymbol{U}_\theta(N) > u,\ \max_{\theta\in\mathcal{T}} \boldsymbol{C}_\theta(N) \leq \frac{u}{2}\right] \\ &\leq \sum_{\theta\in\mathcal{T}} \mathbb{P}\left[\bar{\boldsymbol{U}}_\theta(N) > \frac{u}{2},\ \boldsymbol{Q}_\theta(N) \leq \frac{bu}{2}\right]. \end{aligned} \tag{E.14}$$

The sequence $\{\bar{\boldsymbol{U}}_\theta(N)\}_{N\geq 1}$ is a martingale by construction, since it is a sum of random variables (i.e., $\boldsymbol{u}_\theta(n)$) minus their conditional expectation (i.e., $\boldsymbol{\nu}_\theta(n)$). Moreover, from (E.5) we have the bound:

$$\bar{\boldsymbol{U}}_\theta(N+1) - \bar{\boldsymbol{U}}_\theta(N) \leq \boldsymbol{u}_\theta(N+1) \leq b. \tag{E.15}$$

Therefore, it can be readily checked that the scaled sequence $\{\bar{\boldsymbol{U}}_\theta(N)/b\}_{N\geq 1}$ meets the hypotheses of Theorem 2.1 in [38], and in particular the upper bound obtained by combining Eqs. (10), (11) and (15) in [38], which finally yields:

$$\mathbb{P}\left[\bar{\boldsymbol{U}}_\theta(N) > \frac{u}{2}, \boldsymbol{Q}_\theta(N) \leq \frac{bu}{2}\right] \leq e^{-\frac{3}{16b}u}, \tag{E.16}$$

and the proof is complete. ■

List of Figures

List of Tables

Bibliography

[1] J. Alcock, *Animal Behavior: An Evolutionary Approach,* Sunderland, MA, USA: Sinauer Associates, 2009.

[2] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion," *J. Mach. Learn. Res.*, vol. 13, pp. 2293–2337, Jan. 2012.

[3] A. Anandkumar and R. Valluvan, "Learning loopy graphical models with latent variables: Efficient methods and guarantees," *Ann. Statist.*, vol. 41, no. 2, pp. 401–435, Apr. 2013.

[4] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[5] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience: Exploring the Brain,* Philadelphia, PA, USA: Lippincott, Williams & Wilkins, 2006.

[6] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Vancouver, QC,Canada, Dec. 2010, pp. 172–180.

[7] B. Bollobás, *Random Graphs*, Cambridge, UK: Cambridge University Press, 2001.

[8] B. Bollobás and O. Riordan, "Mathematical results on scale-free random graphs," *Handbook of Graphs and Networks: from the Genome to the Internet*, Hoboken, UK: Wiley, pp. 1–34, 2003.

[9] B. Bollobás and O. Riordan, "Robustness and vulnerability of scale-free random graphs," *Internet Math.*, vol. 1, no. 1, pp. 1–35, 2004.

[10] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, "The degree sequence of a scale-free random graph process," *Random Structures & Algorithms*, vol. 18, no. 3, pp. 279–290, 2001.

[11] V. Bordignon, V. Matta, and A. H. Sayed, "Adaptive social learning," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 6053–6081, Jul. 2021.

[12] N. Bowler, J. Erde, P. Heinig, F. Lehner, and M. Pitz, "A counterexample to the reconstruction conjecture for locally finite trees," *Bull. London Math. Soc.*, vol. 49, no. 4, pp. 630–648, Aug. 2017.

[13] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.,* vol. 3, no. 1, pp. 1–122, 2010.

[14] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006

[15] G. Bresler, F. Koehler, A. Moitra, and E. Mossel, "Learning restricted Boltzmann machines via influence maximization," in *Proc. ACM Symp. Theory Comput. (STOC)*, Phoenix, AZ, USA, Jun. 2019.

[16] F. S. Cattivelli and A. H. Sayed, "Distributed detection over adaptive networks using diffusion adaptation," *IEEE Trans. Signal Process.,* vol. 59, no. 5, pp. 1917–1932, May 2011.

[17] E. Cesàro, "Sur la convergence des séries," *Nouvelles Annales de Mathématiques*, series 3, 7: pp. 49–59, 1888.

[18] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *Ann. Statist.*, vol. 40, no. 4, pp. 1935–1967, Aug. 2012.

[19] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks-Part I: Transient analysis," *IEEE Trans. Inf. Theory,* vol. 61, no. 6, pp. 3487–3517, Jun. 2015.

[20] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks-Part II: Performance analysis," *IEEE Trans. Inf. Theory,* vol. 61, no. 6, pp. 3518–3548, Jun. 2015.

[21] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačevíc, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.,* vol. 63, no. 24, pp. 6510–6523, Dec. 2015.

[22] Y. Chen, Z. Wang, and X. Shen, "An unbiased symmetric matrix estimator for topology inference under partial observability," *IEEE Signal Process. Lett.*, vol. 29, pp. 1257–1261, Sep. 2022.

[23] S. P. Chepuri and G. Leus, "Graph sampling for covariance estimation," *IEEE Trans. Signal Inf. Process. Netw.,* vol. 3, no. 3, pp. 451–466, Sep. 2017.

[24] E. S. C. Ching and H. C. Tam, "Reconstructing links in directed networks from noisy dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, no. 1, pp. 010301-1–010301-5, Jan. 2017.

[25] M. Cirillo, V. Matta, and A. H. Sayed, "Learning Bollobás-Riordan graphs under partial observability," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 5360–5364.

[26] M. Cirillo, V. Matta, and A. H. Sayed, "Estimating the topology of preferential attachment graphs under partial observability," *IEEE Trans. Inf. Theory*, vol. 69, no. 2, pp. 1355–1380, Feb. 2022.

[27] M. Cirillo, V. Matta, and A. H. Sayed, "Learning dynamic graphs under partial observability," submitted to the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023.

[28] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.

[29] I. D. Couzin, "Collective cognition in animal groups," *Trends Cognit. Sci.*, vol. 13, no. 1, pp. 36–43, Jan. 2009.

[30] M. H. DeGroot, "Reaching a consensus," *J. Amer. Statist. Assoc.*, vol. 69, no. 345, pp. 118–121, 1974.

[31] J. A. Deri and J. M. F. Moura, "New York city taxi analysis with graph signal processing," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP),* Washington, DC, USA, Dec. 2016, pp. 1275–1279.

[32] R. Diestel, *Graph Theory*, Heidelberg, Germany: Springer-Verlag, 2005.

[33] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[34] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, Sep. 2017.

[35] P. Erdős and A. Rényi, "On random graphs I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[36] J. Etesami and N. Kiyavash, "Measuring causal relationships in dynamical systems through recovery of functional dependencies," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 4, pp. 650–659, Dec. 2017.

[37] J. Etesami, N. Kiyavash, and T. Coleman, "Learning minimal latent directed information polytrees," *Neural Comput.*, vol. 28, no. 9, pp. 1723–1768, Aug. 2016.

[38] X. Fan, I. Grama, and Q. Liu, "Hoeffding's inequality for supermartingales," *Stoch. Process. Their Appl.*, vol. 122, no. 10, pp. 3545–3559, 2012.

[39] N. Friedman, L. Getoor, D. Koller, A. Pfeffer, "Learning probabilistic relational models," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, Aug. 1999, pp. 1300–1309.

[40] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira, "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC Syst. Biol.,* vol. 1, no. 1, p. 39, 2007.

[41] A. Ganesh, L. Massoulié, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. IEEE INFOCOM,* vol. 2, Mar. 2005, pp. 1455–1466.

[42] P. Geiger, K. Zhang, B. Schölkopf, M. Gong, and D. Janzing, "Causal inference by identification of vector autoregressive processes with hidden components," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, vol. 37, pp. 1917–1925.

[43] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* New York, USA: Springer Science+Business Media, 2001.

[44] G., B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. IEEE*, vol. 106, no. 5, pp. 787–807, May 2018.

[45] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica,* vol. 37, no. 3, pp. 424–438, Aug. 1969.

[46] M. C. Grant and S. P. Boyd, *Graph Implementations for Nonsmooth Convex Programs,* London, UK: Springer-Verlag Limited, 2008, pp. 95–110.

[47] M. C. Grant and S. P. Boyd, "CVX: Matlab software for disciplined convex programming," *available at* http://cvxr.com/cvx.

[48] J. D. Hamilton, *Time Series Analysis,* Princeton, NJ, USA: Princeton Univ. Press, 1994.

[49] F. Han, H. Lu, and H. Liu, "A direct estimation of high dimensional stationary vector autoregressions," *J. Mach. Learn. Res.*, vol. 16, pp. 3115–3150, Dec. 2015.

[50] A. Natali, M. Coutino, E. Isufi, and G. Leus, "Online time-varying topology identification via prediction-correction algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 5400–5404.

[51] A. Jalali and S. Sanghavi, "Learning the dependence graph of time series with latent factors," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Scotland, UK, Jun. 2012, pp. 619–626.

[52] M. Jawed, M. Kaya and R. Alhajj, "Time frame based link prediction in directed citation networks," in *Proc. of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Paris, France, Aug. 2015, pp. 1162–1168.

[53] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. Intl. Conf. Artificial Intell. Statistics*, Cadiz, Spain, May 2016, pp. 920–929.

[54] D. Karger and N. Srebro, "Learning Markov networks: Maximum bounded tree-width graphs," in *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Washington DC, USA, Jan. 2001, pp. 392–401.

[55] U. A. Khan, W. U. Bajwa, A. Nedić, M. G. Rabbat, and A. H. Sayed, *Editors*, "Optimization for Data-Driven Learning and Control," *Proc. IEEE*, vol. 108, no. 11, pp. 1863–1868, Nov. 2020.

[56] G. Kossinets, "Effects of missing data in social networks," *Soc. Netw.*, vol. 28, no. 3, pp. 247–268, Jul. 2006.

[57] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, M. A. Calderwood, M. Vidal, and A. L. Barabási, "Network-based prediction of protein interactions," *Nat. Commun.*, vol. 10, no. 1, pp. 1–8, Mar. 2019.

[58] P.-Y. Lai, "Reconstructing network topology and coupling strengths in directed networks of discrete-time dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, no. 2, pp. 022311-1–022311-13, Feb. 2017.

[59] S. Lee and A. Nedić, "Distributed random projection algorithm for convex optimization," *IEEE J. Sel. Topics Signal Process.,* vol. 7, no. 2, pp. 221–229, Apr. 2013.

[60] R. Liégeois, A. Santos, V. Matta, D. Van de Ville, and A. H. Sayed, "Revisiting correlation-based functional connectivity and its relationship with structural connectivity," *Network Neuroscience*, vol. 4, no. 4, pp. 1235–1251, 2020.

[61] P. L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity," *Ann. Statist.*, vol. 40, no. 3, pp. 1637–1664, Apr. 2012.

[62] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A: Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011.

[63] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*, Berlin, Germany: Springer, 2005.

[64] S. Mahdizadehaghdam, H. Wang, H. Krim, and L. Dai, "Information diffusion of topic propagation in social media," *IEEE Trans. Signal Inf. Process. Netw.,* vol. 2, no. 4, pp. 569–581, Dec. 2016.

[65] B. Manvel, "On reconstructing graphs from their sets of subgraphs," *J. Comb. Theory, Ser. B*, vol. 21, no. 2, pp. 156–165, Oct. 1976.

[66] S. Marano, V. Matta, T. He, and L. Tong, "The embedding capacity of information flows under renewal traffic," *IEEE Trans. Inf. Theory,* vol. 59, no. 3, pp. 1724–1739, Mar. 2013.

[67] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.,* vol. 65, no. 22, pp. 5911–5926, Nov. 2017.

[68] G. Mateos, S. Segarra, A. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019.

[69] D. Materassi and M. V. Salapaka, "On the problem of reconstructing an unknown topology via locality properties of the Wiener filter," *IEEE Trans. Autom. Control*, vol. 57, no. 7, pp. 1765–1777, Jul. 2012.

[70] D. Materassi and M. V. Salapaka, "Network reconstruction of dynamical polytrees with unobserved nodes," in *Proc. IEEE Conf. Decis. Control (CDC)*, Maui, HI, USA, Dec. 2012, pp. 4629–4634.

[71] D. Materassi and M. V. Salapaka, "Identification of network components in presence of unobserved nodes," in *Proc. IEEE Conf. Decis. Control (CDC)*, Osaka, Japan, Dec. 2015, pp. 1563–1568.

[72] V. Matta, V. Bordignon, A. Santos, and A. H. Sayed, "Interplay between topology and social learning over weak graphs," *IEEE Open J. Signal Process.*, vol. 1, pp. 99–119, 2020.

[73] V. Matta, A. Santos, and A. H. Sayed, "Tomography of large adaptive networks under the dense latent regime," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Oct. 2018, pp. 2144–2148.

[74] V. Matta, A. Santos, and A. H. Sayed, "Graph learning with partial observations: Role of degree concentration," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1312–1316.

[75] V. Matta, A. Santos, and A. H. Sayed, "Graph learning over partially observed diffusion networks: Role of degree concentration," *IEEE Open J. Signal Process.*, vol. 3, pp. 335–371, Jul. 2022.

[76] V. Matta and A. H. Sayed, "Estimation and detection over adaptive networks," in *Cooperative and Graph Signal Processing,* P. Djuric and C. Richard, Eds. Amsterdam, The Netherlands: Elsevier, 2018, pp. 69–106.

[77] V. Matta and A. H. Sayed, "Consistent tomography under partial observations over adaptive networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 622–646, Jan. 2019.

[78] V. Matta, A. Santos, and A. H. Sayed, "Graph learning under partial observability," *Proc. IEEE*, vol. 108, no. 11, pp. 2049–2066, Nov. 2020.

[79] A. Mauroy and J. Goncalves, "Linear identification of nonlinear systems: A lifting technique based on the Koopman operator," in *Proc. IEEE Conf. Decis. Control (CDC)*, Las Vegas, NV, USA, Dec. 2016, pp. 6500–6505.

[80] J. Mei and J. Moura, "Signal processing on graphs: Causal modeling of *un*structured data," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 2077–2092, Apr. 2017.

[81] A. Moneta, N. Chlaß, D. Entner, and P. Hoyer, "Causal search in structural vector autoregressive models," in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec. 2009, pp. 95–118.

[82] D. Napoletani and T. D. Sauer, "Reconstructing the topology of sparsely connected dynamical networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 77, no. 2, pp. 026103-1–026103-5, Feb. 2008.

[83] R. Nassif, S. Vlaski, C. Richard, J. Chen, and A. H. Sayed, "Multitask learning over graphs: An approach for distributed, streaming machine learning," *IEEE Signal Process. Mag.,* vol. 37, no. 3, pp. 14–25, May 2020.

[84] A. Nedić and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.,* vol. 12, no. 1, pp. 109–138, Jan. 2001.

[85] A. Nedić and A. Ozdaglar, "Cooperative distributed multi-agent optimization," in *Convex Optimization in Signal Processing and Communications,* Y. Eldar and D. Palomar, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010, pp. 340–386.

[86] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, vol. 64, no. 2, Apr. 2001.

[87] M. Nokleby and W. U. Bajwa, "Stochastic optimization from distributed streaming data in rate-limited networks," *IEEE Trans. Signal Inf. Process. Netw.,* vol. 5, no. 1, pp. 152–167, Mar. 2019.

[88] A. Özcan and Ş. G. Öğüdücü, "Supervised temporal link prediction using time series of similarity measures," in *Proc. of the International Conference on Ubiquitous and Future Networks (ICUFN)*, Milan, Italy, Jul. 2017, pp. 519–521.

[89] B. L. Partridge, "The structure and function of fish schools," *Sci. Amer.*, vol. 246, no. 6, pp. 114–123, Jan. 1982.

[90] B. Pasdeloup, V. Gripon, G. Mercier, D. Pastor, and M. G. Rabbat, "Characterization and inference of graph diffusion processes from observations of stationary signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 3, pp. 481–496, Sep. 2018.

[91] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.,* vol. 65, no. 13, pp. 3462–3477, Jul. 2017.

[92] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Phys. Rev. Lett.,* vol. 109, no. 6, pp. 068702-1–068702-5, Aug. 2012.

[93] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6887–6909, Dec. 2015.

[94] M. G. Rabbat and A. Ribeiro, "Multiagent distributed optimization," in *Cooperative and Graph Signal Processing,* P. Djuric and C. Richard, Eds. Amsterdam, The Netherlands: Elsevier, 2018, pp. 147–167.

[95] M. Rao, A. Kipnis, M. Javidi, Y. Eldar, and A. Goldsmith, "System identification with partial samples: Non-asymptotic analysis," in *Proc. IEEE Conf. Decis. Control (CDC)*, Las Vegas, NV, USA, Dec. 2016, pp. 2938–2944.

[96] J. Ren, W.-X. Wang, B. Li, and Y.-C. Lai, "Noise bridges dynamical correlation and topology in coupled oscillator networks," *Phys. Rev. Lett.*, vol. 104, no. 5, pp. 058701-1–058701-4, Feb. 2010.

[97] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, Austin, TX, USA, Jan. 2015, pp. 4292–4293.

[98] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[99] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Process.,* vol. 62, no. 12, pp. 3042–3054, Jun. 2014.

[100] A. Santos, V. Matta, and A. H. Sayed, "Local tomography of large networks under the low-observability regime," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 587–613, Jan. 2020.

[101] A. H. Sayed, *Inference and Learning from Data*, Volume 3: Learning, Cambridge University Press, 2023.

[102] A. H. Sayed, "Adaptive networks," *Proc. of the IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[103] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Found. Trends Mach. Learn.*, vol. 7, no. 4-5, pp. 311–801, 2014.

[104] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks," *IEEE Signal Process. Mag.,* vol. 30, no. 3, pp. 155–171, May 2013.

[105] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, Sep. 2017.

[106] S. Segarra, M. T. Schaub, and A. Jadbabaie, "Network inference from consensus dynamics," in *Proc. IEEE Conference on Decision and Control (CDC)*, Dec. 2017, pp. 3212–3217.

[107] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.,* vol. 30, no. 3, pp. 83–98, May 2013.

[108] V. Shumovskaia, K. Ntemos, S. Vlaski, and A. H. Sayed, "Online graph learning from social interactions," in *Proc. Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, Pacific Grove, CA, USA, Oct.–Nov. 2021, pp. 1263–1267.

[109] O. Stolz, *Vorlesungen über allgemeine Arithmetik: nach den neueren Ansichten.* Leipzig, Germany: Teubner, 1885.

[110] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control,* vol. 31, no. 9, pp. 803–812, Sep. 1986.

[111] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *IEEE Trans. Signal Process.,* vol. 64, no. 18, pp. 4845–4860, Sep. 2016.

[112] R. Van Der Hofstad, *Random Graphs and Complex Networks*, Berlin, Germany: Springer, 2016.

[113] P. Venkitasubramaniam, T. He, and L. Tong, "Anonymous networking amidst eavesdroppers," *IEEE Trans. Inf. Theory,* vol. 54, no. 6, pp. 2770–2784, Jun. 2008.

[114] S. Vlaski, H. P. Maretić, R. Nassif, P. Frossard, and A. H. Sayed, "Online graph learning from sequential data," in *Proc. IEEE Data Science Workshop,* Jun. 2018, pp. 190–194.

[115] S. Vlaski, A. H. Sayed, "Distributed learning in non-convex environments—Part I: Agreement at a linear rate," *IEEE Trans. Signal Process,* vol. 69, pp. 1242–1256, 2021.

[116] S. Vlaski, A. H. Sayed, "Distributed learning in non-convex environments—Part II: Polynomial escape from saddle-points," *IEEE Trans. Signal Process,* vol. 69, pp. 1257–1270, 2021.

[117] S. Vlaski, A. H. Sayed, "Second-order guarantees of stochastic gradient descent in non-convex optimization," *IEEE Trans. Autom. Control,* vol. 67, no. 12, pp. 6489–6504 Dec. 2022.

[118] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *IEEE Trans. Autom. Control,* vol. 62, no. 8, pp. 3986–3992, Aug. 2017.

[119] C. Xi, V. S. Mai, R. Xin, E. H. Abed, and U. A. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Trans. Autom. Control,* vol. 63, no. 10, pp. 3558–3565, Oct. 2018.

[120] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53, no. 1, pp. 65–78, Sep. 2004.

[121] Y. Yang, T. Luo, Z. Li, X. Zhang, and P. S. Yu, "A robust method for inferring network structures," *Sci. Rep.*, vol. 7, no. 5221, pp. 1–12, Jul. 2017.

[122] K. Yu, W. Chu, S. Yu, V. Tresp, Z. Xu, "Stochastic relational models for discriminative link prediction," in *Proc. of the International Conference on Neural Information Precessing Systems (NIPS)*, Vancouver, Canada, Dec. 2006, pp. 1553–1560.

[123] M. Zhang, Y. Chen, "Link prediction based on graph neural networks," in *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*, Montréal, Canada, Dec. 2018, pp. 5171–5181.