# Graph Learning with Partial Observations: Role of Degree Concentration

Vincenzo Matta
DIEM, University of Salerno
via Giovanni Paolo II, I-84084, Fisciano
Email: vmatta@unisa.it

Augusto Santos and Ali H. Sayed
EPFL, School of Engineering
CH-1015 Lausanne, Switzerland
Email: {augusto.santos, ali.sayed}@epfl.ch

***Abstract*—In this work we consider the problem of learning an Erdős-Rényi graph over a diffusion network when: $i$) data from only a limited subset of nodes are available (*partial observation*); $ii$) and the inferential goal is to discover the graph of interconnections linking the accessible nodes (*local structure learning*). We propose three matrix estimators, namely, the Granger, the one-lag correlation, and the residual estimators, which, when followed by a *universal clustering* algorithm, are shown to retrieve the true subgraph in the limit of large network sizes. Remarkably, it is seen that a fundamental role is played by the *uniform concentration* of node degrees, rather than by sparsity.**

## I. Introduction

The evolutionary dynamics of complex networks is determined by local interactions among neighboring agents. Unveiling the topology of these interactions is one fundamental goal of graph learning [1]–[4]. Graph learning plays a prominent role in many domains including signal processing over graphs [5], [6], social networks [7], and brain connectivity [8].

In this article, we focus on the graph learning problem when observations can be collected only from a subset of the nodes (*partial observation*) and examine different regimes of network connectivity including the often overlooked case of *dense* networks. Under these demanding conditions, we establish that, under certain assumptions on the entries of the combination matrix and the network topology, the problem of graph learning becomes *localized*. In other words, the graph connecting the monitored nodes can be consistently retrieved, in the limit of large network sizes, from knowledge contained solely in the information from samples of the observed agents. Notably, this result holds irrespective of the sparsity of connections. One fundamental conclusion is that the main element enabling consistent local tomography is the node degree concentration and not sparsity.

### A. Networked Dynamical System

Given a network of connected agents, streaming observations from a monitored subset of nodes are gathered. The objective is to discover the graph of interconnections within this subnetwork. The learning process consists of two stages: an estimation stage, where a matrix quantifying the strength of connections is estimated; and a thresholding stage, where node pairs whose strength weights stay above some threshold are classified as connected. A *structurally-consistent* estimator would assign strong weights to connected pairs and weak weights to disconnected pairs. As a result, at the thresholding stage one would be able to retrieve the subgraph correctly.

One relevant dynamical model for graph learning is the diffusion or first-order Vector Autoregressive (VAR) system:

$$\boxed{\boldsymbol{y}_n = A\,\boldsymbol{y}_{n-1} + \boldsymbol{x}_n} \tag{1}$$

where $A$ is some stable $N\times N$ matrix with nonnegative entries, and the random vectors $\boldsymbol{x}_n = [\boldsymbol{x}_1(n), \boldsymbol{x}_2(n), \ldots, \boldsymbol{x}_N(n)]^\top$ and $\boldsymbol{y}_n = [\boldsymbol{y}_1(n), \boldsymbol{y}_2(n), \ldots, \boldsymbol{y}_N(n)]^\top$ collect respectively the the input (e.g., streaming data or noise) and output from all nodes at time $n$. The input variables $\boldsymbol{x}_i(n)$ have zero mean and unit variance, and are independent and identically distributed (i.i.d.), both spatially (i.e., w.r.t. to index $i$) and temporally (i.e., w.r.t. to index $n$). From (1) we readily see that, at time $n$, the output of node $i$ is updated by *combining* the outputs of other nodes from time $n-1$. In particular, node $i$ scales the output of node $\ell$ by using a combination weight $a_{i\ell}$, which implies that the output of agent $\ell$ is *effectively used* by node $i$ if, and only if $a_{i\ell} \neq 0$. After the combination step, the output measurement $\boldsymbol{y}_i(n)$ is adjusted by incorporating the streaming-source value, $\boldsymbol{x}_i(n)$, which is locally available at node $i$ at current time $n$.

### B. Related Work

Most works on graph learning focus on linear dynamical systems, and in particular on autoregressive diffusion models [9], [10]. Some recent works focus on reducing complexity by exploiting structural constraints (e.g., smoothness or sparsity of the signals defined on the graph) [10]. However, these results consider the case in which measurements from the entire network are available. We focus instead on the case in which only *partial observation of the network is permitted*. This learning problem is unfeasible or NP-hard in general, and one fundamental challenge is to find nontrivial graph classes where it can be affordable [11]. There are results holding for specific network graphs (polytrees) [12], or results holding for more general topologies, which are however impractical over large networks since they require precise details of the topology and of the statistical models [9], [13]. In contrast, the thermodynamic regime of large networks is addressed here by exploiting *random* graphs, with conditions on the network

The work of A. H. Sayed was supported in part by NSF grant CCF-1524250.

connectivity being summarized through average indicators such as the probability of drawing an edge. Similar approaches are exploited in the context of graphical models with latent variables [4], [14], which, however, do not match the dynamic graph models considered in this work. For the latter models, recent results are available in [2], [3]. In particular, we focus on the full Erdős-Rényi model addressed in [2], and, hence, we now explain briefly the key contributions of the present work in relation to [2]. Preliminarily, we observe that we adopt a slightly more restrictive condition on the class of combination matrices, summarized in Assumption 1 further ahead. We remark that such assumption is automatically satisfied by the matrices considered in [2], as they arise naturally when the combination weights are collected into a (scaled version of a) symmetric doubly-stochastic matrix whose support graph matches the underlying graph of connections.

The first advance regards the regime of connectivity, because here we deal also with the *dense* regime that is not addressed in [2]. For what concerns the sparse regime, the consistency result in [2] is formulated in terms of the *average* fraction of misclassified node pairs, whereas we are able to establish here the stronger conclusion that the total number (not only the fraction) of misclassified node pairs goes to zero. We make progress also in relation to the learning algorithms that we propose since, along with the Granger estimator examined in [2], we introduce two other estimators: the *one-lag correlation matrix* and the *correlation matrix between the residuals* (i.e., difference between subsequent time samples). It will be seen that all three estimators are structurally consistent, with some remarkable findings in terms of their relative performance. The analysis will reveal that, in contrast to widespread belief, the main enabling feature for consistent graph learning is the *node degree concentration*, and not graph connection sparsity. Preliminary results for the dense regime, and for the Granger estimator, have been reported in [15].

*Notation*. Boldface letters denote random variables, normal font letters their realizations. Capital letters are used for matrices, small letters for vectors. For a matrix $Z$, the submatrix spanning the rows and columns of $Z$ indexed by a subset of indices $\mathcal{S}$ is denoted by $Z_{\mathcal{S}}$ or alternatively by $[Z]_{\mathcal{S}}$.

## II. Thermodynamic Limit of Erdős-Rényi Graphs

We assume that the entire network graph is generated according to the Erdős-Rényi model, where edge drawing obeys a sequence of Bernoulli experiments with identical success (i.e., connection) probability [16], [17]. The partial observability constraint is formalized by saying that the subnetwork of observable measurements, $\mathcal{S}$, has a cardinality $S$ scaling as $(S/N) \overset{N\to\infty}{\longrightarrow} \xi \in (0,1)$, where $\xi$ is the (asymptotic) fraction of monitored nodes.

We now introduce two fundamental graph descriptors. The first descriptor is the adjacency matrix, denoted by $\boldsymbol{G}$, with $\boldsymbol{g}_{ij} = 1$ if nodes $i$ and $j$ are connected, and $\boldsymbol{g}_{ij} = 0$ otherwise. The bold notation highlights that we deal with *random* graphs.

The second descriptor is the *degree*, namely, the cardinality of the node neighborhood (which conventionally includes the node itself). The degree of node $i$ can be represented as $\boldsymbol{d}_i = 1 + \sum_{\ell\neq i} \boldsymbol{g}_{i\ell}$. The minimal and maximal degrees will be denoted by $\boldsymbol{d}_{\min}$ and $\boldsymbol{d}_{\max}$, respectively.

Over large networks, it is important to characterize the average behavior that emerges with high probability in the *thermodynamic limit* as the network size goes to infinity. To this end, the traditional framework of random graph evolution allows the connection probability $p_N$ to scale with $N$. For instance, with a constant $p_N$, the number of neighbors grows linearly with $N$, while a $p_N$ scaling as $(\ln N)/N$ would correspond to a number of neighbors growing logarithmically with $N$. It is useful for our purposes to list briefly the main regimes that are of interest for the forthcoming treatment.

In this work we focus on the regime where the graph is connected with high probability, which is characterized by a connection probability $p_N = (\ln N + c_N)/N$, where $c_N$ is any positive sequence diverging to infinity as $N \to \infty$ [16], [17]. We notice that the concept of connectedness does not specify whether the connection probability vanishes or not as $N$ gets large. In particular, a random graph is connected when $p_N \overset{N\to\infty}{\longrightarrow} p > 0$ (*densely* connected regime), but can be connected even when $p_N$ vanishes (*sparsely* connected regime). The concept of sparsity has been exploited in previous works about topology inference under partial observations. Useful structural consistency results have been proved for our setting under the sparsely connected regime [2], [3].

One fundamental novelty of the present work is characterizing structural consistency by exploiting another feature, namely, node degree *concentration*. We remark that the term "concentration" *does not refer to the number of connections*. This concept is instead often used in statistics to refer to statistical quantities that concentrate around some typical (deterministic) values as $N$ gets large [18]. In particular, in our case it is meaningful to consider the *uniform concentration properties of the minimal and maximal degrees of random graphs*. The uniform concentration regime is attained by choosing the following pairwise connection probability:

$$\boxed{p_N = \omega_N \frac{\ln N}{N} \overset{N\to\infty}{\longrightarrow} p \geq 0} \tag{2}$$

where $\omega_N$ is a positive sequence diverging to $+\infty$ with $N$. We remark that the case of a constant connection probability (i.e., $p_N = p > 0$) is a special case of (2). Under (2), both the minimal and maximal degrees of the graph *concentrate around the expected degree* in the following sense:

$$\boxed{\frac{\boldsymbol{d}_{\min}}{Np_N} \xrightarrow{\text{p}} 1, \quad \frac{\boldsymbol{d}_{\max}}{Np_N} \xrightarrow{\text{p}} 1, \quad \text{[Uniform concentration]}} \tag{3}$$

where $\xrightarrow{\text{p}}$ denotes convergence in probability as $N \to \infty$. In terms of graph properties, Eq. (3) reveals that $\boldsymbol{d}_{\min} \sim Np_N + f_N$ and $\boldsymbol{d}_{\max} \sim Np_N + g_N$, where $f_N$ and $g_N$ are sequences that are asymptotically dominated by $Np_N$. According to (2), we see that the regime of concentration can be either sparse or dense, depending on whether the limiting connection probability, $p$, is zero or positive, respectively.

## III. Main Results

*Assumption 1 (Diffusion matrices):* The combination matrix $\boldsymbol{A}$ is symmetric and its entries fulfill

$$\boxed{\sum_{\ell=1}^{N} \boldsymbol{a}_{i\ell} = \rho, \qquad \frac{\kappa}{\boldsymbol{d}_{\max}}\, \boldsymbol{g}_{ij} \le \boldsymbol{a}_{ij} \le \frac{\kappa}{\boldsymbol{d}_{\min}}\, \boldsymbol{g}_{ij} \quad \forall i \neq j} \tag{4}$$

for some $0 < \rho < 1$ and $0 < \kappa \le \rho$. □

The most common combination matrices used in the literature satisfy Assumption 1 automatically. Some popular choices are the Laplacian and the Metropolis rules — see [2] — which arise naturally in many applications, for instance, they find widespread application over *adaptive* networks [5].

We notice that the second relationship in (4) implies assigning positive weights to neighboring agents, whereas the first relationship sets the self-weights as $\boldsymbol{a}_{ii} = \rho - \sum_{\ell \neq i} \boldsymbol{a}_{i\ell}$.

Next we present the estimators proposed in this work. We introduce preliminarily the steady-state correlation matrix $\boldsymbol{R}_0 = \lim_{n\to\infty} \mathbb{E}\left[\boldsymbol{y}_n \boldsymbol{y}_n^{\top}\right]$, whose existence is guaranteed by the stability of $\boldsymbol{A}$. Exploiting the discrete-time Lyapunov equation and the symmetry of $\boldsymbol{A}$, it is possible to show that $\boldsymbol{R}_0 = (I_N - \boldsymbol{A}^2)^{-1}$, where $I_N$ is the $N \times N$ identity matrix. Likewise, we introduce the steady-state one-lag correlation matrix, $\boldsymbol{R}_1 = \lim_{n\to\infty} \mathbb{E}\left[\boldsymbol{y}_n \boldsymbol{y}_{n-1}^{\top}\right]$, which exploiting the dynamics in (1) can be written as $\boldsymbol{R}_1 = \boldsymbol{A}\boldsymbol{R}_0$. Accordingly, we have the inverse relationship $\boldsymbol{A} = \boldsymbol{R}_1 \boldsymbol{R}_0^{-1}$, a quantity that is also referred to as the Granger estimator [9].

Under the limited observation setting, one natural choice is to compute the Granger estimator only in the subnet $\mathcal{S}$, namely,

$$\widehat{\boldsymbol{A}}_{\mathcal{S}}^{(\text{Gra})} = [\boldsymbol{R}_1]_{\mathcal{S}} ([\boldsymbol{R}_0]_{\mathcal{S}})^{-1} \quad \text{[Granger estimator]} \tag{5}$$

Our second strategy relies on using the one-lag correlation matrix as estimator for the combination matrix:

$$\widehat{\boldsymbol{A}}_{\mathcal{S}}^{(\text{1-lag})} = [\boldsymbol{R}_1]_{\mathcal{S}} \quad \text{[one-lag estimator]} \tag{6}$$

The reason behind such choice is the following series expansion of the one-lag correlation matrix:

$$\boldsymbol{R}_1 = \boldsymbol{A}\boldsymbol{R}_0 = \boldsymbol{A}(I_N - \boldsymbol{A}^2)^{-1} = \boldsymbol{A} + \boldsymbol{A}^3 + \boldsymbol{A}^5 + \ldots, \tag{7}$$

which implies that $\widehat{\boldsymbol{A}}_{\mathcal{S}}^{(\text{1-lag})} = \boldsymbol{A}_{\mathcal{S}}$ plus an error matrix involving the odd powers of matrix $\boldsymbol{A}$.

Finally we introduce the (scaled) residual vector $\boldsymbol{r}_n \triangleq (\boldsymbol{y}_n - \boldsymbol{y}_{n-1})/\sqrt{2}$. Since we have $\mathbb{E}[\boldsymbol{r}_n \boldsymbol{r}_n^{\top}] = \boldsymbol{R}_0 - \boldsymbol{R}_1 = I_N - \boldsymbol{A} + \boldsymbol{A}^2 - \boldsymbol{A}^3 + \ldots$, it makes sense to introduce the estimator:

$$\widehat{\boldsymbol{A}}_{\mathcal{S}}^{(\text{res})} = [\boldsymbol{R}_1]_{\mathcal{S}} - [\boldsymbol{R}_0]_{\mathcal{S}} \quad \text{[residual estimator]} \tag{8}$$

Also in this case we can write $\widehat{\boldsymbol{A}}_{\mathcal{S}}^{(\text{res})} = \boldsymbol{A}_{\mathcal{S}}$ plus an error, which now involves an alternating series of powers of $\boldsymbol{A}$.

In order to ascertain whether or not it is possible to discriminate connected/disconnected agents via observation of their output measurements, we introduce the concept of margins and identifiability gap.

*Definition 1 (Margins):* Let $\widehat{\boldsymbol{A}}_{\mathcal{S}}$ be a certain estimated combination matrix, corresponding to the subset $\mathcal{S}$. The lower and upper margins corresponding to the *disconnected* pairs are defined as, respectively:

$$\underline{\boldsymbol{\delta}}_N \triangleq \min_{\substack{i,j\in\mathcal{S}:\boldsymbol{a}_{ij}=0 \\ i\neq j}} \widehat{\boldsymbol{a}}_{ij}, \quad \overline{\boldsymbol{\delta}}_N \triangleq \max_{\substack{i,j\in\mathcal{S}:\boldsymbol{a}_{ij}=0 \\ i\neq j}} \widehat{\boldsymbol{a}}_{ij}. \tag{9}$$

Likewise, the lower and upper margins corresponding to the *connected* pairs are defined as, respectively:

$$\underline{\boldsymbol{\Delta}}_N \triangleq \min_{\substack{i,j\in\mathcal{S}:\boldsymbol{a}_{ij}>0 \\ i\neq j}} \widehat{\boldsymbol{a}}_{ij}, \quad \overline{\boldsymbol{\Delta}}_N \triangleq \max_{\substack{i,j\in\mathcal{S}:\boldsymbol{a}_{ij}>0 \\ i\neq j}} \widehat{\boldsymbol{a}}_{ij}. \tag{10}$$

□

The aforementioned margins are useful to examine the consistency for an estimator $\widehat{\boldsymbol{A}}_{\mathcal{S}}$.

*Definition 2 (Universal local structural consistency):* Let $\widehat{\boldsymbol{A}}_{\mathcal{S}}$ be an estimated combination matrix. If there exist $s_N > 0$, $\eta \in \mathbb{R}$, and $\Gamma > 0$, such that:

$$\boxed{\begin{array}{ll} s_N\, \underline{\boldsymbol{\delta}}_N \xrightarrow{\text{p}} \eta, & s_N\, \underline{\boldsymbol{\Delta}}_N \xrightarrow{\text{p}} \eta + \Gamma \\ s_N\, \overline{\boldsymbol{\delta}}_N \xrightarrow{\text{p}} \eta, & s_N\, \overline{\boldsymbol{\Delta}}_N \xrightarrow{\text{p}} \eta + \Gamma \end{array}} \tag{11}$$

we say that the estimated matrix $\widehat{\boldsymbol{A}}_{\mathcal{S}}$ achieves universal local structural consistency, with a bias $\eta$, an identifiability gap $\Gamma$, and with a scaling sequence $s_N$. □

*Remark 1 (Identifiability gap):* We see from (11) that $i)$ the *maximum* entry of $s_N \widehat{\boldsymbol{A}}_{\mathcal{S}}$ over the *disconnected* pairs converges to $\eta$; while $ii)$ the *minimum* entry of $s_N \widehat{\boldsymbol{A}}_{\mathcal{S}}$ over the *connected* pairs converges to $\eta + \Gamma$. Therefore, the estimated matrix entries corresponding to connected pairs stand well separated from the entries corresponding to disconnected pairs, and the separation is quantified by the gap $\Gamma$. □

*Remark 2 (Bias):* For the *true* combination matrix, the entries corresponding to disconnected pairs are zero. In contrast, in view of (11), the scaled entries for disconnected pairs are asymptotically close to $\eta$, giving rise to a *bias*. Remarkably, *this bias does not constitute a problem for consistent classification of connected/non-connected nodes.* □

*Remark 3 (Universality):* According to (11), *all entries of $s_N \widehat{\boldsymbol{A}}_{\mathcal{S}}$ corresponding to disconnected pairs* are sandwiched between the pair of (scaled) lower margins, $s_N\, \underline{\boldsymbol{\delta}}_N$ and $s_N\, \overline{\boldsymbol{\delta}}_N$, which both converge to $\eta$. A similar behavior is observed for the scaled entries over the connected pairs, which converge altogether to $\eta + \Gamma$. Thus, the connected and disconnected agent pairs *cluster into well-separated classes* that can be reliably identified by means of a *universal* clustering algorithm. This is a remarkable conclusion because the threshold to separate the classes can be determined in a *fully data-driven* fashion. □

*Remark 4 (Locality):* The qualification "local" is used because reconstruction of the monitored subnetwork is asymptotically achievable through the information contained only in the nodes of that subnetwork. □

*Theorem 1 (Universal local structural consistency under uniform concentration):* Let $\boldsymbol{A}$ fulfill Assumption 1, let the network graph follow an Erdős-Rényi model with $N$ nodes and connection probability $p_N$, and let the fraction of observable nodes be $\xi > 0$. Under the concentration regime in (2), the Granger, the one-lag, and the residual estimators achieve

| Estimator | Error bias $\eta$ | Identifiability gap $\Gamma$ |
|---|---|---|
| Granger | $\kappa^2 p \, \dfrac{(2\rho-\kappa)\,(1-\xi)}{1-(\rho^2-2\rho\kappa\xi+\kappa^2\xi)}$ | $\kappa$ |
| one-lag | $\kappa^2 p \, \dfrac{\rho+\rho\,\zeta^2+2\,\zeta}{(1-\rho^2)(1-\zeta^2)^2}$ | $\dfrac{1+\zeta^2}{(1-\zeta^2)^2}\times\kappa$ |
| residual | $-\dfrac{\kappa^2 p}{(1+\rho)(1+\zeta)^2}$ | $\dfrac{\kappa}{(1+\zeta)^2}$ |

TABLE I
BIASES AND GAPS IN THEOREM 1. WE DEFINE $\zeta=\rho-\kappa$.

| Random variable | Random variable | Limit |
|---|---|---|
| $\mathfrak{M}_a \triangleq \max\limits_{\substack{i,j\in\mathcal{N}\\ i\neq j}} \boldsymbol{a}_{ij}$ | $\mathfrak{m}_a \triangleq \min\limits_{\substack{i,j\in\mathcal{N}\\ i\neq j}} \boldsymbol{a}_{ij}$ | $0$ |
| $\mathfrak{M}_{a,\text{self}} \triangleq \max\limits_{i\in\mathcal{N}} \boldsymbol{a}_{ii}$ | $\mathfrak{m}_{a,\text{self}} \triangleq \min\limits_{i\in\mathcal{N}} \boldsymbol{a}_{ii}$ | $\rho-\kappa$ |
| $\mathfrak{M}_{a_2,\text{self}} \triangleq \max\limits_{i\in\mathcal{N}} \boldsymbol{a}^{(2)}_{ii}$ | $\mathfrak{m}_{a_2,\text{self}} \triangleq \min\limits_{i\in\mathcal{N}} \boldsymbol{a}^{(2)}_{ii}$ | $(\rho-\kappa)^2$ |
| $\mathfrak{M}_{a,\text{sum}} \triangleq \max\limits_{\substack{i,j\in\mathcal{N}\\ i\neq j}} \sum\limits_{\substack{\ell\in\mathcal{N}\\ \ell\neq j}} \boldsymbol{a}_{i\ell}$ | $\mathfrak{m}_{a,\text{sum}} \triangleq \min\limits_{\substack{i,j\in\mathcal{N}\\ i\neq j}} \sum\limits_{\substack{\ell\in\mathcal{N}\\ \ell\neq j}} \boldsymbol{a}_{i\ell}$ | $\rho$ |
| $\mathfrak{M} \triangleq \max\limits_{\substack{i,j\in\mathcal{N}\\ i\neq j}} \sum\limits_{\substack{\ell\in\mathcal{N}\\ \ell\neq i,j}} \boldsymbol{a}_{i\ell}\boldsymbol{a}_{\ell j}$ | $\mathfrak{m} \triangleq \min\limits_{\substack{i,j\in\mathcal{N}\\ i\neq j}} \sum\limits_{\substack{\ell\in\mathcal{N}\\ \ell\neq i,j}} \boldsymbol{a}_{i\ell}\boldsymbol{a}_{\ell j}$ | $\sim \dfrac{\kappa^2 p}{N p_N}$ |

TABLE II
RANDOM VARIABLES RELEVANT FOR THE PROOF OF THEOREM 1. THE LIMITS ARE IN PROBABILITY, AND "$\sim$" MEANS "SCALES AS".

universal local structural consistency, with $s_N = N p_N$, and with the biases and gaps in Table I. ■

*Sketch of proof:* The $(i,j)$-th entry of $\boldsymbol{A}^k$ is denoted by $\boldsymbol{a}^{(k)}_{ij}$, and $\mathcal{N} \triangleq \{1,2,\ldots,N\}$. Table II lists some auxiliary variables used in the proof. For space limitations, we illustrate only the fundamental trick that allows proving the claim, with reference to the one-lag and the residual estimators. The idea is working directly in the graph edge domain, and obtaining uniform (w.r.t. $N$) bounds on the combination matrix entries, which are then exploited to characterize the errors ascribed to the different estimators. We will show that:

$$\underline{\boldsymbol{\alpha}}_k \leq \quad \boldsymbol{a}^{(k)}_{ii} \quad \leq \overline{\boldsymbol{\alpha}}_k, \tag{12}$$

$$\underline{\boldsymbol{\beta}}_k\, \boldsymbol{a}_{ij} + \underline{\boldsymbol{\gamma}}_k\, \mathfrak{m} \leq \quad \boldsymbol{a}^{(k)}_{ij} \quad \leq \overline{\boldsymbol{\beta}}_k\, \boldsymbol{a}_{ij} + \overline{\boldsymbol{\gamma}}_k\, \mathfrak{M} \quad \forall i \neq j, \tag{13}$$

where, for $k \geq 2$, the (random) sequences $\underline{\boldsymbol{\alpha}}_k$, $\overline{\boldsymbol{\alpha}}_k$, $\underline{\boldsymbol{\beta}}_k$, $\overline{\boldsymbol{\beta}}_k$, $\underline{\boldsymbol{\gamma}}_k$, and $\overline{\boldsymbol{\gamma}}_k$, are determined by the following recursions:

$$\begin{aligned}
&\underline{\boldsymbol{\alpha}}_{k+1} = \mathfrak{m}_{a,\text{self}}\, \underline{\boldsymbol{\alpha}}_k, && \overline{\boldsymbol{\alpha}}_{k+1} = \mathfrak{M}_{a,\text{self}}\, \overline{\boldsymbol{\alpha}}_k + \mathfrak{M}_a\, \rho^k,\\
&\underline{\boldsymbol{\beta}}_{k+1} = \underline{\boldsymbol{\alpha}}_k + \mathfrak{m}_{a,\text{self}}\, \underline{\boldsymbol{\beta}}_k, && \overline{\boldsymbol{\beta}}_{k+1} = \overline{\boldsymbol{\alpha}}_k + \mathfrak{M}_{a,\text{self}}\, \overline{\boldsymbol{\beta}}_k,\\
&\underline{\boldsymbol{\gamma}}_{k+1} = \underline{\boldsymbol{\beta}}_k + \mathfrak{m}_{a,\text{sum}}\, \underline{\boldsymbol{\gamma}}_k, && \overline{\boldsymbol{\gamma}}_{k+1} = \overline{\boldsymbol{\beta}}_k + \mathfrak{M}_{a,\text{sum}}\, \overline{\boldsymbol{\gamma}}_k,
\end{aligned} \tag{14}$$

with the initialization choices $\underline{\boldsymbol{\alpha}}_2 = \mathfrak{m}_{a_2,\text{self}}$, $\overline{\boldsymbol{\alpha}}_2 = \mathfrak{M}_{a_2,\text{self}}$, $\underline{\boldsymbol{\beta}}_2 = 2\,\mathfrak{m}_{a,\text{self}}$, $\overline{\boldsymbol{\beta}}_2 = 2\,\mathfrak{M}_{a,\text{self}}$, $\underline{\boldsymbol{\gamma}}_2 = \overline{\boldsymbol{\gamma}}_2 = 1$.

For $k = 2$ the claim in (12) is trivially true with the initialization values for $\underline{\boldsymbol{\alpha}}_2$ and $\overline{\boldsymbol{\alpha}}_2$, because of the definition of $\mathfrak{m}_{a_2,\text{self}}$ and $\mathfrak{M}_{a_2,\text{self}}$. We shall reason by induction to prove that (12) holds for an arbitrary $k$. Since we have $\boldsymbol{a}^{(k+1)}_{ii} = \sum_{\ell\in\mathcal{N}} \boldsymbol{a}_{i\ell}\boldsymbol{a}^{(k)}_{\ell i}$, by using the fact that $\sum_{\ell\in\mathcal{N}} \boldsymbol{a}^{(k)}_{\ell i} = \rho^k$, along with the definitions of the pertinent random variables in Table II, we see that:

$$\underbrace{\boldsymbol{a}_{ii}}_{\geq \mathfrak{m}_{a,\text{self}}}\, \underbrace{\boldsymbol{a}^{(k)}_{ii}}_{\geq \underline{\boldsymbol{\alpha}}_k} \leq \boldsymbol{a}^{(k+1)}_{ii} \leq \underbrace{\boldsymbol{a}_{ii}}_{\leq \mathfrak{M}_{a,\text{self}}}\, \underbrace{\boldsymbol{a}^{(k)}_{ii}}_{\leq \overline{\boldsymbol{\alpha}}_k} + \mathfrak{M}_a\, \rho^k, \tag{15}$$

and we conclude that (12) holds true. We switch to the proof of (13). Again, the suggested initialization choices make claim (13) true in the case $k = 2$. Moreover, we have:

$$\boldsymbol{a}^{(k+1)}_{ij} = \sum_{\ell\in\mathcal{N}} \boldsymbol{a}_{i\ell}\boldsymbol{a}^{(k)}_{\ell j} = \boldsymbol{a}_{ij}\boldsymbol{a}^{(k)}_{jj} + \sum_{\substack{\ell\in\mathcal{N}\\ \ell\neq j}} \boldsymbol{a}_{i\ell}\boldsymbol{a}^{(k)}_{\ell j}. \tag{16}$$

Assuming now that (13) holds for a certain $k \geq 2$, and bounding accordingly the terms $\boldsymbol{a}^{(k)}_{\ell j}$ in (16) we get:

$$\boldsymbol{a}^{(k+1)}_{ij} \leq (\overline{\boldsymbol{\alpha}}_k + \overline{\boldsymbol{\beta}}_k\, \mathfrak{M}_{a,\text{self}})\boldsymbol{a}_{ij} + (\overline{\boldsymbol{\beta}}_k + \overline{\boldsymbol{\gamma}}_k\, \mathfrak{M}_{a,\text{sum}})\, \mathfrak{M},$$

which shows that the rightmost inequality in (13) holds with the sequences $\overline{\boldsymbol{\beta}}_k$ and $\overline{\boldsymbol{\gamma}}_k$ obeying (14). Similar arguments lead to the conclusion that the leftmost inequality in (13) holds true.

From (6) and (8) we see that the errors ascribed to the estimated matrix involve powers of $\boldsymbol{A}$. The inequalities in (13) can be used to compute upper and lower bounding sequences for these errors. Using the limits in Table II we can show that the upper and the lower bounding sequences, scaled by $N p_N$, converge to the same limit: for disconnected pairs this limit is the bias $\eta$, whereas for connected pairs it is $\eta + \Gamma$, with the pertinent values of $\eta$ and $\Gamma$ listed in Table I. ■

*Remark 5 (Role of degree concentration):* Intuitively, under Assumption 1, concentration of the degrees induces concentration of the nonzero entries in the *true* matrix $\boldsymbol{A}_{\mathcal{S}}$, creating an identifiability gap in $\boldsymbol{A}_{\mathcal{S}}$. It is less intuitive that degree concentration induces an identifiability gap in an *estimated* matrix $\widehat{\boldsymbol{A}}_{\mathcal{S}}$, which is the critical property shown in Theorem 1. □

The meaning of Theorem 1 is illustrated in Fig. 1. The three panels refer to the estimators considered in this work, as detailed in the panel titles. We appreciate the emergence of the gap $\Gamma$ and of the bias $\eta$, which, remarkably, match well the theoretical limits summarized in Table I — see dashed lines.

We see from Table I that only the bias of the Granger estimator depends on the fraction of monitored nodes $\xi$, which makes sense since the Granger estimator is based on inversion of a partial matrix (with the latent variables introducing error), whereas the one-lag and the residual estimator are natively determined only by pairwise interactions. We see also that, for the one-lag and of the residual estimator, the term $\kappa$ (i.e., the Granger gap) multiplies a function of $\zeta = \rho - \kappa$. When $\kappa \neq \rho$ (as can happen, e.g., for the Laplacian rule [2]), this factor is greater than one for the one-lag estimator, whereas it is smaller than one for the residual estimator. However, we remark that a magnified/reduced gap do not itself imply any conclusion about the *classification* performance of the pertinent estimators, since it is the spread of the entries, relative to the asymptotic values, that determines the performance. Such spread is not characterized by our analysis, and predicting the performance for finite network sizes is nontrivial. Interesting behaviors can emerge, as we will see in the next section.

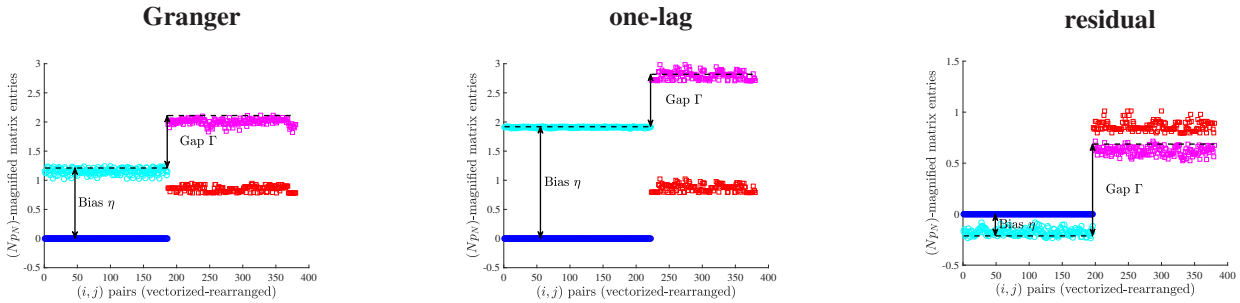


Fig. 1. Illustration of Theorem 1. The entries of the *true* combination matrix are vectorized (column-major ordering), and rearranged with the zero entries (blue) appearing before the nonzero entries (red). The entries of the *estimated* combination matrix are rearranged *with the same ordering used for the true matrix* (disconnected pairs displayed in magenta, connected pairs in cyan). Dashed lines depict theoretical values. Note that the rightmost panel has a different vertical range for better displaying.

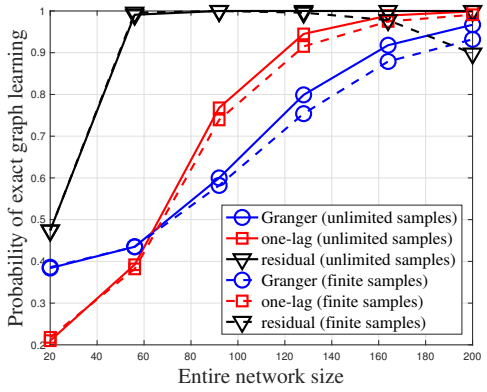


Fig. 2. Graph recovery probability over $10^3$ Monte Carlo runs. The connection probability is $p_N = p = 0.1$, the fraction of monitored nodes is $\xi = 0.2$, and we use a Laplacian rule (see [2]) with $\rho = 0.99$ and $\kappa = 0.891$. Empirical correlations are estimated over $5 \times 10^5$ time samples, with Gaussian data.

## IV. Illustrative Example

In Fig. 2 we consider one example corresponding to the system parameters detailed in the caption. The three estimators proposed in this work are computed by evaluating the theoretical (unlimited amount of data) correlation matrices, $R_0$ and $R_1$, as well as by evaluating the empirical correlations from finite data samples. The classification stage uses a $k$-means algorithm with $k = 2$. Let us start with the unlimited amount of data. We see that all the estimators match well the theoretical predictions. In this particular example, after an initial transient with small network sizes, the Granger estimator is outperformed by the other two estimators. This is a remarkable finding, since it highlights how what is good under the full-observability case (i.e., the Granger estimator) need not be optimal under a partial-observability setting. One reason behind this finding can be the sensitivity of the Granger estimator to the degree of observability. Moving on to examine the case of finite sample sizes, we observe that different estimators exhibit a different sensitivity to the number of samples. In particular, the residual estimator, which offers the best performance in the case of unlimited sample size, needs more samples as the network size increases.

## References


[1] G. Mateos, S. Segarra, A. Marques, and A. Ribeiro, "Connecting the dots," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019.

[2] 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.

[3] A. Santos, V. Matta, and A. H. Sayed, "Local tomography of large networks under the low-observability regime," *submitted for publication,* available online as arXiv:1805.09081v1 [cs.MA], May 2018.

[4] A. Anandkumar and R. Valluvan, "Learning loopy graphical models with latent variables: Efficient methods and guarantees," *The Annals of Statistics*, vol. 41, no. 2, pp. 401–435, Apr. 2013.

[5] V. Matta and A. H. Sayed, "Estimation and detection over adaptive networks," in *Cooperative and Graph Signal Processing*, P. Djuric and C. Richard, Eds. Elsevier, 2018, pp. 69–106.

[6] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[7] H. Salami, B. Ying, and A. H. Sayed, "Social learning over weakly connected graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 2, pp. 222–238, Jun. 2017.

[8] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann, "Predicting human resting-state functional connectivity from structural connectivity," *Proceedings of the National Academy of Sciences*, vol. 106, no. 6, pp. 2035–2040, Feb. 2009.

[9] P. Geiger, K. Zhang, B. Schölkopf, M. Gong, and D. Janzing, "Causal inference by identification of vector autoregressive processes with hidden components," in *Proc. International Conference on Machine Learning (ICML)*, vol. 37, Lille, France, Jul. 2015, pp. 1917–1925.

[10] 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.

[11] G. Bresler, D. Gamarnik, and D. Shah, "Hardness of parameter estimation in graphical models," in *Proc. Neural Information Processing Systems (NIPS)*, Montréal, Canada, Dec. 2014, pp. 1062–1070.

[12] J. Etesami, N. Kiyavash, and T. Coleman, "Learning minimal latent directed information polytrees," *Neural Computation*, vol. 28, no. 9, pp. 1723–1768, Aug. 2016.

[13] D. Materassi and M. V. Salapaka, "Identification of network components in presence of unobserved nodes," in *Proc. IEEE Conference on Decision and Control (CDC)*, Osaka, Japan, Dec. 2015, pp. 1563–1568.

[14] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *The Annals of Statistics*, vol. 40, no. 4, pp. 1935–1967, Aug. 2012.

[15] 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.

[16] P. Erdős and A. Rényi, "On random graphs I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[17] B. Bollobás, *Random Graphs*. Cambridge Univ. Press, 2001.

[18] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, 2013.