Learning Bollobás-Riordan Graphs Under Partial Observability

Michele Cirillo*, Vincenzo Matta*, Ali H. Sayed†

*DIEM, University of Salerno, Italy

†School of Engineering, EPFL, Switzerland

Abstract

This work examines the problem of learning the topology of a network (*graph learning*) from the signals produced at a subset of the network nodes (*partial observability*). This challenging problem was recently tackled assuming that the topology is drawn according to an Erdős-Rényi model, for which it was shown that graph learning under partial observability is achievable, exploiting in particular homogeneity across nodes and independence across edges. However, several real-world networks do not match the optimistic assumptions of homogeneity/independence, for example, high heterogeneity is often observed between very connected nodes (hubs) and scarcely connected peripheral nodes. Random graphs with *preferential attachment* were conceived to overcome these issues. In this work, we discover that, over first-order vector autoregressive systems with a stable Laplacian combination matrix, graph learning is achievable under partial observability, when the network topology is drawn according to a popular preferential attachment model known as the Bollobás-Riordan model.

Considered Scenario

Diffusion Model

Vector AutoRegressive (VAR) dynamics:

$$\boldsymbol{y}_{k,i} = \sum_{\ell=1}^{N} a_{k\ell}\, \boldsymbol{y}_{\ell,i-1} + \boldsymbol{x}_{k,i} \qquad (1)$$

having **Laplacian** combination matrix $A = [a_{k\ell}]$, with:

$$\begin{cases} a_{k\ell} \propto d_{\max}^{-1}, & \text{if nodes } k \text{ and } \ell \text{ are connected} \\ a_{k\ell} = 0, & \text{otherwise} \end{cases}$$

and with a_{kk} chosen to make the matrix doubly stochastic

- Boldface font for random variables
- N: total number of network nodes
- $k, \ell = 1, \dots, N$: node indices
- $i = 1, 2, \dots$: time instants
- $\boldsymbol{x}_{k,i}$: input signal of node k at time i
- $\boldsymbol{y}_{k,i}$: output signal of node k at time i
- $d_{\max}$: maximum node degree

Partial Observability

We observe the signals $\boldsymbol{y}_{k,i}$ emitted **only** by the nodes belonging to a subset $\mathcal{S}$. (A subscript $\mathcal{S}$ denotes sub-matrices relative to $\mathcal{S}$). Let R_0, R_1 be the steady-state covariance and one-lag covariance matrices of $\boldsymbol{y}_{k,i}$

Granger Predictor for $A_{\mathcal{S}}$

$$A_{\mathcal{S}} = [R_1 R_0^{-1}]_{\mathcal{S}}.$$

Here the covariance matrices relative to *all* nodes are necessary

Partial Granger Predictor

$$\widehat{A}_{\mathcal{S}} = [R_1]_{\mathcal{S}} \left([R_0]_{\mathcal{S}}\right)^{-1}.$$

Here only the covariance matrices relative to the *probed* nodes are used

Issue

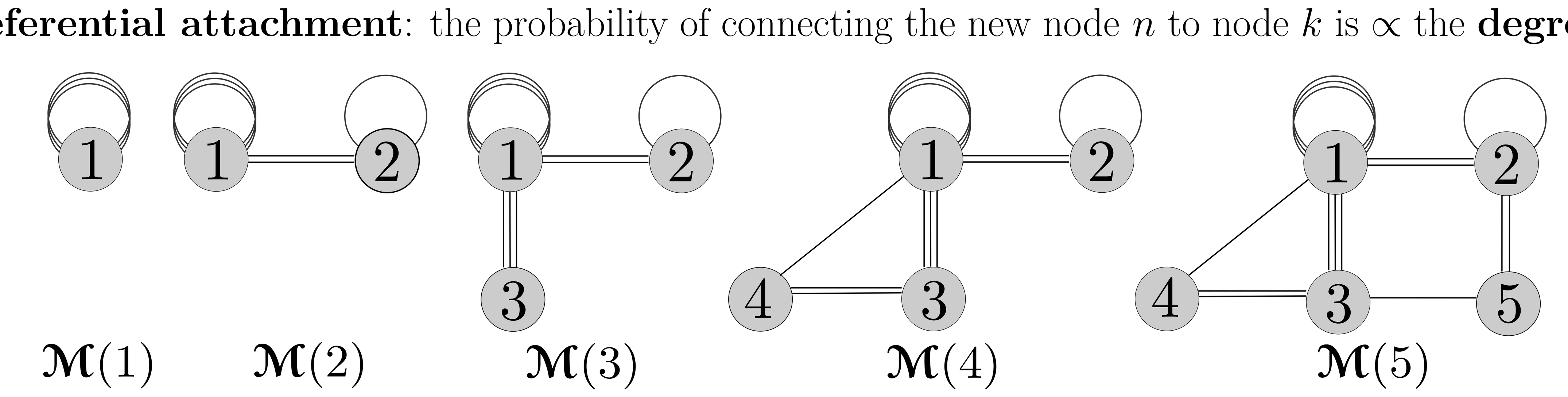
$$A_{\mathcal{S}} \neq \widehat{A}_{\mathcal{S}}.$$

The unobserved nodes introduce an unavoidable error

Bollobás-Riordan Random Graphs

The Bollobás-Riordan construction [1] produces a **multi-graph** (multiple edges allowed)

- Start from an **initial multi-graph** $\mathcal{M}(1)$ with one node and $\eta \in \mathbb{N}$ edges
- For $n = 2, \dots, N$, build the multi-graph $\mathcal{M}(n)$ by adding to $\mathcal{M}(n-1)$ a new node n and η new edges
- **Preferential attachment**: the probability of connecting the new node n to node k is $\propto$ the **degree** of k



- After the multi-graph $\mathcal{M}(N)$ is built, **uproot** the repeated edges and the self-loops from it
- The resulting graph is used in the VAR diffusion model (1)

Main Result

Theorem 1

Let $\epsilon > 0$ and let $\xi \in (0,1)$ be an arbitrary fraction of probed nodes, namely, $|\mathcal{S}|/N \to \xi$ as $N \to \infty$. Then, there exists a random variable $\boldsymbol{\Gamma} > 0$ such that $\widehat{\boldsymbol{A}}_{\mathcal{S}}$ (in bold because the graph is random) satisfies the following properties with high probability as $N \to \infty$. If $k, \ell \in \mathcal{S}$ are connected we have:

$$(1-\epsilon)\boldsymbol{\Gamma} < \sqrt{N}\, \widehat{\boldsymbol{a}}_{k\ell} < (1+\epsilon)\boldsymbol{\Gamma}, \qquad (2)$$

whereas if k and ℓ are unconnected we have:

$$0 < \sqrt{N}\, \widehat{\boldsymbol{a}}_{k\ell} < \epsilon\boldsymbol{\Gamma}. \qquad (3)$$

- The entries of $\sqrt{N}\widehat{\boldsymbol{A}}_{\mathcal{S}}$ corresponding to connected pairs are clustered around a value $\boldsymbol{\Gamma} > 0$, see (2)
- The entries of $\sqrt{N}\widehat{\boldsymbol{A}}_{\mathcal{S}}$ corresponding to unconnected pairs are clustered around zero, see (3)

Take-Away Message

The graph of the probed nodes can be retrieved from $\widehat{\boldsymbol{A}}_{\mathcal{S}}$ via standard clustering algorithms

Illustrative Examples

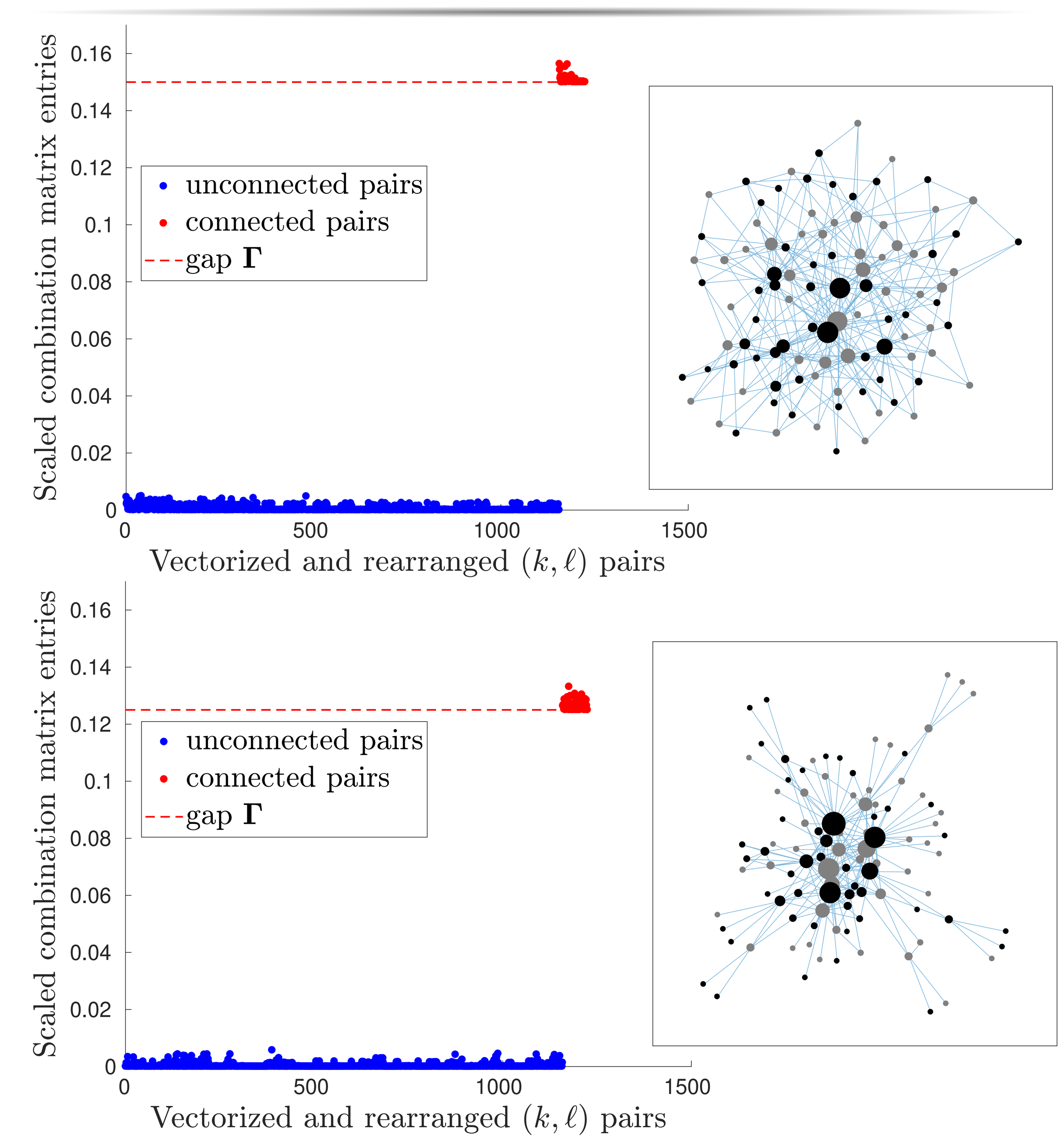


Figure 1: Illustration of Theorem 1 for three networks of $N = 100$ nodes, and fraction of probed nodes $\xi = 0.5$. Probed nodes are displayed with black circles, with radius proportional to their degree. We show the entries of the estimated matrix, scaled by $\sqrt{N}$, ordered so that the entries for unconnected pairs come first. The dashed line displays the gap $\boldsymbol{\Gamma}$. *Top*: realization of a Bollobás-Riordan graph. *Bottom*: graph of 100 nodes extracted from a real-world network of routers [7].

Performance

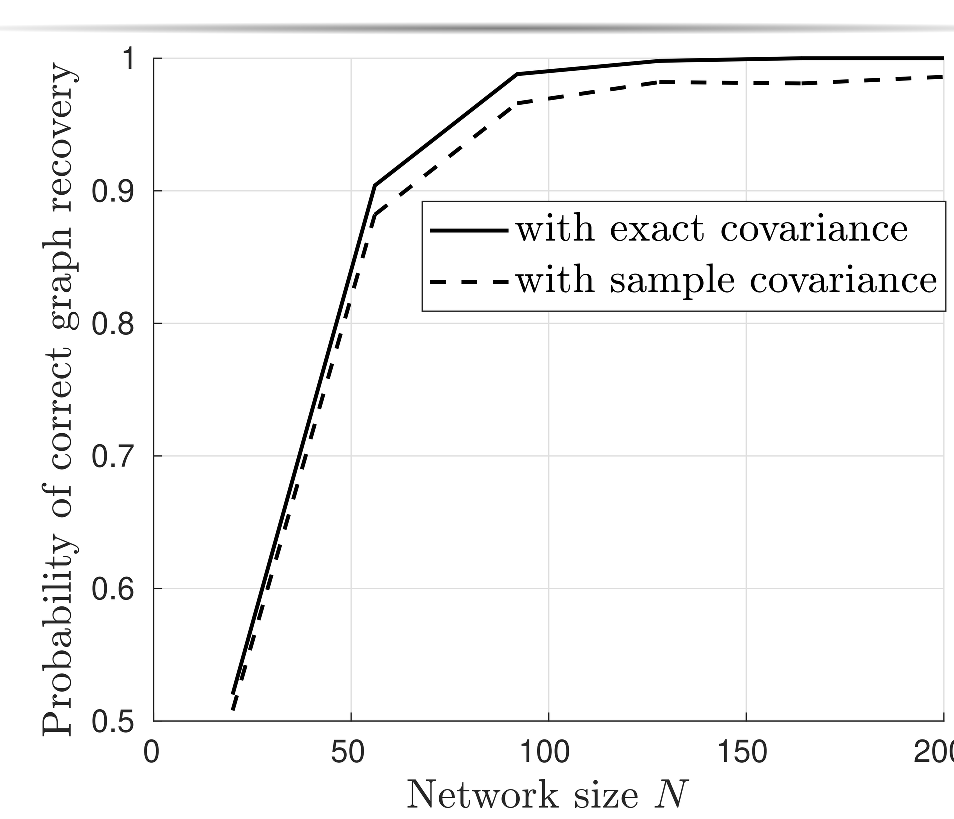


Figure 2: Probability of correct graph recovery, computed over 10^3 Monte Carlo runs, as a function of N. Solid line: *limiting* estimator with true covariances. Dashed line: *empirical* estimator with sample covariances computed over 10^5 samples.

References

[1] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, "The degree sequence of a scale-free random graph process," *Random Struct. Alg.*, 2001

[2] V. Matta, A. Santos, and A. H. Sayed, "Graph learning under partial observability," *Proceedings of the IEEE*, 2020

[3] F. Han, H. Lu, and H. Liu, "A direct estimation of high dimensional stationary vector autoregressions," *JMLR*, 2015

[4] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," IEEE Signal Process. Mag., 2019

[5] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion," *JMLR*, 2012

[6] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *Ann. Statist.*, 2012

[7] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI*, 2015