PSEUDO-LIKELIHOOD-BASED M-ESTIMATION OF RANDOM GRAPHS WITH DEPENDENT EDGES AND PARAMETER VECTORS OF INCREASING DIMENSION

By Jonathan R. Stewart
Florida State University

By Michael Schweinberger
Rice University

An important question in statistical network analysis is how to estimate models of dependent network data without sacrificing computational scalability and statistical guarantees. We demonstrate that scalable estimation of random graph models with dependent edges is possible, by establishing the first consistency results and convergence rates for pseudo-likelihood-based M-estimators for parameter vectors of increasing dimension based on a single observation of dependent random variables. The main results cover models of dependent random variables satisfying weak dependence conditions, and may be of independent interest. To showcase consistency results and convergence rates, we introduce a novel class of generalized β-models with dependent edges and parameter vectors of increasing dimension. We establish consistency results and convergence rates for pseudo-likelihood-based M-estimators of generalized β-models with dependent edges, in dense- and sparse-graph settings.

1. Introduction.

Network data have garnered considerable attention in recent years [49], driven by the growth of the internet and online social networks that can serve as echo chambers and facilitate polarization, and applications in science, technology, and public health (e.g., the spread of infectious diseases through networks of contacts).

Substantial progress has been made on models of network data with independence assumptions, such as the p_1-model and the β-model [e.g., 42, 17, 81, 64, 48, 80, 79, 60, 19]; exchangeable random graph models [e.g., 12, 14, 21]; random graph models with latent structure, such as stochastic block models [e.g., 62, 1, 7, 29, 30], latent space models [e.g., 40, 72], and other latent variable models [e.g., 14, 39]; and exponential-family models of

[*] We acknowledge support from the National Science Foundation (NSF awards DMS-1513644 and DMS-1812119) and are grateful to Sergii Babkin for helpful discussions.

[†] MSC2010 subject classifications. Primary 05C80; secondary 62B05, 62F10, 91D30.

Keywords and phrases: β-model, generalized β-models, exponential random graph models, graphical models, Markov random fields, maximum pseudo-likelihood estimators.

random graphs with constraints on the parameter space [e.g., 69, 44, 43, 65, 61, 68] or the dependence structure [66]. That being said, fundamental questions arising from the statistical analysis of non-standard, dependent, and high-dimensional network data have remained unanswered.

1.1. *Two questions.* Since the dawn of statistical network analysis in the 1980s [42, 28], two questions have loomed large:

I. How can one construct random graph models that do justice to the fact that (a) the propensities of nodes to form edges may vary across nodes and (b) network data are dependent data?
II. How can one estimate random graph models based on a single observation of a random graph with dependent edges and parameter vectors of increasing dimension?

Due to the challenge of constructing models of random graphs with dependent edges without sacrificing computational scalability and statistical guarantees, much of the recent literature has focused on models with independence assumptions, such as the p_1-model (the first probability model for random graphs with directed edges) and the β-model (its relative for random graphs with undirected edges), and latent structure models with conditional independence assumptions, with stochastic block models receiving most of the spotlight [7]. Many of these models can capture heterogeneity in the propensities of nodes to form edges, but make explicit or implicit independence or weak dependence assumptions that may not be satisfied by real-world networks, because network data are dependent data [e.g., 41]. Meanwhile, the question of what can be learned from a single observation of a random graph with dependent edges and parameter vectors of increasing dimension has remained wide open.

Two recent works have taken first steps towards capturing dependence among edges along with statistical guarantees, but both have limitations. Mukherjee [61] considered models with functions of degrees as sufficient statistics, which allow edges to be dependent but have two parameters and do not capture network features other than degrees. Schweinberger and Stewart [68] considered a wide range of models with dependent edges, but constrained the dependence among edges to non-overlapping subpopulations of nodes.

We address here the question of how to construct and estimate random graph models with dependent edges, without sacrificing computational scalability and statistical guarantees. To do so, we first introduce a probabilistic framework that facilitates the construction of random graph models with dependent edges from simple building blocks. We then establish the first consistency results and convergence rates for pseudo-likelihood-based M-

estimators for parameter vectors of increasing dimension based on a single observation of dependent random variables, which may be of independent interest. In applications to random graphs, these results provide the first statistical guarantees for scalable estimation of random graph models with dependent edges and parameter vectors of increasing dimension.

1.2. *Probabilistic framework.* On the modeling side, we introduce a simple and flexible approach to specifying models of random graphs with dependent edges, drawing inspiration from Whittle's [77] work on spatial processes. We demonstrate the probabilistic framework by extending the β-model of Chatterjee et al. [17]—studied by Rinaldo et al. [64], Yan and Xu [81], Karwa and Slavković [48], Mukherjee et al. [60], and Chen et al. [19], among others—to generalized β-models capturing heterogeneity in the propensities of nodes to form edges along with dependence among edges. The number of parameters of the β-model and generalized β-models increases with the number of nodes, but the β-model assumes that edges are independent, whereas generalized β-models allow edges to be dependent. The β-model and generalized β-models with dependent edges serve as running examples throughout the remainder of the paper.

1.3. *Computational scalability and statistical guarantees.* On the statistical side, we demonstrate that computational scalability and statistical guarantees need not be sacrificed in order to estimate random graph models with dependent edges and parameter vectors of increasing dimension.

While it may be tempting to estimate random graph models with dependent edges by using likelihood-based methods, computing likelihood-based estimators is often infeasible, because the likelihood function of many random graph models with dependent edges does not possess convenient factorization properties that would facilitate computations. Approximate likelihood-based estimators—e.g., Monte Carlo maximum likelihood estimators [44] and Bayesian Markov chain Monte Carlo estimators [13]—can be computed, but are expensive in terms of computing time [5, 16]. In spatial statistics, where similar computational problems arise from the dependence among random variables induced by discrete Markov random fields with lattice structure, Besag [4] proposed maximum pseudo-likelihood estimators as scalable alternatives to maximum likelihood estimators. In the literature on Markov random fields and Ising models, consistency and asymptotic normality of maximum pseudo-likelihood estimators has been established [35, 46, 20, 58, 45, 15, 6, 34]. However, all of those results are limited to models with a fixed and finite number of parameters and a specific form of interactions among random variables (e.g., pairwise interactions) and a specific form of structure (e.g., lattice structure). A notable exception is the recent work of Ghosal and Mukherjee [34], who focus on two-parameter Ising models with pairwise interactions, but investigate general conditions on the interaction structure under which consistent estimation is possible.

Frank and Strauss [28] and Strauss and Ikeda [71] adapted maximum pseudo-likelihood estimators to random graph models, without providing statistical guarantees. The literature concerned with dependent network data has viewed maximum pseudo-likelihood estimators with skepticism [e.g., 75], with the exception of the literature on stochastic block models [2]. Some of the skepticism may be rooted in the fact that, during the early days of statistical network analysis, maximum pseudo-likelihood estimators were applied to models that turned out to be ill-posed [5, 65, 16], and maximum pseudo-likelihood estimators may have been blamed for problems rooted in those models [67]. Other concerns stem from the fact that many random graph models do not possess dependence structure that resembles the dependence structure of Markov random fields with lattice structure, and random graph models may need many parameters to capture heterogeneity in the propensities of nodes to form edges. As a consequence, theoretical results on Markov random fields and Ising models with lattice structure and a fixed and finite number of parameters are not applicable to most random graph models—despite the fact that many of those models share the same mathematical foundation, by virtue of being families of strictly positive distributions with countable support and exponential parameterizations [9].

We demonstrate that scalable estimation of random graph models with dependent edges and parameter vectors of increasing dimension is possible, using pseudo-likelihood-based M-estimators. We do so by establishing the first consistency results and convergence rates for pseudo-likelihood-based M-estimators for parameter vectors of increasing dimension based on a single observation of dependent random variables. The main results are stated in terms of random graphs, but could be extended to models of dependent random variables with countable and uncountable sample spaces. In fact, the main results neither assume a specific form of interactions (e.g., pairwise interactions) nor a specific form of structure (e.g., lattice structure), although the main results make weak dependence assumptions and additional structure (e.g., overlapping or non-overlapping subpopulation structure) is helpful for verifying them. We showcase consistency results and convergence rates for pseudo-likelihood-based M-estimators in applications to random graphs, using generalized β-models with dependent edges and parameter vectors of increasing dimension as examples. These results cover dense- and sparse-graph settings.

1.4. *Structure of the paper.* Section 2 introduces the probabilistic framework. Section 3 discusses consistency results and convergence rates for pseudo-likelihood-based M-estimators, along with applications to generalized β-models with dependent edges and parameter vectors of increasing dimension. Simulation results are presented in Section 4.

1.5. *Notation.* Let $\mathcal{N} = \{1, \dots, N\}$ $(N \geq 2)$ be a finite set of nodes and $\boldsymbol{X}$ be a random graph defined on $\mathcal{N}$ with sample space $\mathbb{X} = \{0, 1\}^{\binom{N}{2}}$, where $X_{i,j} = 1$ if nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ are connected by an edge and $X_{i,j} = 0$ otherwise. We focus here on random graphs with undirected edges and without self-edges, although all results reported here can be extended to directed random graphs. To denote subgraphs induced by subsets of nodes $\mathcal{C} \subseteq \mathcal{N}$ and $\mathcal{D} \subseteq \mathcal{N}$, we write $\boldsymbol{X}_{\mathcal{C},\,\mathcal{D}} = (X_{i,j} \,:\, i \neq j, i \in \mathcal{C}, j \in \mathcal{D})$. If $\mathcal{C} = \mathcal{D}$, we write $\boldsymbol{X}_{\mathcal{C}} \equiv \boldsymbol{X}_{\mathcal{C},\,\mathcal{C}} = (X_{i,j} :\; i \neq j,\; i \in \mathcal{C},\; j \in \mathcal{C})$. Two special cases are subgraphs induced by intersections $\mathcal{C} \cap \mathcal{D}$ of subsets of nodes $\mathcal{C}$ and $\mathcal{D}$, denoted by $\boldsymbol{X}_{\mathcal{C} \cap \mathcal{D}} = (X_{i,j} :\; i \neq j,\; i \in \mathcal{C} \cap \mathcal{D},\; j \in \mathcal{C} \cap \mathcal{D})$, and subgraphs induced by unions $\mathcal{C} \cup \mathcal{D}$ of subsets of nodes $\mathcal{C}$ and $\mathcal{D}$, denoted by $\boldsymbol{X}_{\mathcal{C} \cup \mathcal{D}} = (X_{i,j} :\; i \neq j,\; i \in \mathcal{C} \cup \mathcal{D},\; j \in \mathcal{C} \cup \mathcal{D})$. The set $\mathbb{R}^+ = (0, \infty)$ denotes the set of positive real numbers, and the vector $\boldsymbol{0} \in \mathbb{R}^d$ denotes the d-dimensional null vector $(d \geq 1)$. We denote the ℓ_1-, ℓ_2-, and ℓ_∞-norm of vectors in $\mathbb{R}^d$ by $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$, respectively, the spectral norm of matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ by $|||\boldsymbol{A}|||_2 = \sup_{\boldsymbol{u} \in \mathbb{R}^d:\, \|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{A}\,\boldsymbol{u}\|_2$, and the determinant of matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ by $\det(\boldsymbol{A})$. The ℓ_∞-ball in $\mathbb{R}^d$ centered at $\boldsymbol{c} \in \mathbb{R}^d$ with radius $\rho > 0$ is denoted by $\mathcal{B}_\infty(\boldsymbol{c},\, \rho) = \{\boldsymbol{a} \in \mathbb{R}^d :\; \|\boldsymbol{a} - \boldsymbol{c}\|_\infty \leq \rho\}$. For any finite set $\mathcal{S}$, the number of elements of $\mathcal{S}$ is denoted by $|\mathcal{S}|$. The function $\mathbb{1}(\,\cdot\,)$ is an indicator function, which is 1 if its argument is true, and is 0 otherwise. Unless stated otherwise, uppercase letters $A, B, C, \dots$ denote finite constants, which may be recycled from line to line.

2. Probabilistic framework.

Two broad approaches to specifying models of dependent random variables are based on specifying the joint probability density function either directly or indirectly, that is, by specifying conditional probability density functions and obtaining the joint probability density function by invoking the Hammersley-Clifford theorem, assuming the joint probability density function is strictly positive [4, 52]. In the literature on random graph models with dependent edges, the most common approach (leaving aside models with latent structure, such as stochastic block models [62, 7, 29, 30] and latent space models [40, 72]) is to first choose statistics that capture interesting features of real-world networks, and then base statistical inference on the exponential family generated by the chosen statistics [33, 37]. Model construction along these lines can be motivated by the maximum entropy property of exponential families [33, 37].

We introduce here a simple and flexible approach to specifying random graph models with complex dependence from simple building blocks, drawing inspiration from Whittle's [77] work on spatial processes. To elaborate, consider a family of probability measures $\{\mathbb{P}_{\boldsymbol{\theta}_N},\, \boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N\}$ dominated by a σ-finite measure $\nu : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$, with densities of the form

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \;\; \propto \;\; \prod_{i<j}^{N} \varphi_{i,j}(x_{i,j},\, \boldsymbol{x}_{\mathcal{S}_{i,j}};\, \boldsymbol{\theta}_N), \qquad \boldsymbol{x} \in \mathbb{X}, \tag{2.1}$$

where $\varphi_{i,j} : \{0,1\}^{|\mathcal{S}_{i,j}|+1} \times \boldsymbol{\Theta}_N \mapsto \mathbb{R}^+ \cup \{0\}$ is a function that specifies how edge variable $X_{i,j}$ depends on a subset of edge variables,

$$\boldsymbol{X}_{\mathcal{S}_{i,j}} \;\; = \;\; (X_{a,b} : \{a,b\} \subseteq \mathcal{S}_{i,j}), \qquad \mathcal{S}_{i,j} \subset \{\{c,d\} \subset \mathcal{N} :\, c < d\},$$

parameterized by $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$. We assume that $\boldsymbol{\Theta}_N$ is an open subset of $\mathbb{R}^p$, and allow the dimension p of the parameter vector $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ to increase as a function of the number of nodes N. A natural choice of the reference measure ν is counting measure on $\mathbb{X}$, although other reference measures can be chosen, e.g., reference measures that place more mass on sparse, rather than dense, graphs [67]. An example of sparsity-inducing reference measures is mentioned in Section 2.5. It is worth noting that the factorization of probability density function (2.1) does not imply that edges are independent, because each $\varphi_{i,j}$ can be a function of multiple edges and can hence induce dependence among edges. The generalized β-models in Sections 2.3–2.5 demonstrate that.

Remark 1. Relationship to Whittle [77] and undirected graphical models. The factorization of probability density function (2.1) is reminiscent of the factorization of probability density functions of undirected graphical models [52, 57], which can be traced back to Whittle's [77] work on spatial processes [4]. Undirected graphical models of random graphs have been studied elsewhere [52, 53], but undirected graphical models make the additional assumption that the probability density function can be written as a product of non-negative functions defined on complete subsets of an underlying conditional independence graph. While it is possible to construct undirected graphical models of random graphs, we are not interested in undirected graphical models of random graphs, but we are interested in specifying models with complex dependence from simple building blocks. Specifying models by specifying (2.1) facilitates the construction of models with complex dependence, because models can be specified by specifying functions of edges.

Remark 2. Conditional independence properties and local computing. It is well-known that factorization properties of probability density functions imply conditional independence properties, as discussed in the foundational work on conditional independence by Dawid [23, 24]. Conditional independence and factorization properties have computational advantages, and have long been exploited in the graphical model literature for the purpose of local computing on subgraphs of conditional independence graphs [e.g., 54]. While the probabilistic framework introduced here is not limited to graphical models of random graphs, the factorization of probability density function (2.1) does have conditional independence properties, which can be exploited for the purpose of local computing on subgraphs of random graphs. We review selected conditional independence properties in the supplement [70].

2.1. *Parameterizations.* It is convenient to parameterize the functions of edges $\varphi_{i,j}$ by using exponential parameterizations, so that the resulting probability density function is non-negative:

$$\varphi_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}}; \boldsymbol{\theta}_N) \;=\; a_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}}) \exp(\langle \boldsymbol{\theta}_N,\, s_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}})\rangle), \tag{2.2}$$

where $a_{i,j} : \{0,1\}^{|\mathcal{S}_{i,j}|+1} \mapsto \mathbb{R}^+ \cup \{0\}$ is a function of $x_{i,j}$ and $\boldsymbol{x}_{\mathcal{S}_{i,j}}$, which can be used to induce sparsity by penalizing edges, and $\langle \boldsymbol{\theta}_N,\, s_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}})\rangle$ is the inner product of a vector of parameters $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ and a vector of statistics $s_{i,j} : \{0,1\}^{|\mathcal{S}_{i,j}|+1} \mapsto \mathbb{R}^p$ ($\{i,j\} \subset \mathcal{N}$). Such parameterizations are widely used in the related literature on undirected graphical models: see, e.g., Lauritzen [52, pp. 71–72], Geiger et al. [31], and Wainwright and Jordan [76]. We demonstrate in Sections 2.2–2.5 that these parameterizations

- allow the dimension p of the parameter vector $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ to increase as a function of the number of nodes N;
- help specify random graph models with dependent edges that place more mass on sparse than dense graphs;
- allow parameters to depend on the sizes of subpopulations of nodes.

To do so, we extend the β-model—the undirected version of the first probability model for random graphs with directed edges, the p_1-model [42]—to generalized β-models. In contrast to the β-model, generalized β-models allow edges to be dependent. Other classes of generalized β-models were introduced by Rinaldo et al. [64] and Fan et al. [26], but those models assume that edges are independent, as does the original β-model of Chatterjee et al. [17]. In addition to generalized β-models with dependent edges, we introduce generalized β-models that place more mass on sparse than dense graphs, and generalized β-models with size-dependent parameterizations.

2.2. *Model 1: β-model with independent edges.* To introduce generalized β-models with dependent edges, we first review the β-model with independent edges studied by Chatterjee et al. [17] and others. The β-model assumes that edges between nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ are independent Bernoulli$(\mu_{i,j})$ $(\mu_{i,j} \in (0,1))$ random variables, where

$$\log \frac{\mu_{i,j}}{1 - \mu_{i,j}} \;\; = \;\; \theta_i + \theta_j, \qquad \theta_i \in \mathbb{R}, \qquad \theta_j \in \mathbb{R}.$$

The parameters θ_i and θ_j can be interpreted as the propensities of nodes i and j to form edges. The β-model is a special case of the probabilistic framework introduced above, corresponding to

$$\varphi_{i,j}(x_{i,j};\, \boldsymbol{\theta}_N) \;\; = \;\; a_{i,j}(x_{i,j}) \, \exp((\theta_i + \theta_j)\, x_{i,j}),$$

where

$$a_{i,j}(x_{i,j}) \;\; = \;\; \begin{cases} 1 & \text{if } x_{i,j} \in \{0,1\} \\[2ex] 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The β-model captures heterogeneity in the propensities of nodes to form edges, but assumes that edges are independent.

2.3. *Model 2: generalized β-model with dependent edges.* We introduce a generalization of the β-model, which captures heterogeneity in the propensities of nodes to form edges along with dependence among edges induced by brokerage in networks. Brokerage can influence economic and political outcomes of interest and has therefore been studied by economists, political scientists, and other network scientists since at least the 1980s [e.g., 36, 10]. An example of brokerage is given by faculty members of universities with appointments in both computer science and statistics, who can facilitate collaborations among faculty members in computer science and faculty members in statistics. In other words, faculty members with appointments in multiple departments can facilitate interdisciplinary research.

To capture dependence among edges induced by brokerage in networks, consider a finite population of nodes $\mathcal{N}$ consisting of $K \geq 2$ known subpopulations $\mathcal{A}_1, \ldots, \mathcal{A}_K$, which may overlap in the sense that the intersections of the subpopulations are non-empty. As a consequence, nodes may belong to multiple subpopulations: e.g., faculty members of universities may have appointments in multiple departments, which implies that the faculties of departments overlap. Subpopulation structure is inherent to many real-world networks, in part because people tend to build communities, and in part because organizations tend to divide large bodies of people into small bodies of people (e.g., divisions, subdivisions).

Define, for each node $i \in \mathcal{N}$, the subset of all nodes $j \in \mathcal{N}$ that share at least one subpopulation with node $i \in \mathcal{N}$:

$$\mathcal{N}_i \;=\; \{j \in \mathcal{N} \setminus \{i\} :\; \text{exists } k \in \{1, \dots, K\} \text{ such that } i \in \mathcal{A}_k \text{ and } j \in \mathcal{A}_k\},$$

which we call the neighborhood of node $i \in \mathcal{N}$. To capture dependence among edges induced by shared partners in the intersections of neighborhoods, we consider functions of edges $\varphi_{i,j}$ of the form

$$\varphi_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}}; \boldsymbol{\theta}_N) \;=\; a_{i,j}(x_{i,j}) \exp\left((\theta_i + \theta_j)\, x_{i,j} + \theta_{N+1}\, b_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}})\right),$$

where $\mathcal{S}_{i,j} = \{i,j\} \times \{\mathcal{N}_i \cap \mathcal{N}_j\}$, $a_{i,j} : \{0,1\} \mapsto \mathbb{R}^+ \cup \{0\}$ is given by (2.3), and $b_{i,j} : \{0,1\}^{|\mathcal{S}_{i,j}|+1} \mapsto \{0,1\}$ is given by

$$b_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}}) = \begin{cases} 0 & \text{if } \mathcal{N}_i \cap \mathcal{N}_j = \emptyset \\ x_{i,j}\, \mathbb{1}\left(\displaystyle\sum_{h \,\in\, \mathcal{N}_i \,\cap\, \mathcal{N}_j} x_{i,h}\, x_{j,h} \geq 1\right) & \text{if } \mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset. \end{cases} \tag{2.4}$$

Here, $\mathbb{1}(\sum_{h \,\in\, \mathcal{N}_i \,\cap\, \mathcal{N}_j} x_{i,h}\, x_{j,h} \geq 1)$ is an indicator function, which is 1 if nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ have at least one shared partner in the intersection $\mathcal{N}_i \cap \mathcal{N}_j$ of neighborhoods $\mathcal{N}_i$ and $\mathcal{N}_j$, and is 0 otherwise.

Remark 3. Generalized β-model captures brokerage in networks. The generalized β-model captures brokerage in networks, along with heterogeneity in the propensities of nodes to form edges. To demonstrate, consider the two overlapping subpopulations $\mathcal{A}_1$ and $\mathcal{A}_2$ shown in Figure 1. The nodes $1 \in \mathcal{A}_1 \setminus \mathcal{A}_2$ and $2 \in \mathcal{A}_2 \setminus \mathcal{A}_1$ do not belong to the same subpopulation, but the shared partner $3 \in \mathcal{A}_1 \cap \mathcal{A}_2$ in the intersection of subpopulations $\mathcal{A}_1$ and $\mathcal{A}_2$ can facilitate an edge between nodes 1 and 2, provided $\theta_{N+1} > 0$. In the language of network science, nodes in the intersection $\mathcal{A}_1 \cap \mathcal{A}_2$ of subpopulations $\mathcal{A}_1$ and $\mathcal{A}_2$ can act as brokers, facilitating edges between nodes in $\mathcal{A}_1 \setminus \mathcal{A}_2$ and nodes in $\mathcal{A}_2 \setminus \mathcal{A}_1$. In fact, the generalized β-model can capture an excess in the expected number of brokered edges relative to the

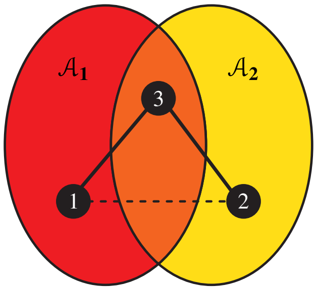

FIG 1. *A graphical representation of the dependencies among edges induced by brokerage. Consider two overlapping subpopulations $\mathcal{A}_1$ and $\mathcal{A}_2$. The nodes $1 \in \mathcal{A}_1 \setminus \mathcal{A}_2$ and $2 \in \mathcal{A}_2 \setminus \mathcal{A}_1$ do not belong to the same subpopulation, but the shared partner $3 \in \mathcal{A}_1 \cap \mathcal{A}_2$ in the intersection of subpopulations $\mathcal{A}_1$ and $\mathcal{A}_2$ can facilitate an edge between nodes 1 and 2, indicated by the dashed line between nodes 1 and 2.*

β-model, in the sense that

$$\underbrace{\mathbb{E}_{\theta_1,\ldots,\theta_N,\theta_{N+1}>0}\, b(\boldsymbol{X})}_{\textit{generalized } \beta\textit{-model}} \;\; > \;\; \underbrace{\mathbb{E}_{\theta_1,\ldots,\theta_N,\theta_{N+1}=0}\, b(\boldsymbol{X})}_{\beta\textit{-model}}, \tag{2.5}$$

where $b(\boldsymbol{X}) = \sum_{i<j}^{N} b_{i,j}(X_{i,j},\, \boldsymbol{X}_{\mathcal{S}_{i,j}})$ and $\mathbb{E}_{\theta_1,\ldots,\theta_N,\theta_{N+1}}\, b(\boldsymbol{X})$ is the expectation of $b(\boldsymbol{X})$ under $(\theta_1, \ldots, \theta_N, \theta_{N+1}) \in \mathbb{R}^{N+1}$. In other words, the generalized β-model with $\theta_{N+1} > 0$ generates graphs that have, on average, more brokered edges than the β-model, assuming that the propensities $\theta_1, \ldots, \theta_N$ of nodes $1, \ldots, N$ to form edges are the same under both models. The inequality in (2.5) follows from the fact that the generalized β-model is an exponential-family model, along with Corollary 2.5 of Brown [9, p. 37].

2.4. *Model 3: generalized β-models with sparsity.* Many real-world networks are sparse in the sense that few of the possible edges are present, which suggests that random graph models should place more mass on sparse graphs with few edges than dense graphs with many edges.

Sparse random graphs have been studied since the pioneering work of Erdős and Rényi [25] [e.g., 50, 51, 11, 64, 60, 61, 19]. More recently, Mukherjee et al. [60] and Chen et al. [19] developed sparse β-models with independent edges. We focus here on generalized β-models with dependent edges in sparse-graph settings. To develop sparse versions of generalized β-models, note that it makes sense to penalize edges among pairs of nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ that are distant in the sense that $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$, without penalizing edges among nodes that are close in the sense that $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$. We therefore induce sparsity by considering Model 2 but replacing (2.3) by

$$a_{i,j}(x_{i,j}) \quad = \quad \begin{cases} N^{-\alpha\, x_{i,j}\, \mathbb{1}(\mathcal{N}_i \,\cap\, \mathcal{N}_j = \emptyset)} & \text{if } x_{i,j} \in \{0, 1\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha \in \mathbb{R}^+ \cup \{0\}$ is a known constant, which may be interpreted as the level of sparsity of the random graph. If $\alpha = 0$, Model 3 reduces to Model 2, whereas $\alpha > 0$ implies that Model 3 imposes a penalty on edges among nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ with $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$, which increases with the number of nodes N. Thus, Model 3 with $\alpha > 0$ places less mass on graphs with edges among nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ satisfying $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$ than Model 3 with $\alpha = 0$ does. It is worth noting that the functions $a_{i,j}$ can be absorbed into the reference measure ν, so that the resulting reference measure places more mass on sparse than dense graphs.

2.5. *Model 4: generalized β-models with size-dependent parameterizations.*

Model 2 assumes that the brokerage parameter does not depend on the sizes of neighborhood intersections. While convenient on mathematical grounds, the assumption that small and large neighborhood intersections have the same brokerage parameter may be unwarranted. To allow small and large neighborhood intersections to have distinct brokerage parameters, we consider functions of edges $\varphi_{i,j}$ of the form

$$\varphi_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}}; \boldsymbol{\theta}_N) = a_{i,j}(x_{i,j}) \exp\left((\theta_i + \theta_j)\, x_{i,j} + \eta_{i,j}(\theta_{N+1})\, b_{i,j}(x_{i,j}, \boldsymbol{x}_{\mathcal{S}_{i,j}})\right),$$

where $a_{i,j}$ and $b_{i,j}$ are defined in (2.3) and (2.4), respectively, and

$$\eta_{i,j}(\theta_{N+1}) \quad = \quad \theta_{N+1} \log\left(1 + \frac{\log |\mathcal{N}_i \,\cap\, \mathcal{N}_j|}{|\mathcal{N}_i \,\cap\, \mathcal{N}_j|}\right), \qquad \theta_{N+1} \in \mathbb{R}. \tag{2.6}$$

Parameterization (2.6) is reminiscent of the size-dependent parameterization of Jonasson [47]. A sparse version of Model 4 can be constructed by changing the reference measure, as described in Section 2.4.

3. Consistency results and convergence rates.

We establish consistency results and convergence rates for pseudo-likelihood-based M-estimators for parameter vectors of increasing dimension based on a single observation of a random graph with dependent edges, in dense- and sparse-graph settings. We first present general consistency results and convergence rates for maximum likelihood and pseudo-likelihood-based M-estimators in Sections 3.2 and 3.3, and then present applications to generalized β-models with dependent edges in Section 3.4. To prepare the ground, we review in Section 3.1 how the dependence among edges and the smoothness of sufficient statistics can be quantified.

Throughout, we denote the data-generating parameter vector by $\boldsymbol{\theta}_N^\star$ and assume that

$$\boldsymbol{\theta}_N^\star \;\in\; \boldsymbol{\Theta}_N \;=\; \left\{\boldsymbol{\theta}_N \in \mathbb{R}^p :\; \int\limits_{\mathbb{X}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}')\, \mathrm{d}\,\nu(\boldsymbol{x}') \;<\; \infty\right\} \;\subseteq\; \mathbb{R}^p.$$

We denote the data-generating probability measure and expectation by $\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\theta}_N^\star}$ and $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\theta}_N^\star}$, respectively. To ease the presentation, we replace the double subscripts of edge variables by single subscripts and write $(X_m)_{1 \le m \le M}$ instead of $(X_{i,j})_{i<j:\, i \in \mathcal{N},\, j \in \mathcal{N}}$, where $M = \binom{N}{2}$.

3.1. *Controlling dependence and smoothness.* To obtain consistency results and convergence rates based on a single observation of a random graph with dependent edges, we need to control the dependence among edges along with the smoothness of the sufficient statistics of the model.

The dependence among edges can be controlled by bounding the total variation distance between conditional probability mass functions, which quantify how much the conditional probability mass functions of subgraphs are affected by changes of edges outside of the subgraphs. To do so, choose any $i \in \{1, \dots, M\}$ and define

$$\mathbb{P}_{\boldsymbol{x}_{1:i-1},0}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a}) \;\overset{\text{def}}{=}\; \mathbb{P}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1}, X_i = 0)$$

and

$$\mathbb{P}_{\boldsymbol{x}_{1:i-1},1}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a}) \;\overset{\text{def}}{=}\; \mathbb{P}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1}, X_i = 1),$$

where $\boldsymbol{X}_{1:i-1} = (X_1, \dots, X_{i-1})$, $\boldsymbol{X}_{i+1:M} = (X_{i+1}, \dots, X_M)$, and $\boldsymbol{a} \in \{0,1\}^{M-i}$. We compare these conditional probability mass functions in terms of total variation distance, which can be written as

$$\begin{aligned} & \|\mathbb{P}_{\boldsymbol{x}_{1:i-1},0} - \mathbb{P}_{\boldsymbol{x}_{1:i-1},1}\|_{\text{TV}} \\ = \;& \frac{1}{2} \sum_{\boldsymbol{a} \in \{0,1\}^{M-i}} |\mathbb{P}_{\boldsymbol{x}_{1:i-1},0}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a}) - \mathbb{P}_{\boldsymbol{x}_{1:i-1},1}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a})|. \end{aligned}$$

If the total variation distance is large, the conditional probability mass function of the subgraph in question is very sensitive to changes of edges outside of the subgraph, indicating strong dependence among edges. The total variation distance is intractable in all but the most trivial cases, but it can be bounded above by coupling. A coupling of two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ defined on a common measurable space $(\Omega, \mathcal{F})$ is a probability measure $\mathbb{Q}$ defined on $(\Omega^2, \mathcal{F}^2)$ such that $\mathbb{P}_1$ and $\mathbb{P}_2$ are marginals of $\mathbb{Q}$ [56, pp. 9–10]. The basic coupling inequality [56, pp. 9–12] implies that the total variation distance of conditional probability mass functions $\mathbb{P}_{\boldsymbol{x}_{1:i-1},0}$ and $\mathbb{P}_{\boldsymbol{x}_{1:i-1},1}$ is bounded above by

$$\|\mathbb{P}_{\boldsymbol{x}_{1:i-1},0} - \mathbb{P}_{\boldsymbol{x}_{1:i-1},1}\|_{\text{TV}} \leq \mathbb{Q}_{\boldsymbol{x}_{1:i-1}}(\boldsymbol{X}^\star \neq \boldsymbol{X}^{\star\star}), \quad (\boldsymbol{X}^\star, \boldsymbol{X}^{\star\star}) \in \Omega^2, \tag{3.1}$$

where $\Omega = \{0,1\}^{M-i}$, $\mathcal{F}$ is the set of all subsets of Ω, and $\mathbb{Q}_{\boldsymbol{x}_{1:i-1}}$ is a probability mass function on $(\Omega^2, \mathcal{F}^2)$ that couples the two conditional probability mass functions $\mathbb{P}_{\boldsymbol{x}_{1:i-1},0}$ and $\mathbb{P}_{\boldsymbol{x}_{1:i-1},1}$. A classic result in the literature on coupling [56, Theorem 5.2, p. 19] shows that, for any two probability measures defined on the same measurable space, there exists an optimal coupling satisfying

$$\|\mathbb{P}_{\boldsymbol{x}_{1:i-1},0} - \mathbb{P}_{\boldsymbol{x}_{1:i-1},1}\|_{\text{TV}} \quad = \quad \mathbb{Q}^{\text{opt}}_{\boldsymbol{x}_{1:i-1}}(\boldsymbol{X}^\star \neq \boldsymbol{X}^{\star\star}).$$

Optimal couplings are not unique, but any optimal coupling will do. A general approach to constructing optimal couplings is reviewed in Lindvall [56, pp. 99–107]. Armed with optimal couplings, we construct the coupling matrix $\mathcal{D}$ with elements $\mathcal{D}_{i,j}$ defined by

$$\mathcal{D}_{i,j} \quad = \quad \begin{cases} 0 & \text{if } j < i \\ 1 & \text{if } j = i \\ \max\limits_{\boldsymbol{x}_{1:i-1} \in \{0,1\}^{i-1}} \mathbb{Q}^{\text{opt}}_{i,\boldsymbol{x}_{1:i-1}}(X_j^\star \neq X_j^{\star\star}) & \text{if } j > i \end{cases}, \quad i, j = 1, \dots, M.$$

We use the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$ to quantify the dependence among edges.

In addition to the dependence among edges, we need to control the smoothness of the sufficient statistics of the model. Observe that the sufficient statistic vector $s : \mathbb{X} \mapsto \mathbb{R}^p$ of models with probability density functions parameterized by (2.1) and (2.2) is given by

$$s(\boldsymbol{x}) \quad = \quad \sum_{i=1}^{M} s_i(x_i, \boldsymbol{x}_{\mathcal{S}_i}), \qquad \boldsymbol{x} \in \mathbb{X},$$

where $\mathcal{S}_i \subset \{1, \ldots, M\}$. To quantify the smoothness of the sufficient statistics, we define, for each $i \in \{1, \ldots, p\}$,

$$\Xi_j \, s_i \;\; = \;\; \sup_{(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{X} \times \mathbb{X}:\; x_k = x_k' \text{ for all } k \neq j} |s_i(\boldsymbol{x}) - s_i(\boldsymbol{x}')|, \qquad j = 1, \ldots, M.$$

We write $\Xi \, s_i = (\Xi_1 \, s_i, \ldots, \Xi_M \, s_i)$ $(i = 1, \ldots, p)$ and define

$$\Psi \;\; = \;\; \max_{1 \leq i \leq p} \|\Xi \, s_i\|_2.$$

The main results on maximum likelihood and pseudo-likelihood-based M-estimators in Sections 3.2 and 3.3 are stated in terms of $|||\mathcal{D}|||_2$ and Ψ. In applications, $|||\mathcal{D}|||_2$ and Ψ need to be bounded. Corollaries 1–4 in Section 3.4 demonstrate how $|||\mathcal{D}|||_2$ and Ψ can be bounded in applications to the β-model and generalized β-models with dependent edges.

3.2. *Maximum likelihood estimators.* We start with maximum likelihood estimators, because the general idea of how convergence rates can be obtained is best explained by using maximum likelihood estimators. These results may be of independent interest, and serve as a convenient stepping stone for the convergence rates for pseudo-likelihood-based M-estimators presented in Section 3.3.

Maximum likelihood estimators, based on a single observation $\boldsymbol{x}$ of a random graph $\boldsymbol{X}$ with dependent edges, are defined by

$$\widehat{\boldsymbol{\theta}}_N \;\; = \;\; \left\{ \boldsymbol{\theta}_N \, \in \, \boldsymbol{\Theta}_N : \; \ell(\boldsymbol{\theta}_N; \, \boldsymbol{x}) \; = \; \sup_{\dot{\boldsymbol{\theta}}_N \, \in \, \boldsymbol{\Theta}_N} \; \ell(\dot{\boldsymbol{\theta}}_N; \, \boldsymbol{x}) \right\},$$

where

$$\ell(\boldsymbol{\theta}_N; \, \boldsymbol{x}) \;\; = \;\; \log f_{\boldsymbol{\theta}_N}(\boldsymbol{x}).$$

Choose $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \, \epsilon) \subset \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ and assume that there exist $C > 0$, $N_0 > 0$, and $\Lambda : \{1, 2, \ldots\} \mapsto \mathbb{R}^+$ such that

$$\inf_{\boldsymbol{\theta}_N \, \in \, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \, \epsilon)} \; \sqrt[p]{\det(\mathcal{I}(\boldsymbol{\theta}_N))} \;\; \geq \;\; C \, \Lambda(N) \;\; > \;\; 0 \quad \text{for all} \quad N > N_0,$$

where $\mathcal{I}(\boldsymbol{\theta}_N) = -\mathbb{E} \, \nabla^2_{\boldsymbol{\theta}_N} \, \ell(\boldsymbol{\theta}_N; \, \boldsymbol{X})$ is the Fisher information matrix. The p-th root determinant $\sqrt[p]{\det(\mathcal{I}(\boldsymbol{\theta}_N))}$ of $\mathcal{I}(\boldsymbol{\theta}_N)$ is bounded below by the smallest eigenvalue $\lambda_{\min}(\boldsymbol{\theta}_N)$ of $\mathcal{I}(\boldsymbol{\theta}_N)$, so $\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)} \lambda_{\min}(\boldsymbol{\theta}_N) \; \geq \; C \, \Lambda(N)$ implies $\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)} \sqrt[p]{\det(\mathcal{I}(\boldsymbol{\theta}_N))} \geq C \, \Lambda(N)$. For convenience, we henceforth write Λ rather than $\Lambda(N)$.

We then obtain the following consistency results and convergence rates.

Theorem 1. *Consider a single observation of a random graph with N nodes and $\binom{N}{2}$ dependent edges. Let $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ and $p \to \infty$ as $N \to \infty$. Assume that, for any $C_0 > 0$, however large, there exists $N_0 > 0$ such that*

$$\Lambda \;\; > \;\; C_0 \, |||\mathcal{D}|||_2 \, \Psi \, \sqrt{\log p} \quad \textit{for all} \quad N > N_0.$$

Then there exist universal constants $A > 0$ and $C > 0$ such that, with at least probability $1 - 2\,\exp(-A\,\log p)$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ exists, is unique, and satisfies

$$\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\; \leq \;\; C \, |||\mathcal{D}|||_2 \sqrt{\frac{\log p}{\Lambda^2 \,/\, \Psi^2}} \;\; \longrightarrow \;\; 0 \quad \textit{as} \quad N \longrightarrow \infty.$$

The rate of growth Λ of the p-th root determinant of the Fisher information matrix in a neighborhood of $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ may be interpreted as the effective sample size, similar to Krivitsky and Kolaczyk [51] and Chen et al. [19]. In other words, Theorem 1 implies that $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty$ converges in probability to 0 as $N \to \infty$ provided that the effective sample size Λ grows faster than $|||\mathcal{D}|||_2 \, \Psi \, \sqrt{\log p}$. We show in Corollary 1 in Section 3.4 that in the special case of the β-model, with at least probability $1 - 2\,\exp(-A\,\log N)$, $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty$ is bounded above by $C\,\sqrt{\log N \,/\, N}$ provided N is large enough, which agrees with the result of Chatterjee et al. [17]. In contrast to the work of Chatterjee et al. [17] and others [61, 68], Theorem 1 is neither limited to models with independent edges [17] nor models that constrain the dependence among edges to non-overlapping subpopulations of nodes [68].

Theorem 1 is based on a novel argument that helps establish convergence rates in scenarios where a single observation of dependent random variables is available and the number of parameters increases without bound. A simple observation shows that the map $\boldsymbol{\mu} : \boldsymbol{\Theta}_N \mapsto \mathbb{R}^p$ defined by $\boldsymbol{\mu}(\boldsymbol{\theta}_N) = \mathbb{E}_{\boldsymbol{\theta}_N}\, s(\boldsymbol{X})$ is a homeomorphism and hence, for all $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon) \subset \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$, there exists $\delta(\epsilon) > 0$ such that $\boldsymbol{\mu}(\boldsymbol{\theta}_N) \in \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$ implies $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$. A second, more important insight reveals that the radius $\delta(\epsilon)$ of the ℓ_∞-ball $\mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$ can be related to the radius ϵ of the ℓ_∞-ball $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)$ as follows:

$$\delta(\epsilon) \;\; \geq \;\; \epsilon\, B \inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)} \sqrt[p]{\det(\mathcal{I}(\boldsymbol{\theta}_N))} \;\; \geq \;\; \epsilon\, C\, \Lambda,$$

where $B > 0$ and $C > 0$ are constants. An immediate consequence is that the convergence rate depends on Λ. In addition, the convergence rate depends on how the probability mass of $s(\boldsymbol{X})$ concentrates in the ℓ_∞-ball

$\mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$. In general, the strength of concentration depends on the dependence among edges in terms of $|||\mathcal{D}|||_2$ and the smoothness of $s(\boldsymbol{X})$ in terms of Ψ, so the convergence rate depends on Λ, $|||\mathcal{D}|||_2$, Ψ, and p.

3.3. *Pseudo-likelihood-based M-estimators.*

Maximum likelihood estimators are a useful benchmark, but computing them is infeasible unless the number of edge variables $M = \binom{N}{2}$ is small. The reason is that the normalizing constant of probability density function (2.1) is a sum over $\exp(M \log 2)$ possible graphs, which cannot be computed by complete enumeration of all $\exp(M \log 2)$ possible graphs unless M is small. At the same time, it is not straightforward to simplify the sum of $\exp(M \log 2)$ possible graphs due to the dependence among edges, and there are no general-purpose approximations with statistical guarantees, leaving aside variational approximations of ill-posed models with a fixed and finite number of parameters [16]. Markov chain Monte Carlo approximations of maximum likelihood estimators can be computed [44], but are expensive in terms of computing time [5, 16].

To facilitate scalable estimation of random graph models with dependent edges, consider replacing the loglikelihood function by the pseudo-loglikelihood function [4, 28, 71], defined by

$$\widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) \;\; = \;\; \log \prod_{i=1}^{M} f_{\boldsymbol{\theta}_N}(x_i \mid \boldsymbol{x}_{-i}),$$

where $f_{\boldsymbol{\theta}_N}(x_i \mid \boldsymbol{x}_{-i})$ denotes the conditional probability of $X_i = x_i$ given $\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}$, that is, the values of all other edge variables $(i = 1, \ldots, M)$. On computational grounds, maximizing $\widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ is preferable to maximizing $\ell(\boldsymbol{\theta}_N;\, \boldsymbol{x})$, because computing $\widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ and its gradient at parameter vectors $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ requires computations of M terms rather than $\exp(M \log 2)$ terms. Despite the computational advantages of maximum pseudo-likelihood estimators, the statistical properties of maximum pseudo-likelihood estimators in single-observation scenarios with $p \to \infty$ parameters are unknown.

That said, not all is lost. The following lemma suggests that maximum pseudo-likelihood estimators do hold promise, even in single-observation scenarios with $p \to \infty$ parameters.

Lemma 1. *The set $\boldsymbol{\Theta}_N$ is a convex set and $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ is a strictly concave function on the convex set $\boldsymbol{\Theta}_N$. In addition,*

$$\boldsymbol{\theta}_N^\star \;\; = \;\; \underset{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N}{\arg\max}\; \mathbb{E}\, \ell(\boldsymbol{\theta}_N;\, \boldsymbol{X}) \;\; = \;\; \underset{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N}{\arg\max}\; \mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) \;\in\; \boldsymbol{\Theta}_N \;\subseteq\; \mathbb{R}^p.$$

Thus, if we could compute the expected loglikelihood and pseudo-loglikelihood function and wanted to estimate the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$, we could obtain $\boldsymbol{\theta}_N^\star$ by maximizing either $\mathbb{E}\, \ell(\boldsymbol{\theta}_N; \boldsymbol{X})$ or $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$, which is equivalent to minimizing

$$\boldsymbol{g}(\boldsymbol{\theta}_N) \;=\; \|\nabla_{\boldsymbol{\theta}_N} \mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})\|_\infty.$$

In practice, assuming $\boldsymbol{g}(\boldsymbol{\theta}_N)$ can be evaluated at all parameter vectors $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$, it would be natural to use a root-finding algorithm to approximate the root of $\boldsymbol{g}(\boldsymbol{\theta}_N)$, choose a convergence criterion $\gamma > 0$, and stop as soon as the root-finding algorithm returns a parameter vector $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ satisfying $\|\boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;\leq\; \gamma$. In other words, it would be natural to report an element of the non-random set

$$\dot{\boldsymbol{\Theta}}_N \;=\; \{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N :\; \|\boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;\leq\; \gamma\}$$

as an educated guess of $\boldsymbol{\theta}_N^\star$. By construction, the set $\dot{\boldsymbol{\Theta}}_N$ contains the data-generating parameter vector $\boldsymbol{\theta}_N^\star$.

It goes without saying that $\boldsymbol{g}(\boldsymbol{\theta}_N) = \nabla_{\boldsymbol{\theta}_N} \mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$—the gradient of the expected pseudo-loglikelihood function with respect to the data-generating probability measure—cannot be computed in practice, but it can be approximated by $\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) = \nabla_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ based on an observation $\boldsymbol{x}$ of a random graph $\boldsymbol{X}$. While replacing the non-random function $\boldsymbol{g}(\boldsymbol{\theta}_N)$ by the random function $\boldsymbol{g}(\boldsymbol{\theta}_N; \boldsymbol{X})$ comes at a cost, the random set

$$\widetilde{\boldsymbol{\Theta}}_N \;=\; \{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N :\; \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X})\|_\infty \;\leq\; \gamma\}$$

inherits the desirable properties of the non-random set $\dot{\boldsymbol{\Theta}}_N$. To demonstrate, observe that

$$\|\boldsymbol{g}(\boldsymbol{\theta}_N^\star;\, \boldsymbol{X})\|_\infty \;=\; \|\boldsymbol{g}(\boldsymbol{\theta}_N^\star;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N^\star)\|_\infty \;\leq\; \sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty.$$

Thus, as long as the event

$$\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;\;\leq\;\; \gamma \tag{3.2}$$

occurs with high probability, the random set $\widetilde{\boldsymbol{\Theta}}_N$ contains the data-generating parameter vector $\boldsymbol{\theta}_N^\star$ with high probability. In other words, the random set $\widetilde{\boldsymbol{\Theta}}_N$ possesses desirable properties by construction, and it takes nothing more than two inequalities to demonstrate them.

The random set $\widetilde{\boldsymbol{\Theta}}_N$ can be viewed as an M-estimator [74] and generalizes maximum pseudo-likelihood estimators. If the random set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty, it consists of the maximizers of the pseudo-likelihood function when $\gamma = 0$, and maximizers as well as near-maximizers when $\gamma > 0$. For convenience, we refer to elements $\widetilde{\boldsymbol{\theta}}_N$ of the random set $\widetilde{\boldsymbol{\Theta}}_N$ as maximum pseudo-likelihood estimators, despite the fact that some elements are maximizers while others are near-maximizers.

To establish consistency results and convergence rates for maximum pseudo-likelihood estimators, we establish a uniform convergence result of the form (3.2), and demonstrate them in applications to generalized β-models with dependent edges and parameter vectors of increasing dimension. Such results—in single-observation scenarios with $p \to \infty$ parameters—are non-trivial, but it turns out that some of the ideas mentioned in Section 3.2 help establish them. Choose $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon) \subset \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ and assume that there exist $C > 0$, $N_0 > 0$, and $\widetilde{\Lambda} : \{1, 2, \ldots\} \mapsto \mathbb{R}^+$ such that

$$\inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \sqrt[p]{|\det(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_N}\; \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))|} \;\geq\; C\, \widetilde{\Lambda}(N) \;>\; 0 \text{ for all } N > N_0.$$

We write henceforth $\widetilde{\Lambda}$ instead of $\widetilde{\Lambda}(N)$ and assume that each edge variable depends on a finite number of other edge variables, which resembles the finite neighborhood assumptions in the single-observation literature on Ising models, Gibbs measures, and Markov random fields [20, 73, 32, 22, 34].

Assumption A. *For each* $i \in \{1, \ldots, M\}$*, there exists a subset* $\mathfrak{N}_i \subset \{1, \ldots, M\} \setminus \{i\}$ *such that*

$$X_i \;\perp\!\!\!\perp\; \boldsymbol{X} \setminus (X_i, \boldsymbol{X}_{\mathfrak{N}_i}) \mid \boldsymbol{X}_{\mathfrak{N}_i},$$

and there exists a constant $C > 0$ *such that* $\max_{1 \leq i \leq M} |\mathfrak{N}_i| < C$ *(*$M = 1, 2, \ldots$*).*

It is challenging, but not impossible to dispense with Assumption A, because we do not have independent replications, and we need to control the dependence among edges to obtain concentration and consistency results, as the single-observation literature on Ising models, Gibbs measures, and Markov random fields does [20, 73, 32, 22, 34]. We provide a more elaborate discussion of Assumption A in Remark 4 following Theorem 2.

Armed with Assumption A, we obtain the following consistency results and convergence rates. In contrast to the single-observation literature on Markov random fields and Ising models [e.g., 35, 46, 20, 58, 45, 15, 6, 34], which is limited to parameter vectors of fixed and finite dimension p, we consider parameter vectors of dimension $p \to \infty$.

Theorem 2. *Consider a single observation of a random graph with* N *nodes and* $\binom{N}{2}$ *dependent edges satisfying Assumption A. Let* $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$

and $p \to \infty$ as $N \to \infty$. Assume that, for any $C_0 > 0$, however large, there exists $N_0 > 0$ such that

$$\widetilde{\Lambda} \;\;>\;\; C_0\, |||\mathcal{D}|||_2\, \Psi\, \sqrt{\log p} \quad \textit{for all} \quad N > N_0.$$

Then there exist universal constants $A > 0$, $B > 0$, and $C > 0$ such that, for all $\gamma \in (0,\, B\, |||\mathcal{D}|||_2\, \Psi\, \sqrt{\log p})$, with at least probability $1 - 2\, \exp(-A \log p)$, the random set $\widetilde{\mathbf{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\mathbf{\Theta}}_N$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\;\leq\;\; C\, |||\mathcal{D}|||_2 \sqrt{\frac{\log\, p}{\widetilde{\Lambda}^2\, /\, \Psi^2}} \;\;\longrightarrow\;\; 0 \quad \textit{as} \quad N \longrightarrow \infty.$$

While Theorem 2 is stated in terms of random graphs with $M = \binom{N}{2}$ dependent edges, Theorem 2 is a general result on maximum pseudo-likelihood estimators of $p \to \infty$ parameters based on a single observation of $M \to \infty$ dependent random variables with finite sample spaces, and could be extended to countable and uncountable sample spaces [see the supplement, 70]. To the best of our knowledge, Theorem 2 is the first result on maximum pseudo-likelihood estimators of $p \to \infty$ parameters based on a single observation of $M \to \infty$ dependent random variables. Theorem 2 demonstrates that scalable statistical learning with statistical guarantees in single-observation scenarios with $p \to \infty$ parameters is possible in great generality—at least within the large class of strictly positive distributions with exponential parameterizations, which includes discrete Markov random fields in spatial statistics and Boltzmann machines in artificial intelligence. While such results were realized in special cases in spatial statistics and related areas [35, 46, 20, 58, 45, 15, 6, 34], the specialized nature of those results (focusing on, e.g., models with lattice structure, pairwise interactions among random variables, and fixed and finite p) have concealed the general nature of the ideas. Theorem 2 suggests that statistical learning in single-observation scenarios with $p \to \infty$ parameters is not limited to models with lattice structure and pairwise interactions among random variables, but is possible as long as the dependence is not too strong and the amount of information is not too low. Having said that, there is no such thing as a free lunch: The finite neighborhood assumption, in the form of Assumption A, is hard to drop—as discussed in Remark 4—and strong dependence and low information can make estimation challenging, if not impossible.

Specific examples of consistency results and convergence rates are presented in Corollaries 1–4 in Section 3.4. These examples include the β-model and generalized β-models with $M = \binom{N}{2}$ dependent edges and $p \geq N$ parameters. In other words, these results cover random graphs models with either independent or dependent edges, in dense- and sparse-graph settings.

Remark 4. Finite neighborhood assumption. Assumption A of Theorem 2 assumes that the conditional distribution of each edge variable depends on a finite number of other edge variables, an assumption that resembles the finite neighborhood assumptions in the single-observation literature on Ising models, Gibbs measures, and Markov random fields [20, 73, 32, 22, 34]: e.g., a Markov random field with a two-dimensional lattice and pairwise nearest-neighbor interactions assumes that the conditional distribution of each random variable depends on random variables located at neighboring vertices of the lattice, so the conditional distribution of each random variable depends on a finite number of other random variables [4]. In the related literature on high-dimensional Ising models [63, 3, 78, 8] and other high-dimensional graphical models [59], it is known that when n independent observations of M dependent random variables are available, the sizes of the neighborhoods of random variables in the underlying conditional independence graph can increase as a function of n. By contrast, we consider scenarios with a single observation of $M \to \infty$ dependent random variables and $p \to \infty$ parameters. In such settings, one needs to control the sizes of neighborhoods to ensure that the dependence among random variables is weak enough to obtain concentration and consistency results. While it is possible to drop the finite neighborhood assumption, it is not straightforward to replace it by verifiable assumptions. The recent work of Ghosal and Mukherjee [34], concerned with conditions under which consistent estimation of the two-parameter Ising model is possible, points into the same direction.

3.4. *Applications: generalized β-models with dependent edges.* We demonstrate how consistency results and convergence rates for maximum pseudo-likelihood estimators of $p \to \infty$ parameters based on a single observation of a random graph with $M = \binom{N}{2} \to \infty$ dependent edges can be obtained, assuming that the number of nodes N grows without bound ($N \to \infty$). As examples, we use generalized β-models with $M = \binom{N}{2}$ dependent edges and $p \geq N$ parameters. We focus here on maximum pseudo-likelihood estimators, because maximum pseudo-likelihood estimators are more scalable than maximum likelihood estimators.

We first consider the β-model of Chatterjee et al. [17], studied by Rinaldo et al. [64], Yan and Xu [81], Karwa and Slavković [48], Mukherjee et al. [60], and Chen et al. [19], among others. While we are more interested in generalized β-models with dependent edges than the β-model with independent edges, the β-model serves as a useful reference point, because consistency results and convergence rates for maximum likelihood estimators have been established in the special case of the β-model. The following corollary of Theorems 1 and 2 shows that the general convergence rates reported in Theorems 1 and 2 agree with the convergence rate reported by Chatterjee et al. [17, Theorem 1.3] in the special case of the β-model. Throughout, we assume that the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ satisfies

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\leq\; U + \frac{1-\vartheta}{8}\log N, \tag{3.3}$$

where $U > 0$ and $\vartheta \in (1/2,\, 1]$ are constants.

Corollary 1. *Consider Model 1, the β-model with independent edges, and assume that the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^N$ satisfies (3.3). Let $\gamma = 0$. Then there exist universal constants $A > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$, with at least probability $1 - 2\exp(-A\log N)$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ and the maximum pseudo-likelihood estimator $\widetilde{\boldsymbol{\theta}}_N$ exist, are unique, and satisfy*

$$\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;=\; \|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\leq\; C\sqrt{\frac{\log N}{N^{2\vartheta-1}}}.$$

Thus $\vartheta \in (1/2,\, 1]$ implies $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty = \|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \xrightarrow{p} 0$ as $N \longrightarrow \infty$.

The identity $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty = \|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty$ follows from the independence of edges under the β-model, which implies that $\widehat{\boldsymbol{\theta}}_N = \widetilde{\boldsymbol{\theta}}_N$ with probability 1. Corollary 1 recovers the convergence rate of Chatterjee et al. [17] in the special case of the β-model: If $\|\boldsymbol{\theta}_N^\star\|_\infty$ is bounded above by a finite constant (which requires $\vartheta = 1$), then $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty$ is bounded above by $C\sqrt{\log N \,/\, N}$ with high probability, provided N is large enough. However, we deduce these results from Theorems 1 and 2, which are much more general than the results of Chatterjee et al. [17] and cover maximum pseudo-likelihood estimators of random graph models with dependent edges and parameter vectors of increasing dimension. To demonstrate the flexibility of Theorems 1 and 2, we turn to generalized β-models with dependent edges.

Corollary 2. *Consider Model 2, the generalized β-model with dependent edges, and assume that the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfies (3.3). Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ and $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ ($D \geq 2$). Then there exist universal constants $A > 0$, $B > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$ and all $\gamma \in (0,\, B\, |||\mathcal{D}|||_2 \sqrt{N \log N})$, with at least probability $1 - 2\exp(-A\log N)$, the random set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$ satisfies*

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\leq\; C\, |||\mathcal{D}|||_2 \sqrt{\frac{\log N}{N^{2\vartheta-1}}}.$$

If, for any $\epsilon > 0$, however small, the coupling matrix $\mathcal{D}$ satisfies

$$|||\mathcal{D}|||_2 \;<\; \epsilon \sqrt{\frac{N^{2\,\vartheta-1}}{\log N}} \quad \textit{for all} \quad N > N_0, \tag{3.4}$$

then $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \xrightarrow{p} 0$ *as* $N \longrightarrow \infty$.

Corollary 2 reveals two remarkable insights into the behavior of maximum pseudo-likelihood estimators for random graph models with dependent edges and parameter vectors of increasing dimension. First, if $|||\mathcal{D}|||_2$ is bounded above by a finite constant, generalized β-models with dependent edges have the same convergence rate as the β-model with independent edges. In other words, dependence does not come at a cost in terms of the convergence rate as long as the dependence is weak enough. Second, $|||\mathcal{D}|||_2$ can grow as a function of N provided that it grows slower than $\sqrt{N^{2\,\vartheta-1} /\, \log N}$ ($\vartheta \in (1/2,\, 1]$). In fact, if $\|\boldsymbol{\theta}_N^\star\|_\infty$ is bounded above by a finite constant (which requires $\vartheta = 1$), then maximum pseudo-likelihood estimators are consistent as long as $|||\mathcal{D}|||_2$ grows slower than $\sqrt{N \,/\, \log N}$.

It is natural to ask under which conditions $|||\mathcal{D}|||_2$ satisfies (3.4). In general, we need to control the dependence among edges to bound $|||\mathcal{D}|||_2$. We outline two possible restrictions on the subpopulation structure that help control the dependence among edges. To state them, it is helpful to introduce a subpopulation graph $\mathcal{G}_\mathcal{A}$ with set of vertices $\mathcal{V}_\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ consisting of subpopulations $\mathcal{A}_1, \ldots, \mathcal{A}_K$. The subpopulation graph $\mathcal{G}_\mathcal{A}$ contains an edge between two distinct subpopulations $\mathcal{A}_k$ and $\mathcal{A}_l$ if and only if $\mathcal{A}_k \cap \mathcal{A}_l \neq \emptyset$. Denote by $d_\mathcal{G} : \mathcal{V}_\mathcal{A} \times \mathcal{V}_\mathcal{A} \mapsto \{0, 1, \ldots\} \cup \{\infty\}$ the graph distance between two distinct subpopulations, defined as the length of the shortest path between the two subpopulations in $\mathcal{G}_\mathcal{A}$. If there is no path of finite length between two subpopulations, the graph distance is ∞. Let $\mathcal{V}_{\mathcal{A}_k,l}$ be the set of subpopulations at graph distance l from $\mathcal{A}_k$ in $\mathcal{G}_\mathcal{A}$:

$$\mathcal{V}_{\mathcal{A}_k,l} \;=\; \{\mathcal{A}^\star \in \{\mathcal{A}_1, \ldots, \mathcal{A}_K\} \setminus \{\mathcal{A}_k\} :\; d_{\mathcal{G}_\mathcal{A}}(\mathcal{A}_k,\, \mathcal{A}^\star) = l\}\,,$$

and define $\pi^\star \in (0, 1)$ by

$$\pi^\star \;\overset{\text{def}}{=}\; \max_{1 \leq i \leq M} \; \max_{\boldsymbol{x}_{-i} \,\in\, \{0,1\}^{M-1}} \; \mathbb{P}(X_i = 1 \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}).$$

The following assumption places restrictions on the subpopulation structure through the subpopulation graph $\mathcal{G}_\mathcal{A}$.

Assumption B. *Assume that the subpopulation structure satisfies one of the two following conditions:*

B.1 Assume that

$$\max_{1 \leq k \leq K} |\mathcal{V}_{\mathcal{A}_k,l}| \;\; \leq \;\; \omega_1 + \frac{\omega_2}{8\, D^2} \log l, \qquad l \in \{1, 2, \dots\},$$

where $0 \leq \omega_1 < \infty$ *and* $0 \leq \omega_2 < 1 \,/\, |\log(1 - \pi^\star)|$.

B.2 Assume that the subpopulation graph $\mathcal{G}_{\mathcal{A}}$ *is a tree (i.e., an undirected, connected graph without cycles) and that*

$$\max_{1 \leq k \leq K} |\mathcal{V}_{\mathcal{A}_k,l}| \;\; \leq \;\; g_{\mathcal{A}}(l), \qquad l \in \{1, 2, \dots\},$$

where $g_{\mathcal{A}}(l)$ *is subexponential in the sense that, for all* $\epsilon > 0$*, however small, there exists* $l_0(\epsilon) > 0$ *such that* $\log\, g_{\mathcal{A}}(l) \,/\, l \leq \epsilon$ *for all* $l \geq l_0(\epsilon)$.

Assumption B restricts the dependence among edges by placing restrictions on the subpopulation structure, without constraining the dependence among edges to non-overlapping subpopulations of nodes. For instance, both assumptions allow the subpopulations to form a chain in the sense that there exists a permutation $\rho(1), \dots, \rho(K)$ of $1, \dots, K$ such that $\mathcal{A}_{\rho(k)} \cap \mathcal{A}_{\rho(k+1)} \neq \emptyset$ $(k = 1, \dots, K-1)$. Chains of subpopulations are simple examples of models where each subpopulation overlaps with one or more other subpopulations, and the dependence among edges is not constrained to non-overlapping subpopulations of nodes.

Assumptions B.1 and B.2 place two distinct, but related restrictions on the subpopulation structure. Assumption B.1 restricts the overlap of subpopulations by constraining the sizes of the sets $\mathcal{V}_{\mathcal{A}_k,l}$, that is, the number of subpopulations at graph distance $l \in \{1, 2, \dots\}$. It does not, however, impose additional restrictions on the overlap of subpopulations. By contrast, Assumption B.2 assumes that the subpopulation graph forms a tree and hence does constrain the overlap of subpopulations, because each subpopulation can overlap with at most two other subpopulations and there cannot be cycles in the subpopulation graph. By constraining the amount of overlap, Assumption B.2 is able to allow $\max_{1 \leq k \leq K} |\mathcal{V}_{\mathcal{A}_k,l}|$ to grow faster than Assumption B.1. Indeed, while Assumption B.1 restricts $\max_{1 \leq k \leq K} |\mathcal{V}_{\mathcal{A}_k,l}|$ to grow at a logarithmic rate, Assumption B.2 allows $\max_{1 \leq k \leq K} |\mathcal{V}_{\mathcal{A}_k,l}|$ to grow at a subexponential rate.

The following corollaries assume that Assumption B is satisfied and that $\vartheta = 1$, that is, $\|\boldsymbol{\theta}_N^\star\|_\infty$ is bounded above by a finite constant. The boundedness assumption helps demonstrate how high the convergence rate can be in the best-case scenario, and simplifies results. It can be removed, but doing so comes at a cost in terms of the convergence rate.

Corollary 3. *Consider Model 3, the generalized β-model with dependent edges and known level of sparsity $\alpha \in [0,\, 1/2)$, and assume that the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfies (3.3) with $\vartheta = 1$. Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$, $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ $(D \geq 2)$, and Assumption B is satisfied. Then there exist universal constants $A > 0$, $B > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$ and all $\gamma \in (0,\, B\, \sqrt{N \log N})$, with at least probability $1 - 2\, \exp(-A\, \log N)$, the random set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$ satisfies*

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\; \leq \;\; C\, \sqrt{\frac{\log N}{N^{1-2\,\alpha}}}.$$

Thus $\alpha \in [0,\, 1/2)$ implies $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \xrightarrow{p} 0$ as $N \longrightarrow \infty$.

Corollary 3 supplies the first consistency result and convergence rate for random graph models with dependent edges and parameter vectors of increasing dimension in sparse-graph settings. While there are consistency results for sparse random graphs with independent edges and parameter vectors of increasing dimension [e.g., 19] and sparse random graphs with dependent edges and two-dimensional parameter vectors [61], we are not aware of consistency results for sparse random graphs with dependent edges and parameter vectors of increasing dimension. Corollary 3 shows that sparsity comes at a cost: The higher the level of sparsity α is, the greater the statistical error of the maximum pseudo-likelihood estimator can be.

Corollary 4. *Consider Model 4, the generalized β-model with dependent edges and size-dependent parameterization, and assume that the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfies (3.3) with $\vartheta = 1$. Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$, $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ $(D \geq 2)$, and Assumption B is satisfied. Then there exist universal constants $A > 0$, $B > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$ and all $\gamma \in (0,\, B\, \sqrt{N \log N})$, with at least probability $1 - 2\, \exp(-A\, \log N)$, the random set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$ satisfies*

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\; \leq \;\; C\, \sqrt{\frac{\log N}{N}} \longrightarrow 0 \quad \textit{as} \quad N \longrightarrow \infty.$$

Corollary 4 demonstrates that theoretical guarantees can be obtained for models with size-dependent parameterizations, which allow parameters to depend on the sizes of neighborhoods.

4. Simulation study.

We complement the theoretical results in Section 3 by simulation results.

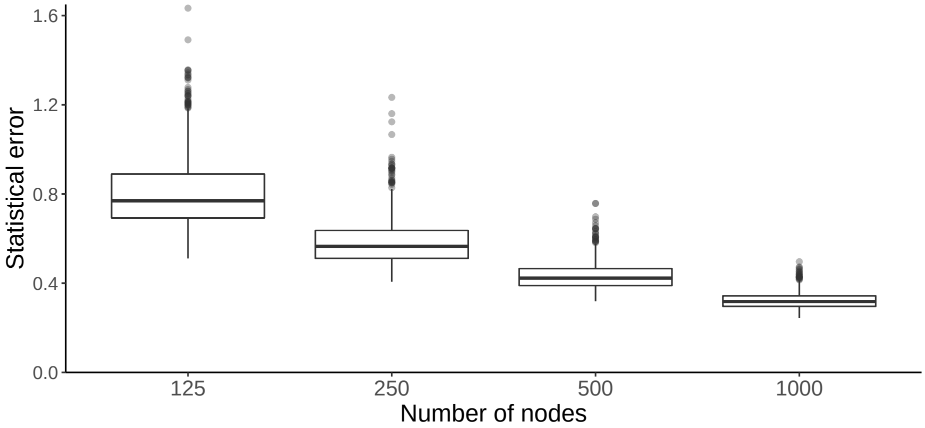

FIG 2. *The statistical error $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty$ of maximum pseudo-likelihood estimator $\widetilde{\boldsymbol{\theta}}_N$ as an estimator of $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ plotted against the number of nodes N.*

We study the performance of maximum pseudo-likelihood estimators by considering populations with $N = 125$, 250, 500, and 1,000 nodes. We focus on maximum pseudo-likelihood estimators, because computing maximum likelihood and Monte Carlo maximum likelihood estimators is too time-consuming when N is large (e.g., when $N = 500$ and $N = 1{,}000$). For each value of N, we generate 1,000 populations with overlapping subpopulations as follows:

- The number of subpopulations K is $N \,/\, 25$.
- Each node $i \in \mathcal{N}$ belongs to $1 + Y_i$ subpopulations, where $Y_i \overset{\text{iid}}{\sim} \text{Binomial}(K-1,\; 1\,/\,K)$ $(i = 1, \dots, N)$.
- For node $i = 1, \dots, N$, the $1 + Y_i$ subpopulation memberships are sampled from the Multinomial$(p_1^{(i)}, \dots, p_K^{(i)})$ distribution with

$$p_k^{(i)} \;=\; \begin{cases} \dfrac{1}{K} & \text{if } i = 1 \\[2ex] \dfrac{1}{K-1}\left(1 - \dfrac{N_k^{(i-1)}}{N_1^{(i-1)} + \ldots + N_K^{(i-1)}}\right) & \text{if } i \in \{2, \dots, N\}, \end{cases}$$

  where $N_k^{(i-1)}$ is the number of nodes in $\{1, \dots, i-1\}$ that belong to subpopulation $\mathcal{A}_k$ $(k = 1, \dots, K)$.

We consider Model 2 with degree parameters $\theta_1^\star, \dots, \theta_N^\star$ drawn from Uniform$(-1.25,\, -.75)$ and brokerage parameter $\theta_{N+1}^\star = .25$. For each value of N and each population of size N, we generate a graph from Model 2

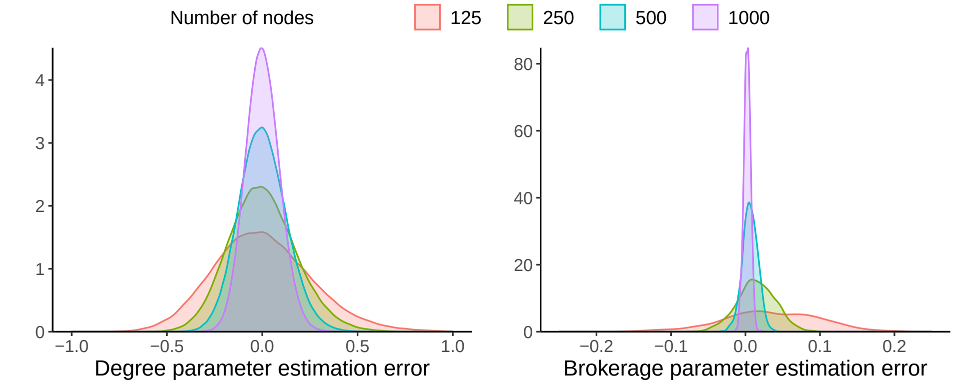

Fig 3. *The maximum deviation* $\max_{1 \leq i \leq N} |\widetilde{\theta}_i - \theta_i^\star|$ *of the maximum pseudo-likelihood estimators* $\widetilde{\theta}_i$ *from the data-generating degree parameters* $\theta_i^\star$ $(i = 1, \ldots, N)$ *(left) and the deviation* $|\widetilde{\theta}_{N+1} - \theta_{N+1}^\star|$ *of the maximum pseudo-likelihood estimator* $\widetilde{\theta}_{N+1}$ *from the data-generating brokerage parameter* $\theta_{N+1}^\star$ *(right).*

and compute the maximum pseudo-likelihood estimator from the generated graph. For each value of N, the gradient descent algorithm used to compute the maximum pseudo-likelihood estimator converged for at least 95% of the simulated data sets, and the following simulation results are based on the simulated data sets for which the gradient descent algorithm converged.

Figure 2 demonstrates that the statistical error $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty$ of $\widetilde{\boldsymbol{\theta}}_N$ as an estimator of the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ decreases as the number of nodes N increases. Figure 3 decomposes the statistical error of $\widetilde{\boldsymbol{\theta}}_N$ into the statistical error of the degree parameter estimators $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_N$ and the statistical error of the brokerage parameter estimator $\widetilde{\theta}_{N+1}$. Figure 3 reveals that the brokerage parameter is estimated with greater accuracy than the degree parameters, which makes sense as the degree parameters are greater in absolute value than the brokerage parameter.

Supplementary materials. Theorems 1 and 2 and Corollaries 1–4 are proved in the supplement [70].

References.

[1] Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008), "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014.

[2] Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, 41, 2097–2122.

[3] Anandkumar, A., Tan, V. Y. F., Huang, F., and Willsky, A. S. (2012), "High-

dimensional structure estimation in Ising models: Local separation criterion," *The Annals of Statistics*, 40, 1346–1375.

[4] Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, 36, 192–225.

[5] Bhamidi, S., Bresler, G., and Sly, A. (2011), "Mixing time of exponential random graphs," *The Annals of Applied Probability*, 21, 2146–2170.

[6] Bhattacharya, B. B., and Mukherjee, S. (2018), "Inference in Ising models," *Bernoulli*, 24, 493–525.

[7] Bickel, P. J., and Chen, A. (2009), "A nonparametric view of network models and Newman-Girvan and other modularities," in *Proceedings of the National Academy of Sciences*, Vol. 106, pp. 21068–21073.

[8] Bresler, G., and Karzand, M. (2020), "Learning a tree-structured Ising model in order to make predictions," *The Annals of Statistics*, 48, 713–737.

[9] Brown, L. (1986), *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, Hayworth, CA, USA: Institute of Mathematical Statistics.

[10] Burt, R. S. (1992), *Structural Holes: The Social Structure of Competition*, Cambridge, MA: Harvard University Press.

[11] Butts, C. T. (2020), "A dynamic process interpretation of the sparse ERGM reference model," *Journal of Mathematical Sociology*.

[12] Cai, D., Campbell, T., and Broderick, T. (2016), "Edge-exchangeable graphs and sparsity," in *Advances in Neural Information Processing Systems*, eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., pp. 4249–4257.

[13] Caimo, A., and Friel, N. (2011), "Bayesian inference for exponential random graph models," *Social Networks*, 33, 41–55.

[14] Caron, F., and Fox, E. B. (2017), "Sparse graphs using exchangeable random measures," *Journal of the Royal Statistical Society, Series B (with discussion)*, 79, 1–44.

[15] Chatterjee, S. (2007), "Estimation in spin glasses: A first step," *The Annals of Statistics*, 35, 1931–1946.

[16] Chatterjee, S., and Diaconis, P. (2013), "Estimating and understanding exponential random graph models," *The Annals of Statistics*, 41, 2428–2461.

[17] Chatterjee, S., Diaconis, P., and Sly, A. (2011), "Random graphs with a given degree sequence," *The Annals of Applied Probability*, 21, 1400–1435.

[18] Chazottes, J. R., Collet, P., Külske, C., and Redig, F. (2007), "Concentration inequalities for random fields via coupling," *Probability Theory and Related Fields*, 137, 201–225.

[19] Chen, M., Kato, K., and Leng, C. (2019), "Analysis of networks via the sparse β-model," *arXiv preprint arXiv:1908.03152*.

[20] Comets, F. (1992), "On consistency of a class of estimators for exponential families of Markov random fields on the lattice," *The Annals of Statistics*, 20, 455–468.

[21] Crane, H., and Dempsey, W. (2018), "Edge exchangeable models for interaction networks," *Journal of the American Statistical Association*, 113, 1311–1326.

[22] Csiszár, I., and Talata, Z. (2006), "Consistent estimation of the basic neighborhood of Markov random fields," *The Annals of Statistics*, 34, 123–145.

[23] Dawid, A. P. (1979), "Conditional independence in statistical theory," *Journal of the Royal Statistical Society, Series B*, 41, 1–31.

[24] — (1980), "Conditional independence for statistical operations," *The Annals of Statistics*, 8, 598–617.

[25] Erdős, P., and Rényi, A. (1960), "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–61.

[26] Fan, Y., Zhang, H., and Yan, T. (2020), "Asymptotic theory for differentially private generalized β-models with parameters increasing," ArXiv:2002.12733v1.

[27] Fortuin, C. M., Kasteleyn, P. W., and Ginibre, J. (1971), "Correlation inequalities on some partially ordered sets," *Communications in Mathematical Physics*, 22, 89–103.

[28] Frank, O., and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

[29] Gao, C., Lu, Y., and Zhou, H. H. (2015), "Rate-optimal graphon estimation," *The Annals of Statistics*, 43, 2624–2652.

[30] Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017), "Achieving optimal misclassification proportion in stochastic block models," *Journal of Machine Learning Research*, 18, 1980–2024.

[31] Geiger, D., Heckerman, D., King, H., and Meek, C. (2001), "Stratified exponential families: graphical models and model selection," *The Annals of Statistics*, 29, 505–529.

[32] Georgii, H. O., Häggström, O., and Maes, C. (2001), "The random geometry of equilibrium phases," in *Phase transitions and critical phenomena*, Elsevier, Vol. 18, pp. 1–142.

[33] Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society, Series B*, 54, 657–699.

[34] Ghosal, P., and Mukherjee, S. (2020), "Joint estimation of parameters in Ising model," *The Annals of Statistics*, 48, 785–810.

[35] Gidas, B. (1986), "Consistency of maximum likelihood and pseudo-likelihood estimation for Gibbs distributions," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, eds. Fleming, W., and Lions, P. L., New York: Springer-Verlag, pp. 1–17.

[36] Gould, R. V., and Fernandez, R. M. (1989), "Structures of mediation: A formal approach to brokerage in transaction networks," *Sociological Methodology*, 19, 89–126.

[37] Handcock, M. S. (2003), "Statistical Models for Social Networks: Inference and Degeneracy," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, eds. Breiger, R., Carley, K., and Pattison, P., Washington, D.C.: National Academies Press, pp. 1–12.

[38] Hillar, C. J., and Wibisono, A. (2015), "A Hadamard-type lower bound for symmetric diagonally dominant positive matrices," *Linear Algebra and its Applications*, 472, 135–141.

[39] Hoff, P. D. (2021), "Additive and multiplicative effects network models," *Statistical Science*, to appear.

[40] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

[41] Holland, P. W., and Leinhardt, S. (1972), "Some evidence on the transitivity of positive interpersonal sentiment," *American Journal of Sociology*, 77, 1205–1209.

[42] — (1981), "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76, 33–65.

[43] Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008), "Goodness of fit of social network models," *Journal of the American Statistical Association*, 103, 248–258.

[44] Hunter, D. R., and Handcock, M. S. (2006), "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, 15, 565–

583.

[45] Jensen, J. L., and Künsch, H. R. (1994), "On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes," *Annals of the Institute of Mathematical Statistics*, 46, 475–486.

[46] Jensen, J. L., and Møller, J. (1991), "Pseudolikelihood for exponential family models of spatial point processes," *The Annals of Applied Probability*, 1, 445–461.

[47] Jonasson, J. (1999), "The random triangle model," *Journal of Applied Probability*, 36, 852–876.

[48] Karwa, V., and Slavković, A. B. (2016), "Inference using noisy degrees: Differentially private β-model and synthetic graphs," *The Annals of Statistics*, 44, 87–112.

[49] Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, New York: Springer-Verlag.

[50] Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011), "Adjusting for network size and composition effects in exponential-family random graph models," *Statistical Methodology*, 8, 319–339.

[51] Krivitsky, P. N., and Kolaczyk, E. D. (2015), "On the question of effective sample size in network modeling: An asymptotic inquiry," *Statistical Science*, 30, 184–198.

[52] Lauritzen, S. (1996), *Graphical Models*, Oxford, UK: Oxford University Press.

[53] Lauritzen, S., Rinaldo, A., and Sadeghi, K. (2018), "Random networks, graphical models and exchangeability," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 481–508.

[54] Lauritzen, S., and Spiegelhalter, D. (1988), "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society, Series B (with discussion)*, 50, 157–224.

[55] Lehmann, E. L. (1983), *Theory of Point estimation*, New York: John Wiley & Sons.

[56] Lindvall, T. (2002), *Lectures On The Coupling Method*, Courier Corporation.

[57] Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2019), *Handbook of Graphical Models*, Boca Raton, Florida: CRC Press.

[58] Mase, S. (1995), "Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes," *The Annals of Applied Probability*, 5, 603–612.

[59] Meinshausen, N., and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the LASSO," *The Annals of Statistics*, 34, 1436–1462.

[60] Mukherjee, R., Mukherjee, S., and Sen, S. (2018), "Detection thresholds for the β-model on sparse graphs," *The Annals of Statistics*, 46, 1288–1317.

[61] Mukherjee, S. (2020), "Degeneracy in sparse ERGMs with functions of degrees as sufficient statistics," *Bernoulli*, 26, 1016–1043.

[62] Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087.

[63] Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, 38, 1287–1319.

[64] Rinaldo, A., Petrović, S., and Fienberg, S. E. (2013), "Maximum likelihood estimation in the β-model," *The Annals of Statistics*, 41, 1085–1110.

[65] Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.

[66] Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society, Series B*, 77, 647–676.

[67] Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. (2020), "Exponential-family models of random graphs: Inference in finite, super, and infi-

nite population scenarios," *Statistical Science*, 35, 627–662.

[68] Schweinberger, M., and Stewart, J. (2020), "Concentration and consistency results for canonical and curved exponential-family models of random graphs," *The Annals of Statistics*, 48, 374–396.

[69] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153.

[70] Stewart, J. R., and Schweinberger, M. (2020), "Supplement: Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension," Tech. rep., Department of Statistics, Florida State University.

[71] Strauss, D., and Ikeda, M. (1990), "Pseudolikelihood estimation for social networks," *Journal of the American Statistical Association*, 85, 204–212.

[72] Tang, M., Sussman, D. L., and Priebe, C. E. (2013), "Universally consistent vertex classification for latent positions graphs," *The Annals of Statistics*, 41, 1406–1430.

[73] van den Berg, J., and Maes, C. (1994), "Disagreement percolation in the study of Markov fields," *The Annals of Probability*, 22, 749–763.

[74] van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

[75] van Duijn, M. A. J., Gile, K., and Handcock, M. S. (2009), "A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models," *Social Networks*, 31, 52–62.

[76] Wainwright, M. J., and Jordan, M. I. (2008), "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1, 1–305.

[77] Whittle, P. (1963), "Stochastic processes in several dimensions," *Bulletin of the International Statistical Institute*, 40, 974–994.

[78] Xue, L., Zou, H., and Cai, T. (2012), "Nonconcave penalized composite conditional likelihood estimation of sparse Ising models," *The Annals of Statistics*, 40, 1403–1429.

[79] Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2019), "Statistical inference in a directed network model with covariates," *Journal of the American Statistical Association*, 114, 857–868.

[80] Yan, T., Leng, C., and Zhu, J. (2016), "Asymptotics in directed exponential random graph models with an increasing bi-degree sequence," *The Annals of Statistics*, 44, 31–57.

[81] Yan, T., and Xu, J. (2013), "A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices," *Biometrika*, 100, 519–524.

Supplement: Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension

By Jonathan Stewart and Michael Schweinberger

Florida State University and Rice University

We prove all results stated in the manuscript. We divide the supplement into the following appendices:

In Appendices A, B, and C, it is convenient to adopt the notation used in Section 3 of the manuscript, denoting the number of edge variables by $M = \binom{N}{2}$ and edge variables by $X_1, \dots, X_M$, where N is the number of nodes. In addition, we denote the data-generating parameter vector by $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$, the data-generating probability measure and expectation by $\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\theta}_N^\star}$ and $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\theta}_N^\star}$, respectively, and the probability density function of $\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\theta}_N^\star}$ with respect to dominating measure ν by $f \equiv f_{\boldsymbol{\theta}_N^\star}$. The parameter space $\boldsymbol{\Theta}_N$ is an open subset of $\mathbb{R}^p$. In Corollaries 1–4, $\boldsymbol{\Theta}_N = \mathbb{R}^p$ and $p \geq N$.

APPENDIX A: BACKGROUND

The factorization of probability density function (2.1), combined with the exponential parameterization (2.2), implies that the probability density function can be written in exponential-family form:

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \quad = \quad a(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\theta}_N)\right) \quad > \quad 0, \qquad \boldsymbol{x} \in \mathbb{X}, \tag{A.1}$$

where $a : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ is given by

$$a(\boldsymbol{x}) \quad = \quad \prod_{i=1}^{M} a_i(x_i,\, \boldsymbol{x}_{\mathcal{S}_i}), \qquad \boldsymbol{x} \in \mathbb{X}$$

and $s : \mathbb{X} \mapsto \mathbb{R}^p$ is the sufficient statistic vector given by

$$s(\boldsymbol{x}) \;\;=\;\; \sum_{i=1}^{M} s_i(x_i,\, \boldsymbol{x}_{\mathcal{S}_i}), \qquad \boldsymbol{x} \in \mathbb{X}.$$

The set $\mathcal{S}_i$ is a subset of $\{1, \ldots, M\} \setminus \{i\}$ $(i = 1, \ldots, M)$. The function $\psi : \mathbb{R}^p \mapsto \mathbb{R} \cup \{\infty\}$ ensures that $\int_{\mathbb{X}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \, \mathrm{d}\, \nu(\boldsymbol{x}) = 1$ and is defined by

$$\psi(\boldsymbol{\theta}_N) \;\;=\;\; \log \int\limits_{\mathbb{X}} a(\boldsymbol{x}') \, \exp\left(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}')\rangle\right) \mathrm{d}\, \nu(\boldsymbol{x}'), \qquad \boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N,$$

and the set of possible parameters $\boldsymbol{\Theta}_N$ is given by

$$\boldsymbol{\Theta}_N \;\;=\;\; \{\boldsymbol{\theta}_N \in \mathbb{R}^p \,:\, \psi(\boldsymbol{\theta}_N) < \infty\}.$$

The parameter vector $\boldsymbol{\theta}_N$ and parameter space $\boldsymbol{\Theta}_N$ are known as the natural parameter vector and natural parameter space of the exponential family, respectively [9]. To ensure that the natural parameter vector $\boldsymbol{\theta}_N$ is identifiable, we assume that the exponential family is minimal in the sense of Brown [9, p. 2], that is, neither the natural parameter vector $\boldsymbol{\theta}_N$ nor the closure of the convex hull of the set $\{s(\boldsymbol{x}) \,:\, \nu(\boldsymbol{x}) > 0\}$ are contained in a proper affine subspace of $\mathbb{R}^p$ (a.e. ν) [9, p. 2]. The assumption of a minimal exponential family involves no loss of generality, because all non-minimal exponential families can be reduced to minimal exponential families by suitable transformations of natural parameters, sufficient statistics, and reference measure [9, Theorem 1.9, p. 13]. In addition, we assume that $\boldsymbol{\Theta}_N$ is an open subset of $\mathbb{R}^p$, which implies that the exponential family is regular in the sense of Brown [9, p. 2]. These assumptions are satisfied by Corollaries 1–4.

APPENDIX B: PROOF OF THEOREM 1

We prove Theorem 1 stated in Section 3.2 of the manuscript.

Proof of Theorem 1. We divide the proof into three parts:

I. Existence and uniqueness of $\widehat{\boldsymbol{\theta}}_N$.

II. Bounding the probability of event $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^\star_N\|_\infty < \epsilon$.

III. Deriving the convergence rate of $\widehat{\boldsymbol{\theta}}_N$.

To prepare the ground, we first review basic facts about the relationship of natural and mean-value parameters in exponential families [9]. By assumption, the exponential family is minimal and its natural parameter space

$\mathbf{\Theta}_N \subseteq \mathbb{R}^p$ is open, so the exponential family is regular in the sense of Brown [9, p. 2]. Since $\mathbf{\Theta}_N \subseteq \mathbb{R}^p$ is open, the expectation $\mathbb{E}_{\boldsymbol{\theta}_N}\, s(\boldsymbol{X})$ of the sufficient statistic vector $s(\boldsymbol{X})$ exists for all $\boldsymbol{\theta}_N \in \mathbf{\Theta}_N$ [9, Theorem 2.2 and Corollary 2.3, pp. 34–36]. Let $\mathbb{M} = \boldsymbol{\mu}(\mathbf{\Theta}_N)$ be the image of $\mathbf{\Theta}_N$ under the mean-value parameter map $\boldsymbol{\mu} : \mathbf{\Theta}_N \mapsto \mathbb{M}$, defined by $\boldsymbol{\mu}(\boldsymbol{\theta}_N) = \mathbb{E}_{\boldsymbol{\theta}_N}\, s(\boldsymbol{X})$; note that $\mathbb{M}$ is the relative interior of the convex hull of the set $\{s(\boldsymbol{x}) : \nu(\boldsymbol{x}) > 0\}$. Observe that the natural parameter space $\mathbf{\Theta}_N$ is a convex set [9, Theorem 1.13, p. 19], and choose $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon) \subset \mathbf{\Theta}_N$. The map $\boldsymbol{\mu} : \mathbf{\Theta}_N \mapsto \mathbb{M}$ is a homeomorphism [9, Theorem 3.6, p. 74], so the inverse map $\boldsymbol{\mu}^{-1} : \mathbb{M} \mapsto \mathbf{\Theta}_N$ exists and both $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{-1}$ are continuous, one-to-one, and onto. The fact that the map $\boldsymbol{\mu} : \mathbf{\Theta}_N \mapsto \mathbb{M}$ is a homeomorphism implies that, for any $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon) \subset \mathbf{\Theta}_N$, there exist $\delta(\epsilon) > 0$ and $\gamma(\delta(\epsilon)) > 0$ such that

$$\boldsymbol{\mu}(\boldsymbol{\theta}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \quad \text{implies} \quad \boldsymbol{\theta}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon) \;\subseteq\; \mathbf{\Theta}_N$$

and

$$\boldsymbol{\theta}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))) \quad \text{implies} \quad \boldsymbol{\mu}(\boldsymbol{\theta}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \;\subseteq\; \mathbb{M}.$$

It is worth noting that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon)))$ cannot be larger than $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$: Otherwise there would exist some $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))) \setminus \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$ such that $\boldsymbol{\mu}(\boldsymbol{\theta}_N) \in \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$, which would contradict the fact that the map $\boldsymbol{\mu} : \mathbf{\Theta}_N \mapsto \mathbb{M}$ is one-to-one and that all $\boldsymbol{\mu}(\boldsymbol{\theta}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$ map to $\boldsymbol{\theta}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$. As a consequence,

$$\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))) \;\;\subseteq\;\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon),$$

which in turn implies that there exists $C_1 \in (0, 1]$ such that

$$\gamma(\delta(\epsilon)) \;\;=\;\; C_1\, \epsilon \;\;\leq\;\; \epsilon.$$

The fundamental relationship between the natural parameter vector $\boldsymbol{\theta}_N$ and the mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\theta}_N)$ helps establish I, II, and III.

I. Existence and uniqueness of maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_N$. In minimal exponential families, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ exists and is unique as long as $s(\boldsymbol{X}) \in \mathbb{M}$ [9, Theorem 5.5, p. 148]. Thus, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ exists and is unique in the event

$$s(\boldsymbol{X}) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \;\subseteq\; \mathbb{M}.$$

II. Bounding the probability of event $\widehat{\boldsymbol{\theta}}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$. To bound the probability of event $\widehat{\boldsymbol{\theta}}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$, we use two facts. First, we have established that

$$\boldsymbol{\mu}(\boldsymbol{\theta}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \quad \text{implies} \quad \boldsymbol{\theta}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon) \;\subseteq\; \boldsymbol{\Theta}_N.$$

Second, exponential-family theory [9, Theorem 5.5, p. 148] implies that the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ exists and is unique in the event $s(\boldsymbol{X}) \in \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \subseteq \mathbb{M}$, and solves

$$\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}_N) \;=\; s(\boldsymbol{X}) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \;\subseteq\; \mathbb{M}.$$

Taken together, these two facts imply that

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)\right) \;&\geq\; \mathbb{P}\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))\right)\\[1em] &=\; \mathbb{P}\left(s(\boldsymbol{X}) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))\right).\end{aligned}$$

To bound the probability of event $s(\boldsymbol{X}) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$, we invoke Theorem 1 of Chazottes et al. [18, p. 207], which implies that

$$\mathbb{P}\left(|s_i(\boldsymbol{X}) - \mathbb{E}\, s_i(\boldsymbol{X})| \;\geq\; \delta(\epsilon)\right) \;\leq\; 2\, \exp\left(-\frac{2\, \delta(\epsilon)^2}{|||\mathcal{D}|||_2^2\; \|\Xi\, s_i\|_2^2}\right),\; i = 1, \ldots, p,$$

where the quantities $\mathcal{D}$ and $\Xi\, s_1,\, \ldots,\, \Xi\, s_p$ are defined in Section 3.2 of the manuscript. A union bound over the p coordinates of $s(\boldsymbol{X})$ shows that

$$\mathbb{P}\left(\|s(\boldsymbol{X}) - \mathbb{E}\, s(\boldsymbol{X})\|_\infty \;\geq\; \delta(\epsilon)\right) \quad\leq\quad 2\, \exp\left(-\frac{2\, \delta(\epsilon)^2}{|||\mathcal{D}|||_2^2\; \Psi^2} + \log p\right),$$

where $\Psi = \max_{1 \leq i \leq p} \|\Xi\, s_i\|_2$. As a result,

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)\right) \;&\geq\; \mathbb{P}\left(s(\boldsymbol{X}) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))\right)\\ &=\; \mathbb{P}\left(\|s(\boldsymbol{X}) - \mathbb{E}\, s(\boldsymbol{X})\|_\infty \;\geq\; \delta(\epsilon)\right)\\ &\geq\; 1 - 2\, \exp\left(-\frac{2\, \delta(\epsilon)^2}{|||\mathcal{D}|||_2^2\; \Psi^2} + \log p\right),\end{aligned}$$

where we used the fact that $\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star) = \mathbb{E}_{\boldsymbol{\theta}_N^\star}\, s(\boldsymbol{X})$ by definition.

III. Deriving the convergence rate of $\widehat{\boldsymbol{\theta}}_N$. To derive the convergence rate of $\widehat{\boldsymbol{\theta}}_N$, we relate the radius $\delta(\epsilon)$ of the ℓ_∞-ball $\mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$ to the radius ϵ of the ℓ_∞-ball $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)$. To do so, observe that

$$\int\limits_{\mathbb{R}^p} \mathbb{1}(\boldsymbol{\mu} \in \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)))\; \mathrm{d}\,\lambda(\boldsymbol{\mu}) \;=\; \lambda(\mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))) \;=\; (2\, \delta(\epsilon))^p,$$

where λ denotes the p-dimensional Lebesgue measure on the Borel subsets of $\mathbb{R}^p$. Let $\mathcal{J}(\boldsymbol{\theta}_N)$ be the Jacobian matrix of $\boldsymbol{\mu}(\boldsymbol{\theta}_N)$ with respect to $\boldsymbol{\theta}_N$, and observe that $\mathcal{J}(\boldsymbol{\theta}_N)$ is equal to the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta}_N)$ in exponential families. Define

$$\mathbb{M}(\gamma(\delta(\epsilon))) \;=\; \{\boldsymbol{\mu}(\boldsymbol{\theta}_N) \in \mathbb{M} : \; \boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon)))\}$$

and observe that

$$\boldsymbol{\theta}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))) \quad \text{implies} \quad \boldsymbol{\mu}(\boldsymbol{\theta}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \;\subseteq\; \mathbb{M},$$

which in turn implies that

$$\mathbb{M}(\gamma(\delta(\epsilon))) \;\subseteq\; \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)).$$

A change of the variable of integration from $\boldsymbol{\mu}$ to $\boldsymbol{\theta}$ shows that

$$\begin{aligned}
&\int\limits_{\mathbb{R}^p} \mathbb{1}(\boldsymbol{\mu} \in \mathcal{B}_\infty(\boldsymbol{\mu}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)))\; \mathrm{d}\,\lambda(\boldsymbol{\mu}) \;\geq\; \int\limits_{\mathbb{R}^p} \mathbb{1}(\boldsymbol{\mu} \in \mathbb{M}(\gamma(\delta(\epsilon))))\; \mathrm{d}\,\lambda(\boldsymbol{\mu})\\
&= \int\limits_{\mathbb{R}^p} \mathbb{1}(\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))))\, |\det(\mathcal{I}(\boldsymbol{\theta}_N))|\; \mathrm{d}\,\lambda(\boldsymbol{\theta}_N)\\
&\geq \left(\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon)))} |\det(\mathcal{I}(\boldsymbol{\theta}_N))|\right) \int\limits_{\mathbb{R}^p} \mathbb{1}(\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))))\; \mathrm{d}\,\lambda(\boldsymbol{\theta}_N)\\
&\geq \inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} |\det(\mathcal{I}(\boldsymbol{\theta}_N))|\;\; (2\,\gamma(\delta(\epsilon)))^p,
\end{aligned}$$

where we exploited the fact that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \gamma(\delta(\epsilon))) \subseteq \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$. In other words, we have shown that

$$(2\,\delta(\epsilon))^p \;\;\geq\;\; \inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} |\det(\mathcal{I}(\boldsymbol{\theta}_N))|\;\; (2\,\gamma(\delta(\epsilon)))^p.$$

By assumption, for any $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon) \subset \boldsymbol{\Theta}_N$, there exists a constant $C_2 > 0$ and a function $\Lambda : \{1, 2, \ldots\} \mapsto \mathbb{R}^+$ of the number of nodes N such that

$$\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)} \sqrt[p]{|\det(\mathcal{I}(\boldsymbol{\theta}_N))|} \;\;\geq\;\; C_2\, \Lambda.$$

As a consequence, we know that $\delta(\epsilon)$ and ϵ are related to each other as follows:

$$\delta(\epsilon) \;\;\geq\;\; \inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)} \sqrt[p]{|\det(\mathcal{I}(\boldsymbol{\theta}_N))|}\; \gamma(\delta(\epsilon)) \;\;\geq\;\; \epsilon\, C_1\, C_2\, \Lambda,$$

using $\gamma(\delta(\epsilon)) = C_1 \, \epsilon$ with $C_1 \in (0, 1]$.

Having related $\delta(\epsilon)$ to ϵ, we establish convergence rates by choosing ϵ so that the event $\widehat{\boldsymbol{\theta}}_N \; \in \; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \, \epsilon)$ occurs with high probability. Choose any $C_3 > 1$ and let

$$\epsilon \;\; = \;\; \frac{\sqrt{C_3}}{\sqrt{2} \, C_1 \, C_2} \; \frac{|||\mathcal{D}|||_2 \; \Psi \, \sqrt{\log p}}{\Lambda} \;\; > \;\; 0,$$

which implies that

$$\delta(\epsilon) \;\; \geq \;\; \sqrt{\frac{C_3}{2}} \, |||\mathcal{D}|||_2 \; \Psi \; \sqrt{\log p} \;\; > \;\; 0.$$

We may therefore conclude that

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_N \; \in \; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \, \epsilon)\right) \;\; \geq \;\; 1 - 2 \, \exp(-A \, \log p),$$

where $A = C_3 - 1 > 0$ follows from $C_3 > 1$.

Last, but not least, assume that, for any $C_0 > 0$, however large, there exists $N_0 > 0$ such that

$$\Lambda \;\; > \;\; C_0 \, |||\mathcal{D}|||_2 \; \Psi \; \sqrt{\log p} \quad \text{for all} \quad N > N_0.$$

Then, with at least probability $1 - 2 \, \exp(-A \, \log p)$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ exists, is unique, and satisfies

$$\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\; \leq \;\; C \; |||\mathcal{D}|||_2 \sqrt{\frac{\log p}{\Lambda^2 \, / \, \Psi^2}} \;\; \longrightarrow \;\; 0 \quad \text{as} \quad N \longrightarrow \infty,$$

where

$$C \;\; = \;\; \frac{\sqrt{C_3}}{\sqrt{2} \, C_1 \, C_2} \;\; > \;\; 0.$$

Remark 1. Extensions to dependent random variables with countable and uncountable sample spaces. Theorem 1 is not restricted to random graphs with dependent edges. It covers models of dependent random variables with finite sample spaces, and can be extended to countable sample spaces: e.g., the concentration result of Chazottes et al. [18] used in Theorem 1 assumes that the sample spaces are finite—motivated by applications to Ising models—but could be extended to countable sample spaces. Uncountable sample spaces could be accomodated by replacing the concentration result of Chazottes et al. [18] by other suitable concentration results, e.g., Subgaussian concentration results. Likewise, the exponential-family properties used in Theorem 1 are neither restricted to finite nor countable sample spaces [9].

APPENDIX C: PROOFS OF LEMMA 1 AND THEOREM 2

We prove Lemma 1 and Theorem 2 stated in Section 3.3 of the manuscript. Auxiliary results are proved in Appendix C.1.

PROOF OF LEMMA 1. Appendix A shows that the family of densities $\{f_{\boldsymbol{\theta}_N},\, \boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N\}$ parameterized by (2.1) and (2.2) is an exponential family of densities. We take advantage of the properties of exponential families [9] to prove Lemma 1, and divide the proof into four parts:

I. $\boldsymbol{\Theta}_N$ is a convex set.

II. $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ is a strictly concave function on the convex set $\boldsymbol{\Theta}_N$.

III. $\boldsymbol{\theta}_N^\star$ is the unique maximizer of $\mathbb{E}\, \ell(\boldsymbol{\theta}_N;\, \boldsymbol{X})$.

IV. $\boldsymbol{\theta}_N^\star$ is the unique maximizer of $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$.

I. $\boldsymbol{\Theta}_N$ is a convex set. By definition of $\boldsymbol{\Theta}_N = \{\boldsymbol{\theta}_N \in \mathbb{R}^p :\, \psi(\boldsymbol{\theta}_N) < \infty\}$ and Hölder's inequality, $\boldsymbol{\Theta}_N$ is a convex set [9, Theorem 1.13, p. 19].

II. $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ is a strictly concave function on the convex set $\boldsymbol{\Theta}_N$. Let $\boldsymbol{x}$ be an observation of a random graph $\boldsymbol{X}$ with dependent edges. Then, by definition,

$$\widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) \;=\; \sum_{i=1}^{M} \widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{x}),$$

where

$$\widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{x}) \;=\; \langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x})\rangle - \psi_i(\boldsymbol{\theta}_N;\, \boldsymbol{x}_{-i})$$

and

$$\psi_i(\boldsymbol{\theta}_N; \boldsymbol{x}_{-i}) = \log\left(\exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-i}, x_i = 0)\rangle) + \exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-i}, x_i = 1)\rangle)\right).$$

We first show that $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) = \sum_{i=1}^{M} \mathbb{E}\, \widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ is a concave function on the convex set $\boldsymbol{\Theta}_N$ by proving that the functions $\mathbb{E}\, \widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ are concave on $\boldsymbol{\Theta}_N$. Observe that the functions $\mathbb{E}\, \widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ are concave provided the functions $\widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ are concave for all $\boldsymbol{x} \in \mathbb{X}$. To show that the functions $\widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ are concave for all $\boldsymbol{x} \in \mathbb{X}$, consider any $i \in \{1, \ldots, M\}$, any $x_i \in \{0, 1\}$, and any $\boldsymbol{x}_{-i} \in \{0, 1\}^{M-1}$. Each $\widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ consists of two terms. The first term, $\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x})\rangle$, is a linear function of $\boldsymbol{\theta}_N$, so $\widetilde{\ell}_i(\boldsymbol{\theta}_N;\, \boldsymbol{x})$ is a concave function of $\boldsymbol{\theta}_N$ if the second term, $\psi_i(\boldsymbol{\theta}_N;\, \boldsymbol{x}_{-i})$, is a convex function of $\boldsymbol{\theta}_N$. Consider any $\boldsymbol{\theta}_N^{(1)} \in \boldsymbol{\Theta}_N$, any $\boldsymbol{\theta}_N^{(2)} \in \boldsymbol{\Theta}_N$, and any $\lambda \in (0, 1)$. Then, by

Hölder's inequality,

$$\psi_i\left(\lambda\, \boldsymbol{\theta}_N^{(1)} + (1-\lambda)\, \boldsymbol{\theta}_N^{(2)};\ \boldsymbol{x}_{-i}\right) \leq \lambda\, \psi_i\left(\boldsymbol{\theta}_N^{(1)};\ \boldsymbol{x}_{-i}\right) + (1-\lambda)\, \psi_i\left(\boldsymbol{\theta}_N^{(2)};\ \boldsymbol{x}_{-i}\right).$$

As a consequence, for any $\boldsymbol{x}_{-i} \in \{0, 1\}^{M-1}$, $\psi_i(\boldsymbol{\theta}_N;\ \boldsymbol{x}_{-i})$ is a convex function on $\boldsymbol{\Theta}_N$. Hence, for all $\boldsymbol{x} \in \mathbb{X}$, $\widetilde{\ell}(\boldsymbol{\theta}_N;\ \boldsymbol{x})$ is a concave function on $\boldsymbol{\Theta}_N$, and so is $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\ \boldsymbol{X})$.

Second, we prove by contradiction that $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\ \boldsymbol{X})$ is a strictly concave function on $\boldsymbol{\Theta}_N$, by showing that there exists $i^\star \in \{1, \ldots, M\}$ such that $\mathbb{E}\, \psi_{i^\star}(\boldsymbol{\theta}_N;\ \boldsymbol{X}_{-i^\star})$ is strictly convex on $\boldsymbol{\Theta}_N$, which implies that $\mathbb{E}\, \widetilde{\ell}_{i^\star}(\boldsymbol{\theta}_N;\ \boldsymbol{X})$ is strictly concave on $\boldsymbol{\Theta}_N$. Suppose that there does not exist any $i^\star \in \{1, \ldots, M\}$ such that $\mathbb{E}\, \psi_{i^\star}(\boldsymbol{\theta}_N;\ \boldsymbol{X}_{-i^\star})$ is strictly convex on $\boldsymbol{\Theta}_N$. Then, for all $i \in \{1, \ldots, M\}$, all $\boldsymbol{x}_{-i} \in \{0, 1\}^{M-1}$, and all $x_i \in \{0, 1\}$,

$$\exp\left(\left\langle \boldsymbol{\theta}_N^{(1)},\ s(\boldsymbol{x}_{-i},\, x_i)\right\rangle\right) \propto \exp\left(\left\langle \boldsymbol{\theta}_N^{(2)},\ s(\boldsymbol{x}_{-i},\, x_i)\right\rangle\right).$$

In other words, for all $\boldsymbol{x} \in \mathbb{X}$,

$$\exp\left(\left\langle \boldsymbol{\theta}_N^{(1)},\, s(\boldsymbol{x})\right\rangle\right) \propto \exp\left(\left\langle \boldsymbol{\theta}_N^{(2)},\, s(\boldsymbol{x})\right\rangle\right). \tag{C.1}$$

The conclusion (C.1) contradicts the assumption that the exponential family is minimal. Therefore, there exists $i^\star \in \{1, \ldots, M\}$ such that $\mathbb{E}\, \psi_{i^\star}(\boldsymbol{\theta}_N;\ \boldsymbol{X}_{-i^\star})$ is strictly convex on $\boldsymbol{\Theta}_N$, which implies that $\mathbb{E}\, \widetilde{\ell}_{i^\star}(\boldsymbol{\theta}_N;\ \boldsymbol{X})$ is strictly concave on $\boldsymbol{\Theta}_N$, and so is $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\ \boldsymbol{X}) = \sum_{i=1}^{M} \mathbb{E}\, \widetilde{\ell}_i(\boldsymbol{\theta}_N;\ \boldsymbol{X})$.

III. $\boldsymbol{\theta}_N^\star$ is the unique maximizer of $\mathbb{E}\, \ell(\boldsymbol{\theta}_N; \boldsymbol{X})$. Maximizing $\mathbb{E}\, \ell(\boldsymbol{\theta}_N; \boldsymbol{X})$ is equivalent to solving

$$\nabla_{\boldsymbol{\theta}_N}\, \mathbb{E}\, \ell(\boldsymbol{\theta}_N;\ \boldsymbol{X}) \quad = \quad \mathbb{E}\, s(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}_N}\, s(\boldsymbol{X}) \quad = \quad \boldsymbol{0}. \tag{C.2}$$

The unique solution of (C.2) is $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$, because $\mathbb{E}\, s(\boldsymbol{X}) \equiv \mathbb{E}_{\boldsymbol{\theta}_N^\star}\, s(\boldsymbol{X}) \in \mathbb{M}$. The fact that the solution is unique follows from the fact the map $\boldsymbol{\mu} : \boldsymbol{\Theta}_N \mapsto \mathbb{M}$ defined by $\boldsymbol{\mu}(\boldsymbol{\theta}_N) = \mathbb{E}_{\boldsymbol{\theta}_N}\, s(\boldsymbol{X})$ is one-to-one [9, Theorem 3.6, p. 74]. As a result, $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ is the unique maximizer of $\mathbb{E}\, \ell(\boldsymbol{\theta}_N; \boldsymbol{X})$.

VI. $\boldsymbol{\theta}_N^\star$ is the unique maximizer of $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\ \boldsymbol{X})$. Observe that, for any $\boldsymbol{x} \in \mathbb{X}$, $\widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{x})$ is a sum of exponential-family loglikelihood functions, because the conditional distributions of edge variables X_i given $\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}$ are Bernoulli distributions ($i = 1, \ldots, M$), and Bernoulli distributions are exponential-family distributions. As a result, $\widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{x})$ is continuously differentiable on $\boldsymbol{\Theta}_N$ for all $\boldsymbol{x} \in \mathbb{X}$ [9], and so is $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$. We have

$$\boldsymbol{g}(\boldsymbol{\theta}_N) \quad = \quad \mathbb{E}\, \nabla_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}) \quad = \quad \mathbb{E} \sum_{i=1}^{M} \left(s(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-i}}\, s(\boldsymbol{X})\right),$$

where $\mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-i}}$ denotes the conditional expectation, computed with respect to the conditional distribution of X_i given $\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}$. By the law of total expectation and the fact that $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\theta}_N^\star}$, we have $\mathbb{E}\, \mathbb{E}_{\boldsymbol{\theta}_N^\star, \boldsymbol{X}_{-i}}\, s(\boldsymbol{X}) = \mathbb{E}\, s(\boldsymbol{X})$, which implies that

$$\boldsymbol{g}(\boldsymbol{\theta}_N^\star) \;=\; \mathbb{E} \sum_{i=1}^{M} \Big(s(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}_N^\star, \boldsymbol{X}_{-i}}\, s(\boldsymbol{X}) \Big) \;=\; \mathbf{0}. \tag{C.3}$$

Thus, a root of $\boldsymbol{g}(\boldsymbol{\theta}_N)$ exists, and $\boldsymbol{\theta}_N^\star$ is a root of $\boldsymbol{g}(\boldsymbol{\theta}_N)$. In addition, $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ is strictly concave on $\boldsymbol{\Theta}_N$, so $\boldsymbol{\theta}_N^\star$ is the unique root of $\boldsymbol{g}(\boldsymbol{\theta}_N)$. As a consequence, the maximizer of $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ as a function of $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ exists and is unique, and is given by $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$.

Proof of Theorem 2. To prepare the ground, we first review some basic facts that help prove Theorem 2.

By Lemma 1, the set $\boldsymbol{\Theta}_N$ is a convex set and $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ is a strictly concave function on the convex set $\boldsymbol{\Theta}_N$. In addition, the maximizer of $\mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ exists and is unique, and is given by $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$. By Lemma 2, its gradient

$$\boldsymbol{g}(\boldsymbol{\theta}_N) \;=\; \nabla_{\boldsymbol{\theta}_N}\, \mathbb{E}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}) \;=\; \mathbb{E}\, \nabla_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$$

exists, is continuous, and is one-to-one, so the inverse of $\boldsymbol{g}(\boldsymbol{\theta}_N)$ exists and is continuous; note that the interchange of differentiation and integration is admissible because the expectation is a finite sum. Consider any $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N$ and any $\epsilon > 0$ small enough so that $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon) \subset \boldsymbol{\Theta}_N = \mathbb{R}^p$. By taking advantage of the continuity and one-to-one nature of $\boldsymbol{g}(\boldsymbol{\theta}_N)$ and its inverse, it can be shown—using an argument along the lines of Theorem 1, with $\boldsymbol{\mu}(\boldsymbol{\theta}_N)$ replaced by $\boldsymbol{g}(\boldsymbol{\theta}_N)$—that there exists $\delta(\epsilon) > 0$ such that

$$\boldsymbol{g}(\boldsymbol{\theta}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{g}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)) \qquad \text{implies} \qquad \boldsymbol{\theta}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon).$$

We divide the remainder of the proof into three parts:

I. Existence.

II. Convergence rate.

III. Uniform convergence of $\boldsymbol{g}(.;\, \boldsymbol{X})$ on $\boldsymbol{\Theta}_N$.

I. Existence. Consider any $\gamma \in (0,\, \delta(\epsilon)\,/\,2]$ and define

$$\mathbb{G} \;=\; \left\{ \boldsymbol{x} \in \mathbb{X} : \sup_{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;\leq\; \gamma \right\} \;\subseteq\; \mathbb{X}.$$

If the event $\boldsymbol{x} \in \mathbb{G}$ occurs, then

$$\begin{array}{lll} \|\boldsymbol{g}(\boldsymbol{\theta}_N^\star;\, \boldsymbol{x})\|_\infty &=& \|\boldsymbol{g}(\boldsymbol{\theta}_N^\star;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N^\star)\|_\infty \\[2ex] &\leq& \displaystyle\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;\;\leq\;\; \gamma, \end{array}$$

because $\boldsymbol{g}(\boldsymbol{\theta}_N^\star) = \mathbf{0}$ by Lemma 1. As a consequence, the set

$$\widetilde{\boldsymbol{\Theta}}_N \;\;=\;\; \{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N :\; \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x})\|_\infty \;\leq\; \gamma\}$$

contains the data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ in the event $\boldsymbol{x} \in \mathbb{G}$. Thus, the set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty in the event $\boldsymbol{x} \in \mathbb{G}$.

II. Convergence rate. Suppose that the event $\boldsymbol{x} \in \mathbb{G}$ occurs, so that the set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty. Consider any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$. By the triangle inequality,

$$\begin{array}{lll} \|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N) - \boldsymbol{g}(\boldsymbol{\theta}_N^\star)\|_\infty &\leq& \|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N) - \boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N;\, \boldsymbol{x})\|_\infty + \|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N^\star)\|_\infty \\[2ex] &=& \|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N)\|_\infty + \|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N;\, \boldsymbol{x})\|_\infty \\[2ex] &\leq& \displaystyle\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma, \end{array}$$

because $\boldsymbol{g}(\boldsymbol{\theta}_N^\star) = \mathbf{0}$ by Lemma 1 and $\|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N;\, \boldsymbol{x})\|_\infty \leq \gamma$ for any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$ by construction of the set $\widetilde{\boldsymbol{\Theta}}_N$. In addition, by construction of the set $\mathbb{G}$ and the choice $\gamma \in (0,\, \delta(\epsilon)\,/\,2]$, we obtain

$$\|\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N) - \boldsymbol{g}(\boldsymbol{\theta}_N^\star)\|_\infty \;\leq\; \sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma \;\leq\; 2\,\gamma \;\leq\; \delta(\epsilon).$$

In other words, any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$ satisfies

$$\boldsymbol{g}(\widetilde{\boldsymbol{\theta}}_N) \;\in\; \mathcal{B}_\infty(\boldsymbol{g}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon)),$$

which implies that

$$\widetilde{\boldsymbol{\theta}}_N \;\in\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$$

and hence

$$\widetilde{\boldsymbol{\Theta}}_N \;\subseteq\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon).$$

To establish convergence rates, we relate the radius $\delta(\epsilon)$ of the ℓ_∞-ball $\mathcal{B}_\infty(\boldsymbol{g}(\boldsymbol{\theta}_N^\star),\, \delta(\epsilon))$ to the radius ϵ of the ℓ_∞-ball $\mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)$. By taking advantage of the change-of-variable argument used in Theorem 1—applied to $\boldsymbol{g}(\boldsymbol{\theta}_N)$ instead of $\boldsymbol{\mu}(\boldsymbol{\theta}_N)$—it can be shown that $\delta(\epsilon)$ is related to ϵ as follows:

$$\delta(\epsilon) \;\;\geq\;\; \epsilon\; C_1 \inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \sqrt[p]{|\det(-\mathbb{E}\; \nabla_{\boldsymbol{\theta}_N}^2\; \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))|} \;\;\geq\;\; \epsilon\; C_1\; C_2\; \widetilde{\Lambda},$$

where $C_1 > 0$ and $C_2 > 0$ are constants. We establish convergence rates by choosing ϵ so that the event $\widetilde{\boldsymbol{\Theta}}_N \subseteq \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$ occurs with high probability. To do so, we need to establish uniform convergence of $\boldsymbol{g}(.;\, \boldsymbol{X})$ on $\boldsymbol{\Theta}_N$.

III. Uniform convergence of $\boldsymbol{g}(.;\, \boldsymbol{X})$ on $\boldsymbol{\Theta}_N$. We have seen that, for any $\boldsymbol{x} \in \mathbb{X}$ such that

$$\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma \;\;\leq\;\; \delta(\epsilon),$$

the set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty and satisfies

$$\widetilde{\boldsymbol{\Theta}}_N \;\subseteq\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon).$$

We therefore conclude that

$$\mathbb{P}\left(\widetilde{\boldsymbol{\Theta}}_N \;\subseteq\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)\right) \;\geq\; \mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma \;\leq\; \delta(\epsilon)\right).$$

The fact that $\delta(\epsilon) \geq \epsilon\, C_1\, C_2\, \widetilde{\Lambda}$ implies that

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma \;\leq\; \delta(\epsilon)\right)\\ \geq\;\; &\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma \;\leq\; \epsilon\; C_1\; C_2\; \widetilde{\Lambda}\right).\end{aligned}$$

We bound the probability of the complement of the event on the right-hand side. Observe that

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty + \gamma \;>\; \epsilon\; C_1\; C_2\; \widetilde{\Lambda}\right)\\ &\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;>\; \frac{\epsilon\, C_1\, C_2\, \widetilde{\Lambda}}{2}\right) + \mathbb{P}\left(\gamma > \frac{\epsilon\, C_1\, C_2\, \widetilde{\Lambda}}{2}\right).\end{aligned}$$

The first term on the right-hand side can be bounded by using Lemma 3, which shows that there exists a constant $C_3 > 0$ such that

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;>\; \frac{\epsilon\;\, C_1\, C_2\; \widetilde{\Lambda}}{2}\right)\\ \leq\;\;& 2\, \exp\left(-\frac{\epsilon^2\; C_1^2\;\; C_2^2\; \widetilde{\Lambda}^2}{2\; C_3\; |||\mathcal{D}|||_2^2\;\; \max\{1,\; \max_{1 \leq i \leq M} |\mathfrak{N}_i|\}^2\;\; \Psi^2} + \log\, p\right),\end{aligned}$$

where $|||\mathcal{D}|||_2 \geq 1$ by construction of $\mathcal{D}$ and $\Psi > 0$ by assumption. To ensure that the event $\sup_{\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N; \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty > \epsilon\, C_1\, C_2\, \widetilde{\Lambda} \,/\, 2$ occurs with low probability, choose any $C_4 > 1$ and let

$$\epsilon \;=\; C\, |||\mathcal{D}|||_2 \sqrt{\frac{\log p}{\widetilde{\Lambda}^2 \,/\, \Psi^2}} \;>\; 0, \tag{C.4}$$

where

$$C \;=\; \frac{\sqrt{2\, C_3\, C_4}}{C_1\, C_2}\; \max\left\{1,\; \max_{1 \leq i \leq M} |\mathfrak{N}_i|\right\} \;>\; 0.$$

Observe that $0 < C < \infty$, because Assumption A ensures that $\max_{1 \leq i \leq M} |\mathfrak{N}_i|$ is bounded above by a finite constant. The choice of $\epsilon > 0$ implies that

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;>\; \frac{\epsilon\; C_1\; C_2\; \widetilde{\Lambda}}{2}\right) \;\leq\; 2\, \exp\left(-A\, \log\, p\right),$$

where $A = C_4 - 1 > 0$. The second term, $\mathbb{P}(\gamma > \epsilon\; C_1\, C_2\; \widetilde{\Lambda} \,/\, 2)$, can be bounded above as follows. First, we have assumed that $\gamma \in (0,\, \delta(\epsilon) \,/\, 2]$ and, second, the definition of $\epsilon > 0$ in (C.4) implies that we have

$$\epsilon\; C_1\; C_2\; \widetilde{\Lambda} \;=\; C_5\; |||\mathcal{D}|||_2\; \Psi\; \sqrt{\log\, p},$$

where $C_5 > 0$ is a constant. As a consequence, choosing

$$\gamma \;<\; \frac{C_5}{2}\; |||\mathcal{D}|||_2\; \Psi\; \sqrt{\log p}$$

ensures that

$$\gamma \;<\; \frac{\epsilon\; C_1\; C_2\; \widetilde{\Lambda}}{2}$$

and hence

$$\mathbb{P}\left(\gamma \;>\; \frac{\epsilon\; C_1\; C_2\; \widetilde{\Lambda}}{2}\right) \;=\; 0.$$

We therefore assume that $\gamma \in (0,\, B\, |||\mathcal{D}|||_2\; \Psi\; \sqrt{\log p})$, where $B = C_5 \,/\, 2 > 0$. Upon collecting terms, we obtain

$$\mathbb{P}\left(\widetilde{\boldsymbol{\Theta}}_N \;\subseteq\; \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)\right) \;\geq\; 1 - 2\, \exp\left(-A\, \log\, p\right).$$

Last, but not least, assume that, for any $C_0 > 0$, however large, there exists $N_0 > 0$ such that

$$\widetilde{\Lambda} \;>\; C_0\; |||\mathcal{D}|||_2\; \Psi\; \sqrt{\log p} \quad \text{for all} \quad N > N_0.$$

Then, with at least probability $1 - 2\,\exp(-A\,\log\,p)$, the random set $\widetilde{\mathbf{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\mathbf{\Theta}}_N$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\; \leq \;\; C\,|||\mathcal{D}|||_2 \sqrt{\frac{\log\,p}{\widetilde{\Lambda}^2\,/\,\Psi^2}} \;\; \longrightarrow \;\; 0 \quad \text{as} \quad N \longrightarrow \infty.$$

Remark 2. Extensions to dependent random variables with countable and uncountable sample spaces. Similar to Theorem 1, Theorem 2 is not restricted to random graphs with dependent edges, but could be extended to dependent random variables with countable and uncountable sample spaces: see Remark 1 in Appendix B.

C.1. Auxiliary results.

We prove auxiliary results.

Lemma 2. *Let $\boldsymbol{g} : \mathbf{\Theta}_N \mapsto \mathbb{R}$ be any continuously differentiable function on the open and convex set $\mathbf{\Theta}_N$. If $\boldsymbol{g}(\boldsymbol{\theta})$ is strictly concave on $\mathbf{\Theta}_N$, then its gradient $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})$ exists, is continuous, and is one-to-one.*

Proof of Lemma 2. The existence and continuity of $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})$ on $\mathbf{\Theta}_N$ follow from the assumption that $\boldsymbol{g}(\boldsymbol{\theta})$ is continuously differentiable on the open and convex set $\mathbf{\Theta}_N$. We prove by contradiction that $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})$ is one-to-one on $\mathbf{\Theta}_N$. Suppose that $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})$ is not one-to-one on $\mathbf{\Theta}_N$, that is, there exists $(\boldsymbol{\theta}_1,\, \boldsymbol{\theta}_2) \in \mathbf{\Theta}_N \times \mathbf{\Theta}_N$ such that $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ and $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1} = \nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_2}$. By the strict concavity of $\boldsymbol{g}(\boldsymbol{\theta})$ on $\mathbf{\Theta}_N$,

$$\langle \nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1},\; \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle \;\; > \;\; \boldsymbol{g}(\boldsymbol{\theta}_2) - \boldsymbol{g}(\boldsymbol{\theta}_1) \tag{C.5}$$

and

$$\langle \nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_2},\; \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \;\; > \;\; \boldsymbol{g}(\boldsymbol{\theta}_1) - \boldsymbol{g}(\boldsymbol{\theta}_2). \tag{C.6}$$

By multiplying both sides of (C.6) by -1, we obtain

$$\langle \nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_2},\; \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle \;\; < \;\; \boldsymbol{g}(\boldsymbol{\theta}_2) - \boldsymbol{g}(\boldsymbol{\theta}_1).$$

If $\nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1} = \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_2}$, then

$$\langle \nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1},\; \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle \;\; < \;\; \boldsymbol{g}(\boldsymbol{\theta}_2) - \boldsymbol{g}(\boldsymbol{\theta}_1). \tag{C.7}$$

The conclusion (C.7) contradicts (C.5), so $\nabla_{\boldsymbol{\theta}}\, \boldsymbol{g}(\boldsymbol{\theta})$ is one-to-one on $\mathbf{\Theta}_N$.

Lemma 3. *Assume that Assumption A is satisfied. Then there exists a universal constant $C > 0$ such that, for all $t > 0$,*

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;>\; t\right)$$

$$\leq\;\; 2\, \exp\left(-\frac{2\, t^2}{C\; |||\mathcal{D}|||_2^2\;\, \max\{1,\; \max_{1 \leq i \leq M} |\mathfrak{N}_i|\}^2\; \Psi^2} + \log\, p\right),$$

where $|||\mathcal{D}|||_2 \geq 1$ *and* $\Psi > 0$.

Proof of Lemma 3. First, consider any $\boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N$ and recall that

$$\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) \;\;=\;\; \nabla_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) \;\;=\;\; \sum_{i=1}^{M} \left(s(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-i}}\, s(\boldsymbol{X})\right),$$

where $\mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-i}}$ denotes the conditional expectation, computed with respect to the conditional distribution of X_i given $\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}$. Consider any $i \in \{1, \ldots, M\}$ and any $(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{X} \times \mathbb{X}$ such that $x_j = x'_j$ for all $j \neq i$. Then, by the triangle inequality, we obtain, for each $k \in \{1, \ldots, p\}$,

$$\begin{aligned}
|g_k(\boldsymbol{\theta}_N;\, \boldsymbol{x}) - g_k(\boldsymbol{\theta}_N;\, \boldsymbol{x}')| \;\;&=\;\; \left|\sum_{j=1}^{M} \left(\mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-j}}\, s_k(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}'_{-j}}\, s_k(\boldsymbol{X})\right)\right| \\
&\leq\;\; \sum_{j=1}^{M} \left|\mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-j}}\, s_k(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}_N, \boldsymbol{x}'_{-j}}\, s_k(\boldsymbol{X})\right| \\
&\leq\;\; \max\left\{1,\; \max_{1 \leq i \leq M} |\mathfrak{N}_i|\right\}\;\, \Psi.
\end{aligned}$$

The last inequality follows from Assumption A, which implies that a change of a single edge variable can affect the conditional distributions of at most $\max_{1 \leq i \leq M} |\mathfrak{N}_i|$ other edge variables. As a result,

$$\sup_{(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{X} \times \mathbb{X}:\, x_j = x'_j \text{ for all } j \neq i} |g_k(\boldsymbol{\theta}_N; \boldsymbol{x}) - g_k(\boldsymbol{\theta}_N; \boldsymbol{x}')| \leq \max\left\{1,\; \max_{1 \leq i \leq M} |\mathfrak{N}_i|\right\} \Psi.$$

By applying Theorem 1 of Chazottes et al. [18] to each coordinate $g_k(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - g_k(\boldsymbol{\theta}_N)$ of $\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)$ $(k = 1, \ldots, p)$, we can conclude that there exists a constant $C > 0$ such that, for all $t > 0$,

$$\mathbb{P}\left(|g_k(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - g_k(\boldsymbol{\theta}_N)| > t\right)$$

$$\leq\;\; 2\, \exp\left(-\frac{2\, t^2}{C\, |||\mathcal{D}|||_2^2\; \max\{1,\; \max_{1 \leq i \leq M} |\mathfrak{N}_i|\}^2\; \Psi^2}\right),$$

where $|||\mathcal{D}|||_2 \geq 1$ by construction of $\mathcal{D}$ and $\Psi > 0$ by assumption. A union bound shows that

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;>\; t\right) \\
\leq\;\; & 2\,\exp\left(-\frac{2\,t^2}{C\;|||\mathcal{D}|||_2^2\;\max\{1,\;\max_{1 \leq i \leq M} |\mathfrak{N}_i|\}^2\;\Psi^2} + \log\,p\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\theta}_N \,\in\, \boldsymbol{\Theta}_N} \|\boldsymbol{g}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}_N)\|_\infty \;>\; t\right) \\
\leq\;\; & 2\,\exp\left(-\frac{2\,t^2}{C\;|||\mathcal{D}|||_2^2\;\max\{1,\;\max_{1 \leq i \leq M} |\mathfrak{N}_i|\}^2\;\Psi^2} + \log\,p\right).
\end{aligned}$$

APPENDIX D: PROOFS OF COROLLARIES 1–4

We prove Corollaries 1–4 stated in Section 3 of the manuscript, using the auxiliary results stated in Appendices D.1 and D.2. To prove them, it is convenient to return to the notation used in Section 2 of the manuscript, denoting edge variables by $X_{i,j}$ ($\{i,j\} \subset \mathcal{N}$).

Proof of Corollaries 1–4. According to Theorem 2, there exist universal constants $A > 0$, $B > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$ and all $\gamma \in (0,\, B\,|||\mathcal{D}|||_2\,\Psi\,\sqrt{\log p})$, with at least probability $1 - 2\,\exp(-A\,\log N)$, the random set $\widetilde{\boldsymbol{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\boldsymbol{\Theta}}_N$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\;\leq\;\; C\;|||\mathcal{D}|||_2\,\sqrt{\frac{\log\,N}{\widetilde{\Lambda}^2\,/\,\Psi^2}}.$$

Observe that $\log\,p$ is replaced by $\log N$, because $p = N$ under Model 1 and $p = N+1$ under Models 2–4, and $N < N + 1 \leq 2\;N$ implies that $\log\,p \leq \log 2\;N \leq 2\,\log N$ ($N \geq 2$). We establish the consistency results and convergence rates stated in Corollaries 1–4 by bounding $\widetilde{\Lambda}$, Ψ, and $|||\mathcal{D}|||_2$. All corollaries assume that

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\;\leq\;\; U + \frac{1 - \vartheta}{8}\,\log N, \tag{D.1}$$

where $U > 0$ and $\vartheta \in (1/2,\, 1]$ are constants. In addition, Corollaries 2–4 assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ and $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ ($D \geq 2$).

Bounding $\widetilde{\Lambda}$. By Lemma 5, $\widetilde{\Lambda} = N^{\vartheta - \alpha}$, where $\vartheta \in (1/2,\, 1]$ and $\alpha = 0$ under Models 1, 2, and 4 and $\alpha \in [0,\, 1/2)$ under Model 3.

Bounding Ψ. Recall the definition of Ψ: For each $a \in \{1, \dots, p\}$ and each pair of nodes $\{i, j\} \subset \mathcal{N}$,

$$\Xi_{\{i,j\}}\, s_a \;=\; \sup_{(\boldsymbol{x}, \boldsymbol{x}') \,\in\, \mathbb{X} \times \mathbb{X}:\; x_{k,l} = x'_{k,l} \;\text{ for all } \{k,l\} \neq \{i,j\}} |s_a(\boldsymbol{x}) - s_a(\boldsymbol{x}')|$$

and

$$\Psi \;=\; \max_{1 \leq a \leq p} \|\Xi\, s_a\|_2.$$

We show that $\Psi \leq \sqrt{N}$ under Model 1 and $\Psi \leq \|s_{N+1}\|_{\text{Lip}} \sqrt{N}$ under Models 2–4 and bound $\|s_{N+1}\|_{\text{Lip}}$, where $\|s_{N+1}\|_{\text{Lip}}$ is the Lipschitz coefficient of $s_{N+1}(\boldsymbol{X})$ with respect to the Hamming metric on $\mathbb{X} \times \mathbb{X}$:

- Models 1–4 have sufficient statistics $s_1(\boldsymbol{X}), \dots, s_N(\boldsymbol{X})$, the degrees of nodes $1, \dots, N$, respectively. Since the degrees of nodes are sums of $N - 1$ edge variables $X_{i,j} \in \{0, 1\}$, we have $\|\Xi\, s_a\|_2 = \sqrt{N-1} \leq \sqrt{N}$ $(a = 1, \dots, N)$.
- Models 2–4 include the additional sufficient statistic $s_{N+1}(\boldsymbol{X}) = \sum_{i<j}^{N} X_{i,j}\, I_{i,j}(\boldsymbol{X})$, where

$$I_{i,j}(\boldsymbol{X}) \;=\; \mathbb{1}\left(\sum_{h \,\in\, \mathcal{N}_i \,\cap\, \mathcal{N}_j} X_{i,h}\, X_{j,h} \;>\; 0\right), \qquad \{i, j\} \subset \mathcal{N}.$$

  By the definition of $s_{N+1}(\boldsymbol{X})$, we have $\Xi_{\{i,j\}}\, s_{N+1} = 0$ for all pairs of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \,\cap\, \mathcal{N}_j = \emptyset$. The number of pairs of nodes $\{i, j\} \subset \mathcal{N}$ satisfying $\mathcal{N}_i \,\cap\, \mathcal{N}_j \neq \emptyset$ is bounded above by $N\, D$: For each of the N nodes $i \in \mathcal{N}$, there are at most D distinct nodes $j \in \mathcal{N}_i$ such that $\mathcal{N}_i \,\cap\, \mathcal{N}_j \neq \emptyset$. In addition, it is not too hard to show that $\|s_{N+1}\|_{\text{Lip}} \leq 2\, D + 1 \leq 3\, D$ $(D \geq 1)$, which implies that

$$\|\Xi\, s_{N+1}\|_2 \;\leq\; \sqrt{N\, D\, \|s_{N+1}\|^2_{\text{Lip}}} \;\leq\; 3\, D^2 \sqrt{N}.$$

As a result, Models 1–4 have

$$\Psi \;=\; \max_{1 \leq a \leq p} \|\Xi\, s_a\|_2 \;\leq\; \max\{1,\; 3\, D^2\} \sqrt{N}.$$

Bounding $|||\mathcal{D}|||_2$. Under Model 1, edge variables are independent, which implies that $|||\mathcal{D}|||_2 = 1$ by the construction of the coupling matrix $\mathcal{D}$.

To bound $|||\mathcal{D}|||_2$ under Model 2 with $\alpha = 0$, consider any $\epsilon > 0$. We want to ensure that the event

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\leq\; C\, \max\{1,\, 3\, D^2\}\, |||\mathcal{D}|||_2 \sqrt{\frac{\log N}{N^{2\,\vartheta-1}}} \;<\; \epsilon$$

occurs with at least probability $1 - 2\, \exp(-A \log N)$, which implies that $|||\mathcal{D}|||_2$ must satisfy

$$|||\mathcal{D}|||_2 \;<\; \frac{\epsilon}{C\, \max\{1,\, 3\, D^2\}} \sqrt{\frac{N^{2\,\vartheta-1}}{\log N}}.$$

Define

$$\epsilon^\star \;=\; \frac{\epsilon}{C\, \max\{1,\, 3\, D^2\}} \;>\; 0 \tag{D.2}$$

and assume that there exists $N_0(\epsilon^\star) > 0$ such that

$$|||\mathcal{D}|||_2 \;<\; \epsilon^\star \sqrt{\frac{N^{2\,\vartheta-1}}{\log N}} \quad \text{for all} \quad N > N_0(\epsilon^\star). \tag{D.3}$$

Then the event $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty < \epsilon$ occurs with at least probability $1 - 2\, \exp(-A \log N)$, as desired.

Under Models 3 and 4 and the assumptions of Corollaries 3 and 4, $|||\mathcal{D}|||_2$ is bounded above by a universal constant $1 \leq C < \infty$ by Lemma 10.

Consistency results and convergence rates. By collecting terms, we obtain the following consistency results and convergence rates; note that the constants are recycled from line to line and may vary from model to model.

- **Corollary 1:** We have $\alpha = 0$, $p = N$, $\Lambda = \widetilde{\Lambda} = N^\vartheta$, and $\Psi = \sqrt{N}$. Choose $\gamma = 0$. The independence of edge variables under Model 1 implies that $|||\mathcal{D}|||_2 = 1$ and $\widehat{\boldsymbol{\theta}}_N = \widetilde{\boldsymbol{\theta}}_N$ with probability 1. Thus, by Theorem 1, there exist universal constants $A > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$, with at least probability $1 - 2\, \exp(-A \log N)$, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ and the maximum pseudo-likelihood estimator $\widetilde{\boldsymbol{\theta}}_N$ exist, are unique, and satisfy

$$\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;=\; \|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\leq\; C \sqrt{\frac{\log N}{N^{2\,\vartheta-1}}}.$$

  Thus $\vartheta \in (1/2,\, 1]$ implies $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty = \|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \xrightarrow{\text{p}} 0$ as $N \to \infty$.

- **Corollary 2:** We have $\alpha = 0$, $p = N + 1$, $\widetilde{\Lambda} = N^{\vartheta}$, and $\Psi \leq \max\{1, 3\, D^2\} \sqrt{N}$. To choose $\gamma > 0$, note that

$$B\, |||\mathcal{D}|||_2\, \Psi\, \sqrt{\log\, p} \quad \geq \quad B\, |||\mathcal{D}|||_2\, \sqrt{N \log N}, \qquad B > 0,$$

where $B > 0$ corresponds to the constant $C_5 \,/\, 2 > 0$ in the proof of Theorem 2. We can hence choose any $\gamma \in (0,\; B\, |||\mathcal{D}|||_2\, \sqrt{N\, \log N})$. By Theorem 2, there exist universal constants $A > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$ and all $\gamma \in (0,\; B\, |||\mathcal{D}|||_2\, \sqrt{N\, \log N})$, with at least probability $1 - 2\, \exp(-A\, \log N)$, the random set $\widetilde{\mathbf{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\mathbf{\Theta}}_N$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \quad \leq \quad C\, |||\mathcal{D}|||_2\, \sqrt{\frac{\log\, N}{N^{2\,\vartheta - 1}}}.$$

Thus $\vartheta \in (1/2,\, 1]$ implies $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \xrightarrow{\text{p}} 0$ as $N \to \infty$.

- **Corollary 3:** By assumption, $\alpha \in [0,\, 1/2)$ and $\vartheta = 1$. We have $p = N + 1$, $\widetilde{\Lambda} = N^{1-\alpha}$, and $\Psi \leq \max\{1, 3\, D^2\} \sqrt{N}$. By Lemma 10, $|||\mathcal{D}|||_2$ is bounded above by a universal constant $1 \leq C_1 < \infty$. To choose $\gamma > 0$, note that

$$B\, |||\mathcal{D}|||_2\, \Psi\, \sqrt{\log p} \quad \geq \quad B\, \sqrt{N \log N}, \qquad B > 0,$$

where $B > 0$ corresponds to the constant $C_5 \,/\, 2 > 0$ in the proof of Theorem 2. We can hence choose any $\gamma \in (0,\, B\sqrt{N \log N})$. By Theorem 2, there exist universal constants $A > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$ and all $\gamma \in (0,\, B\, \sqrt{N \log N})$, with at least probability $1 - 2\, \exp(-A\, \log N)$, the random set $\widetilde{\mathbf{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\mathbf{\Theta}}_N$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \quad \leq \quad C\, \sqrt{\frac{\log\, N}{N^{1 - 2\,\alpha}}}.$$

Thus $\alpha \in [0,\, 1/2)$ implies $\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \xrightarrow{\text{p}} 0$ as $N \to \infty$.

- **Corollary 4:** By assumption, $\vartheta = 1$. We have $\alpha = 0$, $p = N + 1$, $\widetilde{\Lambda} = N$, and $\Psi \leq \max\{1, 3\, D^2\} \sqrt{N}$. By an argument along the lines of Corollary 3, $|||\mathcal{D}|||_2$ is bounded above by a universal constant $1 \leq C_1 < \infty$ and γ can be chosen as $\gamma \in (0,\, B\, \sqrt{N \log N})$ $(B > 0)$. Thus, by Theorem 2, there exist universal constants $A > 0$, $C > 0$, and $N_0 > 0$ such that, for all $N > N_0$, with at least probability $1 -$

$2\,\exp(-A\,\log N)$, the random set $\widetilde{\mathbf{\Theta}}_N$ is non-empty and any element $\widetilde{\boldsymbol{\theta}}_N$ of $\widetilde{\mathbf{\Theta}}_N$ satisfies

$$\|\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_N^\star\|_\infty \;\;\leq\;\; C\,\sqrt{\frac{\log\, N}{N}} \longrightarrow 0 \quad \text{as} \quad N \longrightarrow \infty.$$

D.1. Bounding $\widetilde{\Lambda}$.

By definition, $\widetilde{\Lambda} \equiv \widetilde{\Lambda}(N)$ satisfies

$$\inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \sqrt[p]{|\det(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\,\widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))|} \;\;\geq\;\; C\,\widetilde{\Lambda}(N) \;\;>\;\; 0,$$

where $C > 0$ is a universal constant. To bound $\widetilde{\Lambda}$, we establish Lemmas 4–9. To do so, we first introduce the sparse β-model with independent edges, with probability mass function

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \;\;=\;\; \prod_{i<j}^{N} \frac{\exp((\theta_i + \theta_j)\, x_{i,j})\, N^{-\alpha\, x_{i,j}}}{1 + \exp(\theta_i + \theta_j)\, N^{-\alpha}}, \tag{D.4}$$

where $\alpha \in [0,\, 1/2)$ is a known constant, which may be interpreted as the level of sparsity of the random graph.

Lemma 4 bounds the smallest eigenvalue of the expected Hessian $-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\,\widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ under the sparse β-model; note that Model 1 is the special case of the sparse β-model with $\alpha = 0$. Under Models 2 and 3, the expected Hessian $-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\,\widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ is of the form

$$-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\,\widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X}) \;\;=\;\; \begin{pmatrix} \boldsymbol{A}(\boldsymbol{\theta}_N) & \boldsymbol{c}(\boldsymbol{\theta}_N) \\ \boldsymbol{c}(\boldsymbol{\theta}_N)^\top & v(\boldsymbol{\theta}_N) \end{pmatrix}, \tag{D.5}$$

where

- the entries $A_{i,j}(\boldsymbol{\theta}_N)$ of the matrix $\boldsymbol{A}(\boldsymbol{\theta}_N) \in \mathbb{R}^{N\times N}$ are given by

$$A_{i,j}(\boldsymbol{\theta}_N) \;\;=\;\; \sum_{a<b}^{N} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{a,b\}}}(s_i(\boldsymbol{X}),\, s_j(\boldsymbol{X})), \qquad i,j = 1,\dots,N;$$

- the entries $c_i(\boldsymbol{\theta}_N)$ of the vector $\boldsymbol{c}(\boldsymbol{\theta}_N) \in \mathbb{R}^N$ are given by

$$c_i(\boldsymbol{\theta}_N) \;\;=\;\; \sum_{a<b}^{N} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{a,b\}}}(s_i(\boldsymbol{X}),\, s_{N+1}(\boldsymbol{X})), \qquad i = 1,\dots,N;$$

- $v(\boldsymbol{\theta}_N) \in \mathbb{R}^+$ is given by

$$v(\boldsymbol{\theta}_N) \;\;=\;\; \sum_{a<b}^{N} \mathbb{E}\,\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{a,b\}}}\, s_{N+1}(\boldsymbol{X}).$$

Under Model 4, the expected Hessian $-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ involves derivatives of the size-dependent natural parameters $\eta_{i,j}(\theta_{N+1})$, defined by

$$\eta_{i,j}(\theta_{N+1}) \;\;=\;\; \theta_{N+1}\, \log\left(1 + \frac{\log|\mathcal{N}_i\, \cap\, \mathcal{N}_j|}{|\mathcal{N}_i\, \cap\, \mathcal{N}_j|}\right), \qquad \theta_{N+1} \in \mathbb{R}.$$

That said, the canonical form of exponential families is not unique [55, p. 23], so the terms $\log(1 + \log|\mathcal{N}_i \cap \mathcal{N}_j| \,/\, |\mathcal{N}_i \cap \mathcal{N}_j|)$ can be absorbed into the sufficient statistic $s_{N+1}(\boldsymbol{X})$. The advantage of doing so is that the expected expected Hessian $-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{X})$ then has the same form under Models 2, 3, and 4. We therefore absorb the terms $\log(1 + \log|\mathcal{N}_i \cap \mathcal{N}_j| \,/\, |\mathcal{N}_i \cap \mathcal{N}_j|)$ into $s_{N+1}(\boldsymbol{X})$ to streamline the proofs of Lemmas 4–9. As a consequence, we can ignore the terms $\log(1 + \log|\mathcal{N}_i \cap \mathcal{N}_j| \,/\, |\mathcal{N}_i \cap \mathcal{N}_j|)$, because the terms are bounded above and do not affect the rate of growth of $\widetilde{\Lambda}$.

Lemma 4. *Consider the sparse β-model with independent edges, with data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^N$ and a known level of sparsity $\alpha \in [0,\, 1/2)$. Let $\lambda_{\min}(\boldsymbol{\theta}_N)$ be the smallest eigenvalue of the matrix $\boldsymbol{A}(\boldsymbol{\theta}_N) \in \mathbb{R}^{N\times N}$. Then, for all $\epsilon > 0$ and all $N \geq 3$,*

$$\inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \lambda_{\min}(\boldsymbol{\theta}_N) \;\;\geq\;\; \frac{N^{1-\alpha}}{3\, (1 + \exp(2\, (\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)))^2}.$$

If $\boldsymbol{\theta}_N^\star \in \mathbb{R}^N$ satisfies (D.1), then there exists a universal constant $C > 0$ such that

$$\inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \sqrt[N]{|\det(\boldsymbol{A}(\boldsymbol{\theta}_N)|} \;\;\geq\;\; C\, N^{\vartheta-\alpha}.$$

Proof of Lemma 4. By definition,

$$\widetilde{\ell}(\boldsymbol{\theta}_N;\, \boldsymbol{x}) \;\;=\;\; \sum_{i<j}^{N} \log\, \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = x_{i,j} \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}), \qquad \boldsymbol{x} \in \mathbb{X}.$$

It is straightforward to calculate, for each pair of nodes $\{i, j\} \subset \mathcal{N}$ and each pair of coordinates $(t, l) \in \{1, \dots, N\} \times \{1, \dots, N\}$ of $-\nabla^2_{\boldsymbol{\theta}_N} \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{x})$,

$$\begin{aligned} & -\sum_{i<j}^{N} \frac{\partial}{\partial \theta_t \, \partial \theta_l} \log \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = x_{i,j} \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \\ = \; & \sum_{i<j}^{N} \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X})), \end{aligned}$$

where $\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X}))$ denotes the conditional covariance of $s_t(\boldsymbol{X})$ and $s_l(\boldsymbol{X})$, computed with respect to the conditional distribution of $X_{i,j}$ given $\boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}$. We have, for all pairs of nodes $\{i, j\} \subset \mathcal{N}$ and all $\boldsymbol{x}_{-\{i,j\}} \in \{0, 1\}^{\binom{N}{2}-1}$,

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X})) \;\; = \;\; \sum_{h_1 \,\in\, \mathcal{N} \setminus \{t\}} \; \sum_{h_2 \,\in\, \mathcal{N} \setminus \{l\}} \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(X_{t,h_1}, \, X_{l,h_2}).$$

For each pair of nodes $\{i, j\} \subset \mathcal{N}$, we distinguish two cases:

1. If $t \not\in \{i, j\}$ or $l \not\in \{i, j\}$, then $s_t(\boldsymbol{X})$ and $s_l(\boldsymbol{X})$ cannot be both a function of $X_{i,j}$. It then follows that, conditional on $\boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}$,

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X})) \;\; = \;\; 0.$$

2. If $\{t, l\} \subseteq \{i, j\}$, then either $\{t, l\} = \{i, j\}$ or $t = l \in \{i, j\}$. In both cases, $s_t(\boldsymbol{X})$ and $s_l(\boldsymbol{X})$ are functions of $X_{i,j}$. Conditional on $\boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}$, edge variables $X_{a,b}$ corresponding to pairs of nodes $\{a, b\} \neq \{i, j\}$ are almost surely constant, implying

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(X_{t,h_1}, \, X_{l,h_2}) \;\; = \;\; 0$$

for all $\{t, h_1\} \neq \{i, j\}$ and all $\{l, h_2\} \neq \{i, j\}$. We then have

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X})) \;\; = \;\; \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}} \, X_{i,j}.$$

In the special case when $t = l$,

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X})) \;\; = \;\; \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}} \, s_t(\boldsymbol{X}) \;\; = \;\; \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}} \, X_{i,j}.$$

As a result, for all $t \neq l$,

$$\sum_{i<j}^{N} \mathbb{E} \, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_l(\boldsymbol{X})) \;\; = \;\; \mathbb{E} \, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,l\}}} \, X_{t,l}$$

and

$$\sum_{i<j}^{N} \mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}\, s_t(\boldsymbol{X}) \;\; = \;\; \sum_{l \,\in\, \mathcal{N} \setminus \{t\}} \mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,l\}}}\, X_{t,l}.$$

Observe that

$$\mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}\, X_{i,j} \;\; = \;\; \mathbb{V}_{\boldsymbol{\theta}_N}\, X_{i,j} \;\; = \;\; \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1)\, (1 - \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1)),$$

where the conditional variance $\mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}\, X_{i,j}$ equals the unconditional variance $\mathbb{V}_{\boldsymbol{\theta}_N}\, X_{i,j}$ for all $\{i, j\} \subset \mathcal{N}$ and all $\boldsymbol{x}_{-\{i,j\}} \in \{0, 1\}^{\binom{N}{2}-1}$ due to the independence of edge variables under the sparse β-model. Since $N^{-\alpha} \leq 1$ for all $N \geq 1$ and all $\alpha \in [0, 1/2)$ and $-(\theta_i + \theta_j) \leq |\theta_i + \theta_j| \leq 2\, \|\boldsymbol{\theta}_N\|_\infty$,

$$\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1) \;\; = \;\; \frac{\exp(\theta_i + \theta_j)\, N^{-\alpha}}{1 + \exp(\theta_i + \theta_j)\, N^{-\alpha}} \;\; \geq \;\; \frac{N^{-\alpha}}{1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty)}$$

and

$$1 - \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1) \;\; = \;\; \frac{1}{1 + \exp(\theta_i + \theta_j)\, N^{-\alpha}} \;\; \geq \;\; \frac{1}{1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty)}.$$

We thus obtain, for all pairs of nodes $\{i, j\} \subset \mathcal{N}$ and all $\boldsymbol{x}_{-\{i,j\}} \in \{0, 1\}^{\binom{N}{2}-1}$,

$$\mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}\, X_{i,j} \;\; \geq \;\; \frac{N^{-\alpha}}{(1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty))^2}$$

and

$$\mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}\, X_{i,j} \;\; \geq \;\; \frac{N^{-\alpha}}{(1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty))^2}.$$

By invoking Lemma 2.1 of Hillar and Wibisono [38], the smallest eigenvalue $\lambda_{\min}(\boldsymbol{\theta}_N)$ of the matrix $\boldsymbol{A}(\boldsymbol{\theta}_N) \in \mathbb{R}^{N \times N}$ at $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)$ satisfies

$$\lambda_{\min}(\boldsymbol{\theta}_N) \;\; \geq \;\; \frac{N^{-\alpha}\, (N - 2)}{(1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty))^2}.$$

Using the inequality $N - 2 \geq N \,/\, 3$ $(N \geq 3)$, we obtain, for all $N \geq 3$,

$$\lambda_{\min}(\boldsymbol{\theta}_N) \;\; \geq \;\; \frac{N^{-\alpha}\, (N - 2)}{(1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty))^2} \;\; \geq \;\; \frac{N^{1-\alpha}}{3\, (1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty))^2}.$$

By the reverse triangle inequality, $\|\boldsymbol{\theta}_N\|_\infty \leq \|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon$, which implies that

$$\inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \lambda_{\min}(\boldsymbol{\theta}_N) \;\; \geq \;\; \frac{N^{1-\alpha}}{3\, (1 + \exp(2\, (\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon))^2}.$$

We want to show that there exists a universal constant $C > 0$ such that

$$\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \lambda_{\min}(\boldsymbol{\theta}_N) \;\geq\; \frac{N^{1-\alpha}}{3\,(1+\exp(2\,(\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon))^2} \;\geq\; C\, N^{\vartheta-\alpha},$$

where $\vartheta > \alpha$ because $\vartheta \in (1/2,\, 1]$ and $\alpha \in [0,\, 1/2)$. Upon re-arranging terms and using the fact that $\epsilon > 0$, we obtain

$$\frac{N^{(1-\vartheta)/2}}{(3\,C)^{1/2}} \;\geq\; 1 + \exp(2\,(\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)) \;\geq\; \exp(\|\boldsymbol{\theta}_N^\star\|_\infty),$$

so

$$\frac{1-\vartheta}{2}\,\log\, N - \frac{1}{2}\,\log\, 3\,C \;\geq\; \|\boldsymbol{\theta}_N^\star\|_\infty.$$

Choose any $U > 0$ and let $C = \exp(-2\,U)\,/\,3 > 0$. Then there exists a universal constant $C > 0$ such that, for all $N \geq 3$,

$$\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \lambda_{\min}(\boldsymbol{\theta}_N) \;\geq\; C\, N^{\vartheta-\alpha},$$

provided

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\leq\; U + \frac{1-\vartheta}{8}\,\log N \;\leq\; U + \frac{1-\vartheta}{2}\,\log N.$$

Using properties of determinants, we obtain, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$|\det(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\, \tilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))| \;=\; |\det(\boldsymbol{A}(\boldsymbol{\theta}_N))| \;\geq\; |\lambda_{\min}(\boldsymbol{\theta}_N)|^N \;\geq\; (C\, N^{\vartheta-\alpha})^N,$$

so

$$\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\epsilon)} \sqrt[N]{|\det(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\, \tilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))|} \;\geq\; C\, N^{\vartheta-\alpha},$$

provided

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\leq\; U + \frac{1-\vartheta}{8}\,\log N.$$

Lemma 5. *Consider Models 1–4, with data-generating parameter vector* $\boldsymbol{\theta}_N^\star \in \mathbb{R}^p$ *satisfying* (D.1)*. Assume that* $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ *and* $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ *(*$D \geq 0$*). Then, for all* $\epsilon > 0$*, there exist universal constants* $C > 0$ *and* $N_0 > 0$ *such that, for all* $N > N_0$*,*

$$\inf_{\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \sqrt[p]{|\det(-\mathbb{E}\,\nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))|} \;\geq\; C\, N^{\vartheta-\alpha},$$

where $\alpha = 0$ and $\vartheta \in (1/2,\, 1]$ under Models 1, 2, and 4, whereas $\alpha \in [0,\, 1/2)$ and $\vartheta = 1$ under Model 3. As a result,

$$\widetilde{\Lambda} \;\;=\;\; N^{\vartheta - \alpha}.$$

Proof of Lemma 5. Using (D.5), the absolute value of the determinant of the expected Hessian $-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X})$ can be written as

$$|\det(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))| \;=\; \det(\boldsymbol{A}(\boldsymbol{\theta}_N))\; (v(\boldsymbol{\theta}_N) - \boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{A}(\boldsymbol{\theta}_N)^{-1}\, \boldsymbol{c}(\boldsymbol{\theta}_N)).$$

We note that the matrix $\boldsymbol{A}(\boldsymbol{\theta}_N)$ takes the same form as in Lemma 4, although the conditional probabilities of edges do not reduce to the marginal probabilities as in Lemma 4, because Models 2–4 induce dependence amongst edges. Choose $\epsilon > 0$ and assume $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}^\star_N, \epsilon) \subset \mathbb{R}^{N+1}$, and let $\lambda_{\min}(\boldsymbol{\theta}_N)$ be the smallest eigenvalue of $\boldsymbol{A}(\boldsymbol{\theta}_N)$. The proof of Lemma 4 can then be modified to show that there exists a universal constant $C_1 > 0$ such that, for all $\epsilon > 0$, all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}^\star_N,\, \epsilon) \subset \mathbb{R}^{N+1}$ satisfying (D.1), and all $N \geq 3$,

$$\lambda_{\min}(\boldsymbol{\theta}_N) \;\;\geq\;\; C_1\, N^{\vartheta - \alpha}. \tag{D.6}$$

The single deviation from the proof of Lemma 4 is that the edge variables are dependent under Models 2–4, so we need to bound the conditional probabilities of edges rather than the marginal probabilities. The conditional probabilities of edges can be bounded from below by using Lemma 11:

$$\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \;\;\geq\;\; \frac{N^{-\alpha}}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}$$

and

$$1 - \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \;\;\geq\;\; \frac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)},$$

allowing us to conclude there exists a universal constant $C_1 > 0$ such that

$$\det(\boldsymbol{A}(\boldsymbol{\theta}_N)) \;\;\geq\;\; \lambda_{\min}(\boldsymbol{\theta}_N)^N \;\;\geq\;\; (C_1\, N^{\vartheta - \alpha})^N.$$

Thus, under Model 1 with $\alpha = 0$,

$$|\det(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))| \;\;\geq\;\; (C_1\, N^{\vartheta})^N,$$

whereas under Models 2–4,

$$\begin{aligned} |\det(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}_N}\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))| \;\;&\geq\;\; (C_1\, N^{\vartheta - \alpha})^N\, (v(\boldsymbol{\theta}_N) - \boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{A}(\boldsymbol{\theta}_N)^{-1}\, \boldsymbol{c}(\boldsymbol{\theta}_N)) \\ =\;\; (C_1\, N^{\vartheta - \alpha})^N\, (\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{c}(\boldsymbol{\theta}_N))\, &\left(\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{c}(\boldsymbol{\theta}_N)} - R(\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1},\, \boldsymbol{c}(\boldsymbol{\theta}_N))\right), \end{aligned}$$

where

$$R(\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1},\, \boldsymbol{c}(\boldsymbol{\theta}_N)) \;=\; \frac{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{A}(\boldsymbol{\theta}_N)^{-1}\, \boldsymbol{c}(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{c}(\boldsymbol{\theta}_N)}$$

is the Rayleigh quotient of the matrix $\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1} \in \mathbb{R}^{N\times N}$, assuming $\boldsymbol{c}(\boldsymbol{\theta}_N) \in \mathbb{R}^N \setminus \mathbf{0}$. Here, $\mathbf{0} \in \mathbb{R}^N$ denotes the N-dimensional null vector.

We bound the terms $\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)$, $v(\boldsymbol{\theta}_N) \,/\, (\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N))$, and $R(\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1},\, \boldsymbol{c}(\boldsymbol{\theta}_N))$ one by one. First, by Lemma 6, there exists a universal constant $C_2 > 0$ such that, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\geq\; C_2\, N^{\vartheta} \;>\; 0.$$

Second, by Lemma 8, there exists a universal constant $C_3 > 0$ such that, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\geq\; \frac{C_3}{N^{1-\vartheta}} \;>\; 0.$$

Last, but not least, to bound the Rayleigh quotient $R(\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1},\, \boldsymbol{c}(\boldsymbol{\theta}_N))$, note that if $\lambda_1(\boldsymbol{\theta}_N), \ldots, \lambda_N(\boldsymbol{\theta}_N)$ are the eigenvalues of $\boldsymbol{A}(\boldsymbol{\theta}_N)$, then $1 \,/\, \lambda_1(\boldsymbol{\theta}_N), \ldots, 1 \,/\, \lambda_N(\boldsymbol{\theta}_N)$ are the eigenvalues of $\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1}$. Let $\lambda_{\min}(\boldsymbol{\theta}_N)$ be the smallest eigenvalue of $\boldsymbol{A}(\boldsymbol{\theta}_N)$, so $1 \,/\, \lambda_{\min}(\boldsymbol{\theta}_N)$ is the largest eigenvalue of $\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1}$. Since the Rayleigh quotient is bounded above by the largest eigenvalue of $\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1}$, we obtain, using (D.6),

$$R(\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1},\, \boldsymbol{c}(\boldsymbol{\theta}_N)) \;\leq\; \frac{1}{\lambda_{\min}(\boldsymbol{\theta}_N)} \;\leq\; \frac{1}{C_1\, N^{\vartheta-\alpha}}.$$

As a result,

$$\begin{aligned}
&(C_1\, N^{\vartheta-\alpha})^N\, (\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)) \left(\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)} - R(\boldsymbol{A}(\boldsymbol{\theta}_N)^{-1},\, \boldsymbol{c}(\boldsymbol{\theta}_N))\right)\\
\geq\;& (C_1\, N^{\vartheta-\alpha})^N\, C_2\, N^{\vartheta} \left(\frac{C_3}{N^{1-\vartheta}} - \frac{1}{C_1\, N^{\vartheta-\alpha}}\right)\\
=\;& C_1^N\, C_2\, N^{(\vartheta-\alpha)\, N}\, \frac{N^{\vartheta}}{N^{1-\vartheta}} \left(C_3 - \frac{1}{C_1\, N^{2\,\vartheta-\alpha-1}}\right)\\
=\;& C_1^N\, C_2\, N^{(\vartheta-\alpha)\,(N+1)}\, \frac{N^{\alpha}}{N^{1-\vartheta}} \left(C_3 - \frac{1}{C_1\, N^{2\,\vartheta-\alpha-1}}\right).
\end{aligned}$$

By assumption, $2\,\vartheta - \alpha - 1 > 0$, because $\alpha = 0$ and $\vartheta \in (1/2,\, 1]$ under Models 1, 2, and 4, whereas $\alpha \in [0,\, 1/2)$ and $\vartheta = 1$ under Model 3. Thus,

$N^{2\,\vartheta-\alpha-1} \to \infty$ as $N \to \infty$ and there exist universal constants $C_4 > 0$, $C_5 > 0$, and $N_0 > 0$ such that, for all $N > N_0$,

$$\begin{aligned}\inf_{\boldsymbol{\theta}_N \,\in\, \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)} \sqrt[N+1]{|\det(-\mathbb{E}\,\nabla_{\boldsymbol{\theta}_N}^2\, \widetilde{\ell}(\boldsymbol{\theta}_N; \boldsymbol{X}))|} &\geq C_4\, N^{\vartheta-\alpha} \left(\frac{N^\alpha}{N^{1-\vartheta}}\right)^{\frac{1}{N+1}} \\ &\geq C_5\, N^{\vartheta-\alpha}.\end{aligned}$$

We conclude that $\widetilde{\Lambda} = N^{\vartheta-\alpha}$.

Lemma 6. *Consider Models 2–4, with data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfying (D.1). Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ and $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ ($D \geq 2$). Then, for all $\epsilon > 0$, all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$, and all $N \geq 3$,*

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\; \geq \;\; C\, N^\vartheta,$$

where $C > 0$ is a universal constant.

Proof of Lemma 6. By Lemma 7, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\; \geq \;\; \frac{N}{(1 + \exp((3 + 2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)))^8}.$$

Observe that the lower bound on $\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)$ does not involve α, because the elements of the vector $\boldsymbol{c}(\boldsymbol{\theta}_N) \in \mathbb{R}^N$ correspond to the covariances of the degrees of nodes and the number of brokered edges of nodes with intersecting neighborhoods, and edges among nodes in intersecting neighborhoods are not penalized by α (neither under Model 3 nor under Models 1, 2, and 4).

We want to demonstrate that there exists a universal constant $C > 0$ such that

$$\frac{N}{(1 + \exp((3 + 2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)))^8} \;\; \geq \;\; C\,N^\vartheta.$$

Upon re-arranging terms and using $D \geq 0$ along with $\epsilon > 0$, we obtain

$$\frac{N^{(1-\vartheta)/8}}{C^{1/8}} \;\; \geq \;\; 1 + \exp\left((3 + 2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)\right) \;\; \geq \;\; \exp\left(\|\boldsymbol{\theta}_N^\star\|_\infty\right),$$

so

$$\frac{1-\vartheta}{8} \log N - \frac{1}{8} \log C \;\; \geq \;\; \|\boldsymbol{\theta}_N^\star\|_\infty.$$

Choose any $U > 0$ and let $C = \exp(-8\,U) > 0$. Then, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\; \geq \;\; C\,N^\vartheta,$$

provided

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\leq\; U + \frac{1-\vartheta}{8}\log N.$$

Lemma 7. *Consider Models 2–4, with data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfying (D.1). Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ and $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ ($D \geq 2$). Then, for all $\epsilon > 0$, all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star, \epsilon)$, and all $N \geq 3$,*

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\geq\; \frac{N}{(1+\exp((3+2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)))^8}.$$

Proof of Lemma 7. The coordinates of the vector $\boldsymbol{c}(\boldsymbol{\theta}_N) \in \mathbb{R}^N$ are given by

$$\begin{aligned}
c_t(\boldsymbol{\theta}_N) &= \sum_{i<j}^N \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_{N+1}(\boldsymbol{X})) \\
&= \sum_{i<j}^N \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}\left(\sum_{a \in \mathcal{N}\setminus\{t\}} X_{t,a},\; s_{N+1}(\boldsymbol{X})\right) \\
&= \sum_{i<j}^N \sum_{a \in \mathcal{N}\setminus\{t\}} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}(X_{t,a},\; s_{N+1}(\boldsymbol{X})) \\
&= \sum_{a \in \mathcal{N}\setminus\{t\}} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\; s_{N+1}(\boldsymbol{X})) \\
&= \sum_{a \in \mathcal{N}\setminus\{t\}:\, \mathcal{N}_a \cap \mathcal{N}_t \neq \emptyset} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\; s_{N+1}(\boldsymbol{X})),
\end{aligned}$$

noting that edges $X_{t,a}$ ($\{t,a\} \neq \{i,j\}$) are almost surely constant conditional on $\boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}$ and that $s_{N+1}(\boldsymbol{X})$ is not a function of edges $X_{t,a}$ satisfying $\mathcal{N}_t \cap \mathcal{N}_a = \emptyset$, and thus is almost surely constant under the conditional distribution of $X_{t,a}$ given $\boldsymbol{X}_{-\{t,a\}} = \boldsymbol{x}_{-\{t,a\}}$. We obtain

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;=\; \sum_{t=1}^N \left(\sum_{a \in \mathcal{N}\setminus\{t\}:\, \mathcal{N}_a \cap \mathcal{N}_t \neq \emptyset} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\; s_{N+1}(\boldsymbol{X}))\right)^2.$$

It is therefore enough to demonstrate that

$$\begin{aligned} &\left(\sum_{a \,\in\, \mathcal{N} \setminus \{t\}:\, \mathcal{N}_a \,\cap\, \mathcal{N}_t \,\neq\, \emptyset} \mathbb{E}\, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X}))\right)^2 \\ \geq\;\; & \frac{1}{(1 + \exp((3 + 2\, D)\, (\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)))^8}. \end{aligned}$$

Let

$$I_{t,a}(\boldsymbol{X}) \;=\; \mathbb{1}\left(\sum_{h \,\in\, \mathcal{N}_t \,\cap\, \mathcal{N}_a} X_{t,h}\, X_{a,h} \;>\; 0\right), \qquad \{t, a\} \subset \mathcal{N},$$

and note that

$$s_{N+1}(\boldsymbol{X}) \;=\; \sum_{i<j}^{N} X_{i,j}\, I_{i,j}(\boldsymbol{X}),$$

where $I_{i,j}(\boldsymbol{X}) = 0$ almost surely for all $\{i, j\} \subset \mathcal{N}$ satisfying $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$. Using properties of covariances,

$$\mathbb{E}\, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})) \;=\; \sum_{i<j}^{N} \mathbb{E}\, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, X_{i,j}\, I_{i,j}(\boldsymbol{X})).$$

The FKG inequality [27] implies that

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{t,a\}}}(X_{t,a},\, I_{i,j}(\boldsymbol{X})) \;\geq\; 0 \quad \text{for all} \quad \boldsymbol{x}_{-\{t,a\}} \in \{0,1\}^{\binom{N}{2}-1},$$

because the conditional covariance is computed with respect to the conditional distribution of $X_{t,a}$ and both $X_{t,a}$ and $I_{i,j}(\boldsymbol{X})$ are monotone non-decreasing functions of $X_{t,a}$. Hence

$$\mathbb{E}\, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})) \;\geq\; \mathbb{E}\, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, X_{t,a}\; I_{t,a}(\boldsymbol{X})).$$

Each node $t \in \mathcal{N}$ belongs to one or more subpopulations $\mathcal{A}_k$ for some $k \in \{1, \ldots, K\}$. Since $|\mathcal{A}_l| \geq 3$ for all $l \in \{1, \ldots, K\}$, there exists a node $b \in \mathcal{N}_i \,\cap\, \mathcal{N}_j$ such that

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}_N}(X_{t,a}\; I_{t,a}(\boldsymbol{X}) = 1 \mid \boldsymbol{X}_{-\{t,a\}} = \boldsymbol{x}_{-\{t,a\}}) \\ \geq\;\; & \mathbb{P}_{\boldsymbol{\theta}_N}(X_{t,a}\; X_{t,b}\; X_{a,b} = 1 \mid \boldsymbol{X}_{-\{t,a\}} = \boldsymbol{x}_{-\{t,a\}}), \end{aligned}$$

because the event $X_{t,a}\, X_{t,b}\, X_{a,b} = 1$ implies the event $X_{t,a}\, I_{t,a}(\boldsymbol{X}) = 1$. By Lemma 11,

$$\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \quad\geq\quad \dfrac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}$$

and

$$\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 0 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \quad\geq\quad \dfrac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}$$

for all pairs of nodes $\{i, j\} \subset \mathcal{N}$ satisfying $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$. We can partition the sample space of $\boldsymbol{X}_{-\{t,a\}}$ based on whether $I_{t,a}(\boldsymbol{X}) = 0$ or $I_{t,a}(\boldsymbol{X}) = 1$. When $I_{t,a}(\boldsymbol{X}) = 0$,

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, X_{t,a}\, I_{t,a}(\boldsymbol{X})) \quad=\quad \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, 0) \quad=\quad 0$$

and when $I_{t,a}(\boldsymbol{X}) = 1$,

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, X_{t,a}\, I_{t,a}(\boldsymbol{X})) \quad=\quad \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}\, X_{t,a}.$$

Using the above bounds, we obtain

$$\mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}\, X_{t,a} \quad\geq\quad \left(\dfrac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}\right)^2.$$

Thus, using the law of total expectation,

$$\begin{aligned}
&\mathbb{E}\, \mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, X_{t,a}\, I_{t,a}(\boldsymbol{X})) \\
=\quad& \mathbb{P}(I_{t,a}(\boldsymbol{X}) = 1)\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}\, X_{t,a} \\
\geq\quad& \mathbb{P}(X_{t,b}\, X_{a,b} = 1)\, \left(\dfrac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}\right)^2 \\
\geq\quad& \left(\dfrac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N^\star\|_\infty)}\right)^2 \left(\dfrac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}\right)^2 \\
\geq\quad& \dfrac{1}{(1 + \exp((3 + 2\, D)\, \max\{\|\boldsymbol{\theta}_N^\star\|_\infty,\, \|\boldsymbol{\theta}_N\|_\infty\}))^4}.
\end{aligned}$$

Thus, we have shown that, for all $t \in \{1, \ldots, N\}$,

$$c_t(\boldsymbol{\theta}_N)^2 \quad\geq\quad \dfrac{1}{(1 + \exp((3 + 2\, D)\, \max\{\|\boldsymbol{\theta}_N^\star\|_\infty,\, \|\boldsymbol{\theta}_N\|_\infty\}))^8},$$

which in turn implies that

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\; \geq \;\; \frac{N}{(1+\exp((3+2\,D)\,\max\{\|\boldsymbol{\theta}_N^\star\|_\infty,\,\|\boldsymbol{\theta}_N\|_\infty\}))^8}.$$

Since $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$, we know that $\|\boldsymbol{\theta}_N\|_\infty \leq \|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon$ by the reverse triangle inequality, which implies that, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \;\; \geq \;\; \frac{N}{(1+\exp((3+2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty+\epsilon)))^8}.$$

Lemma 8. *Consider Models 2–4, with data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfying (D.1). Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ and $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ $(D \geq 2)$. Then, for all $\epsilon > 0$, all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$, and all $N \geq 3$,*

$$\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\; \geq \;\; \frac{C}{N^{1-\vartheta}},$$

where $C > 0$ is a universal constant.

Proof of Lemma 8. By Lemma 9, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\; \geq \;\; \frac{1}{D^8\,(1+\exp((3+2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty+\epsilon)))^4}.$$

We show that there exists a universal constant $C > 0$ such that

$$\frac{1}{D^8\,(1+\exp((3+2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty+\epsilon)))^4} \;\; \geq \;\; C\,\frac{1}{N^{1-\vartheta}}.$$

Upon rearranging terms, we obtain

$$\frac{N^{(1-\vartheta)/4}}{(C\,D^8)^{1/4}} \;\; \geq \;\; 1+\exp((3+2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty+\epsilon)) \;\; \geq \;\; \exp(\|\boldsymbol{\theta}_N^\star\|_\infty),$$

using $D \geq 0$ and $\epsilon > 0$. Upon taking the natural logarithm on both sides, we obtain

$$\frac{1-\vartheta}{4}\,\log N - \frac{1}{4}\,\log C\,D^8 \;\; \geq \;\; \|\boldsymbol{\theta}_N^\star\|_\infty.$$

Choose any $U > 0$ and let $C = \exp(-4\,U)\,/\,D^8 > 0$. Then

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\; \leq \;\; U + \frac{1-\vartheta}{4}\,\log N.$$

Thus, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\;\geq\;\; \frac{C}{N^{1-\vartheta}},$$

provided

$$\|\boldsymbol{\theta}_N^\star\|_\infty \;\;\leq\;\; U + \frac{1-\vartheta}{8}\log N \;\;\leq\;\; U + \frac{1-\vartheta}{4}\log N.$$

Lemma 9. *Consider Models 2–4, with data-generating parameter vector $\boldsymbol{\theta}_N^\star \in \mathbb{R}^{N+1}$ satisfying (D.1). Assume that $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$ and $\max_{1 \leq i \leq N} |\mathcal{N}_i| < D$ ($D \geq 2$). Then, for all $\epsilon > 0$, all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$, and all $N \geq 3$,*

$$\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\;\geq\;\; \frac{1}{D^8\,(1+\exp((3+2\,D)\,(\|\boldsymbol{\theta}_N^\star\|_\infty+\epsilon)))^4}.$$

Proof of Lemma 9. We have

$$\begin{aligned} v(\boldsymbol{\theta}_N) \;&=\; \sum_{i<j}^{N} \mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}\, s_{N+1}(\boldsymbol{X}) \\ &=\; \sum_{i<j}^{N} \mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}} \left(\sum_{a<b}^{N} X_{a,b}\, I_{a,b}(\boldsymbol{X})\right), \end{aligned}$$

where

$$I_{a,b}(\boldsymbol{X}) \;=\; \mathbb{1}\left(\sum_{h\,\in\,\mathcal{N}_a\,\cap\,\mathcal{N}_b} X_{a,h}\, X_{b,h} \;>\; 0\right), \qquad \{a,b\} \subset \mathcal{N}.$$

Given any pair of nodes $\{i,j\} \subset \mathcal{N}$, notice that each random variable $X_{a,b}\, I_{a,b}(\boldsymbol{X})$ is a monotone non-decreasing function of $X_{i,j}$ and, conditional on $\boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}$, all other edge variables are constant. As a consequence, the FKG inequality [27] implies that

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{x}_{-\{i,j\}}}(X_{a,b}\, I_{a,b}(\boldsymbol{X}),\; X_{r,t}\, I_{r,t}(\boldsymbol{X})) \;\;\geq\;\; 0$$

for all pairs of nodes $\{a,b\} \subset \mathcal{N}$ and $\{r,t\} \subset \mathcal{N}$. Thus,

$$\begin{aligned} v(\boldsymbol{\theta}_N) \;&=\; \sum_{i<j}^{N} \mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}} \left(\sum_{a<b}^{N} X_{a,b}\, I_{a,b}(\boldsymbol{X})\right) \\ &\geq\; \sum_{i<j}^{N} \sum_{a<b}^{N} \mathbb{E}\, \mathbb{V}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}} \left(X_{a,b}\, I_{a,b}(\boldsymbol{X})\right). \end{aligned}$$

Using the law of total expectation,

$$
\begin{aligned}
& \mathbb{E}\,\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,(X_{i,j}\,I_{a,b}(\boldsymbol{X})) \\
= \;& \mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{a,b}\,I_{a,b}(\boldsymbol{X}) \mid I_{a,b}(\boldsymbol{X}) = 1)\;\mathbb{P}(I_{a,b}(\boldsymbol{X}) = 1) \\
+ \;& \mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{a,b}\,I_{a,b}(\boldsymbol{X}) \mid I_{a,b}(\boldsymbol{X}) = 0)\;\mathbb{P}(I_{a,b}(\boldsymbol{X}) = 0),
\end{aligned}
$$

which shows

$$
\mathbb{E}\,\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,(X_{i,j}\,I_{a,b}(\boldsymbol{X})) \;\; = \;\; \mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{a,b} \mid I_{a,b}(\boldsymbol{X}) = 1)\;\mathbb{P}(I_{a,b}(\boldsymbol{X}) = 1),
$$

owing to the fact that $X_{a,b}\,I_{a,b}(\boldsymbol{X}) = 0$ almost surely conditional on the event $I_{a,b}(\boldsymbol{X}) = 0$. Hence

$$
\begin{aligned}
v(\boldsymbol{\theta}_N) \;\; & \geq \;\; \sum_{i<j}^{N} \sum_{a<b}^{N} \mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{a,b} \mid I_{a,b}(\boldsymbol{X}) = 1)\;\; \mathbb{P}(I_{a,b}(\boldsymbol{X}) = 1) \\
& = \;\; \sum_{i<j}^{N} \mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{i,j} \mid I_{i,j}(\boldsymbol{X}) = 1)\;\; \mathbb{P}(I_{i,j}(\boldsymbol{X}) = 1) \\
& = \;\; \sum_{i<j\,:\,\mathcal{N}_i \,\cap\, \mathcal{N}_j \neq \emptyset}^{N} \mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{i,j} \mid I_{i,j}(\boldsymbol{X}) = 1)\;\; \mathbb{P}(I_{i,j}(\boldsymbol{X}) = 1),
\end{aligned}
$$

noting that $\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{a,b} = 0$ for all $\{a,b\} \neq \{i,j\}$ and $\mathbb{P}(I_{i,j}(\boldsymbol{X}) = 1) = 0$ for all $\{i,j\} \subset \mathcal{N}$ satisfying $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$. We bound

$$
\mathbb{E}\,(\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{X}_{-\{i,j\}}}\,X_{i,j} \mid I_{i,j}(\boldsymbol{X}) = 1)
$$

from below by noting that, for any $\boldsymbol{x} \in \{0,1\}^{\binom{N}{2}}$ with $I_{i,j}(\boldsymbol{x}) = 1$,

$$
\mathbb{V}_{\boldsymbol{\theta}_N,\boldsymbol{x}_{-\{i,j\}}}\,X_{i,j} \;\; \geq \;\; \frac{1}{(1 + \exp((3 + 2\,D)\,\|\boldsymbol{\theta}_N\|_\infty))^2}. \tag{D.7}
$$

The lower bound in (D.7) follows from Lemma 11, which shows that, for all $\boldsymbol{x}_{-\{i,j\}} \in \{0,1\}^{\binom{N}{2}-1}$,

$$
\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 1 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \;\; \geq \;\; \frac{1}{1 + \exp((3 + 2\,D)\,\|\boldsymbol{\theta}_N\|_\infty)}
$$

and

$$
\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = 0 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \;\; \geq \;\; \frac{1}{1 + \exp((3 + 2\,D)\,\|\boldsymbol{\theta}_N\|_\infty)}.
$$

Hence

$$\begin{aligned} v(\boldsymbol{\theta}_N) \;&\geq\; \sum_{i<j:\, \mathcal{N}_i \,\cap\, \mathcal{N}_j \,\neq\, \emptyset}^{N} \frac{1}{(1+\exp((3+2\,D)\,\|\boldsymbol{\theta}_N\|_\infty))^2}\; \mathbb{P}(I_{i,j}(\boldsymbol{X})=1) \\ &\geq\; \sum_{i<j:\, \mathcal{N}_i \,\cap\, \mathcal{N}_j \,\neq\, \emptyset}^{N} \frac{1}{(1+\exp((3+2\,D)\, \max\{\|\boldsymbol{\theta}_N\|_\infty, \|\boldsymbol{\theta}_N^\star\|_\infty\}))^4}. \end{aligned}$$

The lower bound on the probability of event $I_{i,j}(\boldsymbol{X}) = 1$ follows from the observation that, for any given node $h \in \mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$, the event $X_{i,h}\, X_{j,h} = 1$ implies the event $I_{i,j}(\boldsymbol{X}) = 1$, so

$$\mathbb{P}(I_{i,j}(\boldsymbol{X})=1) \;\geq\; \mathbb{P}(X_{i,h}\, X_{j,h}=1) \;=\; \mathbb{P}(X_{i,h}=1 \mid X_{j,h}=1)\; \mathbb{P}(X_{j,h}=1).$$

By using the law of total probability along with the observation that, for all $\boldsymbol{x}_{-\{i,j\}} \in \{0,1\}^{\binom{N}{2}-1}$,

$$\mathbb{P}(X_{i,j}=1 \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \;\;\geq\;\; \frac{1}{1+\exp((3+2\,D)\,\|\boldsymbol{\theta}_N^\star\|_\infty)},$$

one can show that both the conditional and unconditional probabilities are bounded below by $1\,/\,(1+\exp((3+2\,D)\,\|\boldsymbol{\theta}_N^\star\|_\infty))$. Last, but not least, we consider the number of pairs of nodes $\{i,j\} \subset \{1,\ldots,N\}$ for which $\mathcal{N}_i \,\cap\, \mathcal{N}_j \,\neq\, \emptyset$. There exists, for each node $i \in \mathcal{N}$, a node $h \in \mathcal{N} \setminus \{i\}$ such that $\mathcal{N}_i \,\cap\, \mathcal{N}_h \neq \emptyset$, so

$$\begin{aligned} &\sum_{i<j:\, \mathcal{N}_i \,\cap\, \mathcal{N}_j \,\neq\, \emptyset}^{N} \frac{1}{(1+\exp((3+2\,D)\, \max\{\|\boldsymbol{\theta}_N\|_\infty, \|\boldsymbol{\theta}_N^\star\|_\infty\}))^2} \\ \geq\;\; &\frac{N}{(1+\exp((3+2\,D)\, \max\{\|\boldsymbol{\theta}_N\|_\infty, \|\boldsymbol{\theta}_N^\star\|_\infty\}))^4}, \end{aligned}$$

which implies that

$$v(\boldsymbol{\theta}_N) \;\;\geq\;\; \frac{N}{(1+\exp((3+2\,D)\, \max\{\|\boldsymbol{\theta}_N\|_\infty, \|\boldsymbol{\theta}_N^\star\|_\infty\}))^4}.$$

Next, we proceed to demonstrate that there exists a universal constant $C_1 > 0$ such that

$$\boldsymbol{c}(\boldsymbol{\theta}_N)^\top\, \boldsymbol{c}(\boldsymbol{\theta}_N) \;\;\leq\;\; C_1\, N.$$

The coordinates of the vector $\boldsymbol{c}(\boldsymbol{\theta}_N) \in \mathbb{R}^N$ are given by

$$c_t(\boldsymbol{\theta}_N) \;=\; \sum_{i<j}^{N} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_{N+1}(\boldsymbol{X})), \qquad t = 1, \ldots, N,$$

which may be written as

$$\begin{aligned}
&\sum_{i<j}^{N} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{i,j\}}}(s_t(\boldsymbol{X}), s_{N+1}(\boldsymbol{X})) \\
&= \sum_{a \,\in\, \mathcal{N} \setminus \{t\}} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})) \\
&= \sum_{a \,\in\, \mathcal{N} \setminus \{t\}:\, \mathcal{N}_a \,\cap\, \mathcal{N}_t \,\neq\, \emptyset} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})),
\end{aligned}$$

noting $s_{N+1}(\boldsymbol{X})$ is not a function of edge variables $X_{t,a}$ when $\mathcal{N}_t \cap \mathcal{N}_a = \emptyset$, and is hence almost surely constant under the conditional distribution of $X_{t,a}$ given $\boldsymbol{X}_{-\{t,a\}} = \boldsymbol{x}_{-\{t,a\}}$. Using $\max_{1 \leq i \leq N} |\mathcal{N}_i| \leq D$ $(D \geq 0)$, we have

$$\begin{aligned}
&\sum_{a \,\in\, \mathcal{N} \setminus \{t\}:\, \mathcal{N}_a \,\cap\, \mathcal{N}_t \,\neq\, \emptyset} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})) \\
&\leq\; D^2 \max_{a \,\in\, \mathcal{N} \setminus \{t\}:\, \mathcal{N}_a \,\cap\, \mathcal{N}_t \,\neq\, \emptyset} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})),
\end{aligned}$$

because $\mathcal{N}_a \cap \mathcal{N}_t \neq \emptyset$ if one of the two following conditions is satisfied:

- $a \in \mathcal{N}_t$, the number of which is bounded above by $\max_{1 \leq i \leq N} |\mathcal{N}_i| \leq D$;
- $a \notin \mathcal{N}_t$, but there exists $h \in \mathcal{N}_t$ such that $h \in \mathcal{N}_a$; the number of such a for each $h \in \mathcal{N}_t$ is bounded above by $\max_{1 \leq i \leq N} |\mathcal{N}_i| \leq D$. Noting that $|\mathcal{N}_t| \leq \max_{1 \leq i \leq N} |\mathcal{N}_i| \leq D$, the total number of such $a \in \mathcal{N} \setminus \{t\}$ satisfying $\mathcal{N}_a \cap \mathcal{N}_t \neq \emptyset$ is bounded above by D^2.

To bound $\max_{a \,\in\, \mathcal{N} \setminus \{t\}:\, \mathcal{N}_a \,\cap\, \mathcal{N}_t \,\neq\, \emptyset} \mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X}))$, recall that $s_{N+1}(\boldsymbol{X}) = \sum_{i<j}^{N} X_{i,j}\, I_{i,j}(\boldsymbol{X})$. The number of indicator functions $I_{i,j}(\boldsymbol{X})$ that are functions of $X_{t,a}$ and are not constant conditional on $X_{-\{t,a\}}$ is bounded above by D^2, applying the same argument as above. Hence,

$$\mathbb{E}\,\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})) \;\leq\; D^2,$$

because

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, X_{i,j}\, I_{i,j}(\boldsymbol{X})) \;\leq\; 1$$

for all $\{i, j\} \subset \{1, \dots, N\}$, implying

$$\mathbb{C}_{\boldsymbol{\theta}_N, \boldsymbol{X}_{-\{t,a\}}}(X_{t,a},\, s_{N+1}(\boldsymbol{X})) \;\; \leq \;\; D^4.$$

Thus, $c_t(\boldsymbol{\theta}_N) \leq D^4$, so $\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N) \leq D^8\, N$.

Collecting terms shows that

$$\begin{aligned} \frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\; &\geq \;\; \frac{N}{(1 + \exp((3 + 2\, D)\, \max\{\|\boldsymbol{\theta}_N\|_\infty,\, \|\boldsymbol{\theta}_N^\star\|_\infty\}))^4} \left(\frac{1}{D^8\, N}\right) \\ &= \;\; \frac{1}{D^8\, (1 + \exp((3 + 2\, D)\, \max\{\|\boldsymbol{\theta}_N\|_\infty,\, \|\boldsymbol{\theta}_N^\star\|_\infty\}))^4}. \end{aligned}$$

Since $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon) \subset \mathbb{R}^{N+1}$, the reverse implies that $\|\boldsymbol{\theta}_N\|_\infty \leq \|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon$, which in turn implies that, for all $\boldsymbol{\theta}_N \in \mathcal{B}_\infty(\boldsymbol{\theta}_N^\star,\, \epsilon)$,

$$\frac{v(\boldsymbol{\theta}_N)}{\boldsymbol{c}(\boldsymbol{\theta}_N)^\top \boldsymbol{c}(\boldsymbol{\theta}_N)} \;\; \geq \;\; \frac{1}{D^8\, (1 + \exp((3 + 2\, D)\, (\|\boldsymbol{\theta}_N^\star\|_\infty + \epsilon)))^4}.$$

D.2. Bounding the spectral norm of the coupling matrix. To bound the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$, we first review undirected graphical models encoding the conditional independence structure of generalized β-models with dependent edges in Appendices D.2.1 and D.2.2, and then bound $|||\mathcal{D}|||_2$ by using the conditional independence properties in Appendix D.2.3. Auxiliary results can be found in Appendix D.2.4.

D.2.1. *Undirected graphical models of random graphs.* Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be an undirected graph with a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. An undirected graphical model of a random graph [53] is a family of probability measures $\{\mathbb{P}_{\boldsymbol{\theta}_N},\, \boldsymbol{\theta}_N \in \boldsymbol{\Theta}_N\}$ dominated by a σ-finite measure $\nu : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$, with densities of the form

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \;\; \propto \;\; \prod_{\mathcal{C} \in \mathfrak{C}} g_{\mathcal{C}}(\boldsymbol{x}_{\mathcal{C}};\, \boldsymbol{\theta}_N), \qquad \boldsymbol{x} \in \mathbb{X}, \tag{D.8}$$

where $\mathfrak{C}$ is the set of all maximal complete subsets of the conditional independence graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with set of vertices $\mathcal{V} = \{X_1, \dots, X_M\}$ and set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. The functions $g_{\mathcal{C}} : \mathbb{X} \times \boldsymbol{\Theta}_N \mapsto \mathbb{R}^+ \cup \{0\}$ are nonnegative functions defined on the maximal complete subsets $\mathcal{C} \in \mathfrak{C}$ of the conditional independence graph $\mathcal{G}$. Here, as elsewhere [e.g., 52], a complete subset of the conditional independence graph $\mathcal{G}$ is a subset of vertices such that each pair of vertices is connected by an edge, and a complete subset is

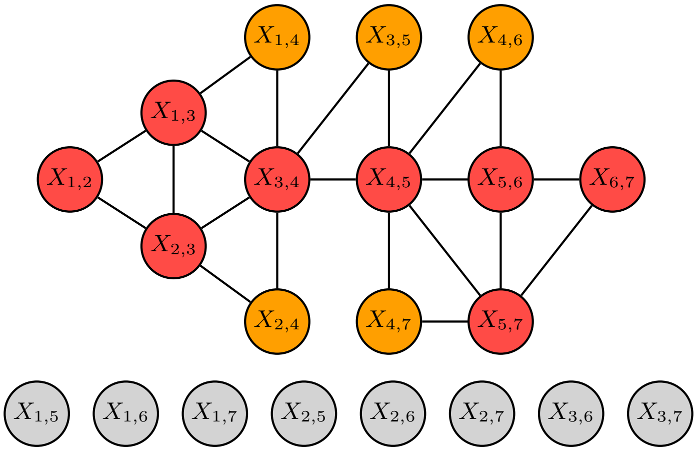

Fig 4. *The conditional independence graph of Models 2–4 with population of nodes* $\mathcal{N} = \{1, \ldots, 7\}$ *consisting of overlapping subpopulations* $\mathcal{A}_1 = \{1, 2, 3\}$, $\mathcal{A}_2 = \{3, 4\}$, $\mathcal{A}_3 = \{4, 5\}$, *and* $\mathcal{A}_4 = \{5, 6, 7\}$. *If nodes* i *and* j *belong to the same subpopulation, edge variable* $X_{i,j}$ *is colored red. If nodes* i *and* j *do not belong to the same subpopulation, edge variable* $X_{i,j}$ *is colored orange if the subpopulations of* i *and* j *overlap and is colored gray otherwise.*

maximal complete if no vertices can be added without losing the property of completeness. It is well-known that the factorization property of probability density function (D.8) implies conditional independence properties, that is, Markov properties. The Markov properties of undirected graphical models are reviewed in Lauritzen [52].

The probability density functions introduced in Section 2 are of the form

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \;\propto\; \prod_{i<j}^{N} \varphi_{i,j}(\boldsymbol{x}_{\mathcal{S}_{i,j}};\, \boldsymbol{\theta}_N), \qquad \boldsymbol{x} \in \mathbb{X}, \tag{D.9}$$

where $\mathcal{S}_{i,j} \subset \mathcal{N} \setminus \{i, j\}$ $(i < j = 1, \ldots, N)$. Probability density functions of the form (D.9) can be represented as probability density functions of the form (D.8) by grouping the functions $\varphi_{i,j}$ in accordance with the maximal complete subsets of conditional independence graph $\mathcal{G}$. The conditional independence graph $\mathcal{G}$ depends on the model: e.g., the conditional independence graph of Model 1 has no edges, because all edge variables are independent. By contrast, the conditional independence graphs of Models 2–4 have edges, representing conditional dependencies induced by brokerage in networks. A graphical representation of the conditional independence graphs of Models 2–4 is shown in Figure 4—note that all three models have the same conditional independence graph.

To distinguish the random graph of interest (representing data structure) from the graph $\mathcal{G}$ (representing conditional independence structure, i.e., model structure), we call $\mathcal{G}(\mathcal{V}, \mathcal{E})$ the conditional independence graph, elements of $\mathcal{V}$ vertices rather than nodes, and elements of $\mathcal{E}$ edges rather than edge variables.

D.2.2. *Conditional independence properties.* We prove selected conditional independence properties that help establish consistency results and convergence rates for generalized β-models with dependent edges.

Generalized β-models with dependent edges constrain the dependence among edges to the intersections $\mathcal{N}_i \cap \mathcal{N}_j$ of neighborhoods $\mathcal{N}_i$ and $\mathcal{N}_j$ of nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$, and hence possess the following property:

Definition 1. Neighborhood intersection property. *Consider a random graph model with a probability density function parameterized by* (2.1) *and* (2.2)*. If $\mathcal{S}_{i,j} = \{i, j\} \times \{\mathcal{N}_i \cap \mathcal{N}_j\}$ for all pairs of nodes $\{i, j\} \subset \mathcal{N}$, then the random graph is said to satisfy the neighborhood intersection property.*

The neighborhood intersection property implies conditional independence properties, including—but not limited to—the following.

Proposition 1. *A random graph with overlapping subpopulations $\mathcal{A}_k$ of sizes $|\mathcal{A}_k| \geq 3$ ($k = 1, \ldots, K$) satisfying the neighborhood intersection property possesses the following conditional independence properties:*

1. *For all pairs of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$:*

$$X_{i,j} \perp\!\!\!\perp \boldsymbol{X} \setminus X_{i,j}.$$

2. *For all pairs of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$ and there exists $k \in \{1, \ldots, K\}$ such that $\{i, j\} \subset \mathcal{A}_k$ (i.e., i and j share one or more subpopulations):*

$$X_{i,j} \quad \perp\!\!\!\perp \quad \boldsymbol{X} \setminus \boldsymbol{X}_{\mathcal{N}_i \cup \mathcal{N}_j} \mid \boldsymbol{X}_{\mathcal{N}_i \cup \mathcal{N}_j} \setminus X_{i,j}.$$

3. *For all pairs of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$ and there does not exist $k \in \{1, \ldots, K\}$ such that $\{i, j\} \subset \mathcal{A}_k$ (i.e., i and j do not share any subpopulation, but belong to subpopulations that overlap):*

$$X_{i,j} \;\perp\!\!\!\perp\; \boldsymbol{X} \setminus (X_{i,j}, \boldsymbol{X}_{\mathcal{N}_i \cap \mathcal{N}_j}) \mid \boldsymbol{X}_{\mathcal{N}_i \cap \mathcal{N}_j}.$$

Proof of Proposition 1. In the following, we use the characterizations of conditional independence due to Dawid [23, 24] and others [e.g., 52], which relate factorization properties of probability density functions to conditional independence properties. Using these characterizations of conditional independence, we can establish the conditional independence properties stated in Proposition 1 by showing that, for each pair of nodes $\{i, j\} \subseteq \mathcal{N}$, there exists a subset of nodes $\mathfrak{S} \subset \mathcal{N} \setminus \{i, j\}$ and non-negative functions $g : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ and $h : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ such that the probability density function can be written as

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \quad \propto \quad g(x_{i,j},\, \boldsymbol{x}_{\mathfrak{S}})\; h(\boldsymbol{x} \setminus (x_{i,j},\, \boldsymbol{x}_{\mathfrak{S}}),\, \boldsymbol{x}_{\mathfrak{S}}),$$

which implies that

$$X_{i,j} \quad \perp\!\!\!\perp \quad \boldsymbol{X} \setminus (X_{i,j},\, \boldsymbol{X}_{\mathfrak{S}}) \mid \boldsymbol{X}_{\mathfrak{S}}.$$

Proposition 1 assumes that the neighborhood intersection property is satisfied, which implies that

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \quad \propto \quad \prod_{a\,<\,b}^{N} \varphi_{a,b}(x_{a,b},\, \boldsymbol{x}_{\{a,b\},\, \mathcal{N}_a \cap \mathcal{N}_b}).$$

Condition 1: Consider any pair of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$, that is, nodes i and j neither share subpopulations nor belong to distinct subpopulations that overlap. Since $\{i, j\} \times \{\mathcal{N}_i \cap \mathcal{N}_j\} = \emptyset$,

$$\varphi_{i,j}(x_{i,j},\, \boldsymbol{x}_{\{i,j\},\, \mathcal{N}_i \cap \mathcal{N}_j}) \quad = \quad \varphi_{i,j}(x_{i,j}).$$

It remains to check whether any $\varphi_{a,b}$ ($\{a, b\} \neq \{i, j\}$) can be a function of $x_{i,j}$, which would require that

$$(i, j) \in \{a, b\} \times \{\mathcal{N}_a \cap \mathcal{N}_b\} \quad \text{or} \quad (j, i) \in \{a, b\} \times \{\mathcal{N}_a \cap \mathcal{N}_b\}.$$

We prove by contradiction that $\varphi_{a,b}$ ($\{a, b\} \neq \{i, j\}$) cannot be a function of $x_{i,j}$. Consider any $a \in \mathcal{N} \setminus \{i, j\}$. Then $\varphi_{a,i}$ is a function of $x_{i,j}$ if $(i, j) \in \{a, i\} \times \{\mathcal{N}_a \cap \mathcal{N}_i\}$ or $(j, i) \in \{a, i\} \times \{\mathcal{N}_a \cap \mathcal{N}_i\}$, which would require that $j \in \mathcal{N}_a \cap \mathcal{N}_i$ and hence $j \in \mathcal{N}_i$. By definition of $\mathcal{N}_i$, $j \in \mathcal{N}_i$ is possible if and only if i and j share one or more subpopulations or belong to distinct subpopulations that overlap, in which case $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$, violating the assumption that $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$. Therefore, there cannot exist a pair of nodes $\{a, b\} \neq \{i, j\}$ such that $\varphi_{a,b}$ is a function of $x_{i,j}$. As a consequence, taking

$$g(x_{i,j}) \quad = \quad \varphi_{i,j}(x_{i,j})$$

and

$$h(\boldsymbol{x} \setminus x_{i,j}) \quad = \prod_{a<b:\ \{a,b\} \neq \{i,j\}}^{N} \varphi_{a,b}(x_{a,b},\, \boldsymbol{x}_{\{a,b\},\, \mathcal{N}_a \cap \mathcal{N}_b})$$

shows that $f_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ can be written as

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \quad \propto \quad g(x_{i,j})\, h(\boldsymbol{x} \setminus x_{i,j}),$$

which implies that

$$X_{i,j} \perp\!\!\!\perp \boldsymbol{X} \setminus X_{i,j}.$$

Condition 2: Consider any pair of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$ and there exists $k \in \{1, \ldots, K\}$ such that $\{i, j\} \subset \mathcal{A}_k$, that is, i and j share one or more subpopulations. By definition, $\varphi_{i,j}$ is a function of $x_{i,j}$. For any $\varphi_{a,b}$ ($\{a, b\} \neq \{i, j\}$) to be a function of $x_{i,j}$, we must have either

$$(i, j) \in \{a, b\} \times \{\mathcal{N}_a \cap \mathcal{N}_b\} \quad \text{or} \quad (j, i) \in \{a, b\} \times \{\mathcal{N}_a \cap \mathcal{N}_b\},$$

which requires that two conditions hold:

- $\{a, b\} \cap \{i, j\} \neq \emptyset$;
- either $i \in \mathcal{N}_a \cap \mathcal{N}_b$ or $j \in \mathcal{N}_a \cap \mathcal{N}_b$, which requires either that $i \in \mathcal{N}_a$ and $i \in \mathcal{N}_b$ or that $j \in \mathcal{N}_a$ and $j \in \mathcal{N}_b$.

Let $c = \{a, b\} \setminus \{i, j\}$ and $d = \{a, b\} \cap \{i, j\}$ and assume, without loss, that $c < d$. Then the subset of pairs of nodes $\{c, d\}$ such that $\varphi_{c,d}$ is a function of $x_{i,j}$ is the subset of all nodes c such that either $c \in \mathcal{N}_i$ or $c \in \mathcal{N}_j$ or both, so c must satisfy $c \in \mathcal{N}_i \cup \mathcal{N}_j \setminus \{i, j\}$. Therefore, there exist non-negative functions $g : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ and $h : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ such that the probability density function can be written as follows:

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \quad \propto \quad g(\boldsymbol{x}_{\{i,j\},\, \mathcal{N}_i \cup \mathcal{N}_j})\, h(\boldsymbol{x} \setminus \boldsymbol{x}_{\{i,j\},\, \mathcal{N}_i \cup \mathcal{N}_j}),$$

which implies that

$$X_{i,j} \quad \perp\!\!\!\perp \quad \boldsymbol{X} \setminus \boldsymbol{X}_{\mathcal{N}_i \cup \mathcal{N}_j} \mid \boldsymbol{X}_{\mathcal{N}_i \cup \mathcal{N}_j} \setminus X_{i,j}.$$

Condition 3: Consider any pair of nodes $\{i, j\} \subset \mathcal{N}$ such that $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$ and there does not exist $k \in \{1, \ldots, K\}$ such that $\{i, j\} \subset \mathcal{A}_k$, that is, i and j do not share any subpopulation, but belong to subpopulations that overlap. Since i and j do not share any subpopulation, $\varphi_{a,b}$ can be a function of $x_{i,j}$ only if $\{a, b\} \cap \{i, j\} \neq \emptyset$ and $c = \{a, b\} \setminus \{i, j\}$ is contained in $\mathcal{N}_i \cap \mathcal{N}_j$.

As a result, there exist non-negative functions $g : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ and $h : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ such that the probability density function can be written as follows:

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \;\;\propto\;\; g(x_{i,j},\, \boldsymbol{x}_{\{i,j\},\, \mathcal{N}_i \cap \mathcal{N}_j})\, h(\boldsymbol{x} \setminus (x_{i,j},\, \boldsymbol{x}_{\{i,j\},\, \mathcal{N}_i \cap \mathcal{N}_j})),$$

which implies that

$$X_{i,j} \;\;\perp\!\!\!\perp\;\; \boldsymbol{X} \setminus (X_{i,j},\, \boldsymbol{X}_{\{i,j\},\, \mathcal{N}_i \cap \mathcal{N}_j}) \mid \boldsymbol{X}_{\{i,j\},\, \mathcal{N}_i \cap \mathcal{N}_j}.$$

D.2.3. *Bounding the spectral norm of the coupling matrix.* We bound the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$. Throughout, we adopt the notation used in Section 3 of the manuscript, that is, we denote the number of edge variables by $M = \binom{N}{2}$ and edge variables by $X_1, \ldots, X_M$, where N is the number of nodes. In a mild abuse of notation, we denote the vertices $X_1, \ldots, X_M$ of the conditional independence graph $\mathcal{G}$ by $1, \ldots, M$. The degree of a vertex is the number of other vertices connected to it.

By Hölder's inequality, the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$ is bounded above by

$$|||\mathcal{D}|||_2 \;\;\leq\;\; \sqrt{|||\mathcal{D}|||_1\; |||\mathcal{D}|||_\infty},$$

where

$$|||\mathcal{D}|||_1 \;\;=\;\; \max_{1 \leq j \leq M} \sum_{i=1}^{M} |\mathcal{D}_{i,j}|$$

is the maximum absolute column sum of $\mathcal{D}$, and

$$|||\mathcal{D}|||_\infty \;\;=\;\; \max_{1 \leq i \leq M} \sum_{j=1}^{M} |\mathcal{D}_{i,j}|$$

is the maximum absolute row sum of $\mathcal{D}$.

The following lemma bounds $|||\mathcal{D}|||_2$ by bounding $|||\mathcal{D}|||_1$ and $|||\mathcal{D}|||_\infty$. Auxiliary results are proved in Appendix D.2.4.

Lemma 10. *Consider a random graph with the neighborhood intersection property stated in Definition 1, with data-generating parameter vector* $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ *satisfying* (D.1) *with* $\vartheta = 1$. *Assume that* $\min_{1 \leq k \leq K} |\mathcal{A}_k| \geq 3$, $\max_{1 \leq i \leq N} |\mathcal{N}_i| \leq D$ $(D \geq 2)$, *and Assumption B is satisfied. Then there exists a universal constant* $1 \leq C < \infty$ *such that* $|||\mathcal{D}|||_2 \leq C$.

Proof of Lemma 10. We adapt the coupling approach of van den Berg and Maes [73] from Gibbs measures and Markov random fields to coupling conditional distributions of subgraphs of random graphs. Let $i \in \mathcal{V}$ be any vertex of the conditional independence graph $\mathcal{G}$ and consider any $\boldsymbol{x}_{1:i-1} \in \{0,1\}^{i-1}$. Define

$$\mathbb{P}_{\boldsymbol{x}_{1:i-1},0}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a}) \overset{\text{def}}{=} \mathbb{P}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1}, X_i = 0)$$

and

$$\mathbb{P}_{\boldsymbol{x}_{1:i-1},1}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a}) \overset{\text{def}}{=} \mathbb{P}(\boldsymbol{X}_{i+1:M} = \boldsymbol{a} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1}, X_i = 1),$$

where $\boldsymbol{X}_{1:i-1} = (X_1, \ldots, X_{i-1})$, $\boldsymbol{X}_{i+1:M} = (X_{i+1}, \ldots, X_M)$, and $\boldsymbol{a} \in \{0,1\}^{M-i}$. We divide the proof into three parts:

I. Coupling conditional distributions of subgraphs.

II. Bounding the elements of the coupling matrix $\mathcal{D}$.

III. Bounding the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$.

I. Coupling conditional distributions of subgraphs. Given any vertex $i \in \mathcal{V}$ of the conditional independence graph $\mathcal{G}$ and any $\boldsymbol{x}_{1:i-1} \in \{0,1\}^{i-1}$, we construct a coupling $(\boldsymbol{X}^\star, \boldsymbol{X}^{\star\star}) \in \{0,1\}^{M-i} \times \{0,1\}^{M-i}$ of the conditional distributions $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},0}$ and $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},1}$. Some background on coupling can be found in Section 3.2 of the manuscript.

To simplify the notation, we assume that the couple $(\boldsymbol{X}^\star, \boldsymbol{X}^{\star\star})$ takes on values in $\{0,1\}^M \times \{0,1\}^M$ rather than $\{0,1\}^{M-i} \times \{0,1\}^{M-i}$ and set $(\boldsymbol{X}^\star_{1:i-1}, X^\star_i) = (\boldsymbol{x}_{1:i-1}, 0)$ and $(\boldsymbol{X}^{\star\star}_{1:i-1}, X^{\star\star}_i) = (\boldsymbol{x}_{1:i-1}, 1)$. As a consequence, the random vectors $\boldsymbol{X}^\star \in \{0,1\}^M$ and $\boldsymbol{X}^{\star\star} \in \{0,1\}^M$ have the same dimension as random vector $\boldsymbol{X} \in \{0,1\}^M$. We then construct a coupling of the conditional distributions $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},0}$ and $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},1}$ as follows:

1. Initialize $\mathfrak{V} = \{1, \ldots, i\}$.

2. Check whether there exists a vertex $j \in \mathcal{V} \setminus \mathfrak{V}$ that is connected to a vertex $v \in \mathfrak{V}$ in $\mathcal{G}$ and satisfies $X^\star_v \neq X^{\star\star}_v$:

   (a) If such a vertex j exists, pick the smallest such vertex, and let $(X^\star_j, X^{\star\star}_j)$ be distributed according to an optimal coupling of $\mathbb{P}(X_j = \cdot \mid \boldsymbol{X}_{\mathfrak{V}} = \boldsymbol{x}^\star_{\mathfrak{V}})$ and $\mathbb{P}(X_j = \cdot \mid \boldsymbol{X}_{\mathfrak{V}} = \boldsymbol{x}^{\star\star}_{\mathfrak{V}})$.

   (b) If no such vertex j exists, select the smallest $j \in \mathcal{V} \setminus \mathfrak{V}$ and let $(X^\star_j, X^{\star\star}_j)$ be distributed according to an optimal coupling of the distributions $\mathbb{P}(X_j = \cdot \mid \boldsymbol{X}_{\mathfrak{V}} = \boldsymbol{x}_{\mathfrak{V}}^{\star})$ and $\mathbb{P}(X_j = \cdot \mid \boldsymbol{X}_{\mathfrak{V}} = \boldsymbol{x}_{\mathfrak{V}}^{\star\star})$. The optimal coupling ensures that $X_j^{\star} = X_j^{\star\star}$ with probability 1, so the total variation distance is 0.

In both steps, an optimal coupling exists [56, Theorem 5.2, p. 19], but may not be unique. However, any optimal coupling will do.

3. Replace $\mathfrak{V}$ by $\mathfrak{V} \cup \{j\}$ and repeat Step 2 until $\mathcal{V} \setminus \mathfrak{V} = \emptyset$.

Denote the resulting coupling distribution by $\mathbb{Q}_{i,\boldsymbol{x}_{1:i-1}}^{\text{opt}}$. Lemma 12 verifies that the algorithm above constructs a valid coupling of the conditional distributions $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},0}$ and $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},1}$, in the sense that the marginal distributions of $\boldsymbol{X}^{\star}$ and $\boldsymbol{X}^{\star\star}$ are $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},1}$ and $\mathbb{P}_{i,\boldsymbol{x}_{1:i-1},0}$, respectively.

By construction, the coupling has useful properties. For any two distinct vertices $i \in \mathcal{V}$ and $j \in \{i+1, \ldots, M\}$ of the conditional independence graph $\mathcal{G}$, define the event $i \not\leftrightarrow j$ to be the event that there exists a path from i to j in $\mathcal{G}$ such that $X_v^{\star} \neq X_v^{\star\star}$ for all vertices v along the path. Such paths are known as *paths of disagreement* in the probability literature on Gibbs measures and Markov random fields. Theorem 1 of van den Berg and Maes [73, p. 753] shows that

$$\mathbb{Q}_{i,\boldsymbol{x}_{1:i-1}}^{\text{opt}}(X_j^{\star} \neq X_j^{\star\star}) \;=\; \mathbb{Q}_{i,\boldsymbol{x}_{1:i-1}}^{\text{opt}}(i \not\leftrightarrow j) \;\leq\; \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j), \tag{D.10}$$

where $\mathbb{B}_{\boldsymbol{\pi}}$ is a Bernoulli product measure on $\{0,1\}^M$ with probability vector $\boldsymbol{\pi} \in [0,1]^M$. The coordinates π_v of $\boldsymbol{\pi}$ are given by

$$\pi_v \;=\; \begin{cases} 0 & \text{if } v \in \{1, \ldots, i-1\} \\ 1 & \text{if } v = i \\ \max\limits_{(\boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}) \,\in\, \{0,1\}^{M-1} \times \{0,1\}^{M-1}} \pi_{v,\, \boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}} & \text{if } v \in \{i+1, \ldots, M\}, \end{cases}$$

where

$$\pi_{v,\, \boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}} \;=\; \|\mathbb{P}(\,\cdot \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v}) - \mathbb{P}(\,\cdot \mid \boldsymbol{X}'_{-v} = \boldsymbol{x}'_{-v})\|_{\text{TV}}.$$

The Bernoulli product measure $\mathbb{B}_{\boldsymbol{\pi}}$ assumes that independent Bernoulli experiments are carried out at vertices $v \in \{1, \ldots, M\}$. The Bernoulli experiment at vertex $v \in \{i+1, \ldots, M\}$ has two possible outcomes: Either vertex v is *open,* corresponding to the event that $X_v^{\star} \neq X_v^{\star\star}$ and hence vertex v allows a path of disagreement from i to j to pass through, or vertex v is *closed.* A vertex v is open with probability π_v, and closed with probability $1-\pi_v$. By construction, vertices $v \in \{1, \ldots, i-1\}$ are closed with probability one, and vertex i is open with probability one.

The coupling argument of van den Berg and Maes [73] is useful, in that it translates the hard problem of bounding probabilities of events involving dependent random variables into the more convenient problem of bounding probabilities of events involving independent random variables. Indeed, we can bound the above-diagonal elements $\mathcal{D}_{i,j}$ of the coupling matrix $\mathcal{D}$ by

$$\mathcal{D}_{i,j} \;=\; \sup_{\boldsymbol{x}_{1:i-1} \in \{0,1\}^{i-1}} \mathbb{Q}^{\text{opt}}_{i,\boldsymbol{x}_{1:i-1}}(X_j^\star \neq X_j^{\star\star}) \;\leq\; \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j). \tag{D.11}$$

By construction of $\mathcal{D}$, the below-diagonal and diagonal elements of $\mathcal{D}$ are known to be 0 and 1, respectively. We define $\pi^\star \in (0,1)$ by

$$\pi^\star \;=\; \max_{1 \leq v \leq M} \; \max_{\boldsymbol{x}_{-v} \in \{0,1\}^{M-1}} \mathbb{P}(X_v = 1 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v}),$$

and note that Lemma 13 along with the assumption that $\boldsymbol{\theta}_N^\star \in \boldsymbol{\Theta}_N \subseteq \mathbb{R}^p$ satisfies (D.1) with $\vartheta = 1$ implies that

$$\pi^\star \;\leq\; \frac{1}{1 + \exp(-(3 + 2\,D)\,U)} \;<\; 1,$$

where $D \geq 2$ and $U > 0$ are constants. In other words, there exists a constant $\zeta \in (0,1)$, given by

$$\zeta \;=\; \frac{\exp(-(3 + 2\,D)\,U)}{1 + \exp(-(3 + 2\,D)\,U)} \;\in\; (0,1),$$

such that

$$\pi^\star \;<\; 1 - \zeta \;\in\; (0,1). \tag{D.12}$$

II. Bounding the elements of the coupling matrix $\mathcal{D}$. To bound the elements $\mathcal{D}_{i,j}$ of the coupling matrix $\mathcal{D}$, we bound the probability on the right-hand side of (D.11). We do so by constructing, for each pair of vertices $(i,j) \in \mathcal{V} \times \mathcal{V}$ $(i < j)$ of the conditional independence graph, a graphical cover $\mathcal{G}_{i,j}^\star$ of the conditional independence graph $\mathcal{G}$ as follows:

1. Initialize $\mathcal{G}_{i,j}^\star$ by $\mathcal{G}$, that is, $\mathcal{G}_{i,j}^\star$ is a graph with the same set of vertices and the same set of edges as $\mathcal{G}$.
2. For each pair of subpopulations $\mathcal{A}_v$ and $\mathcal{A}_w$ such that $\mathcal{A}_v \cap \mathcal{A}_w \neq \emptyset$, add edges in $\mathcal{G}_{i,j}^\star$ between each pair of edge variables in subgraph $\boldsymbol{X}_{\mathcal{A}_v \cup \mathcal{A}_w}$ that are not connected in $\mathcal{G}$.

3. If $d_{\mathcal{G}_{i,j}^\star}(i,j) = l$ $(1 \leq l < \infty)$, let $\mathcal{W} \subseteq \{1, \ldots, M\} \setminus \{i, j\}$ be the subset of vertices for which $d_{\mathcal{G}_{i,j}^\star}(i,w) = l - 1$ $(w \in \mathcal{W})$. We add edges in $\mathcal{G}_{i,j}^\star$ between all unconnected pairs of vertices X_j and X_w $(w \in \mathcal{W})$. As a result, any path of disagreement $i \not\leftrightarrow j$ must pass through the vertices in $\mathcal{W}$.

Since $\mathcal{G}_{i,j}^\star$ is a graphical cover of $\mathcal{G}$, any path of disagreement $i \not\leftrightarrow j$ in $\mathcal{G}$ is a path of disagreement in $\mathcal{G}_{i,j}^\star$, which implies that

$$\mathcal{D}_{i,j} \;\; \leq \;\; \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}) \;\; \leq \;\; \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}_{i,j}^\star).$$

We bound $\mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}_{i,j}^\star)$ under Assumptions B.1 and B.2. To do so, define

$$\mathcal{V}_{i,k} \;\; = \;\; \left\{ v \in \mathcal{V} \setminus \{i\} \, : \, d_{\mathcal{G}_{i,j}^\star}(i,v) = k \right\}, \qquad k = 1, \ldots, M-1.$$

The set $\mathcal{V}_{i,k} \subseteq \mathcal{V}$ is the subset of vertices in $\mathcal{G}_{i,j}^\star$ with graph distance k from i in $\mathcal{G}_{i,j}^\star$. Let $g : \{1, 2, \ldots\} \mapsto \mathbb{R}^+$ be such that, for all subpopulations $\mathcal{A}_r$,

$$|\{\mathcal{A}_w \in \{\mathcal{A}_1, \ldots, \mathcal{A}_K\} \, : \, d_{\mathcal{G}_{\mathcal{A}}}(\mathcal{A}_r, \mathcal{A}_w) = k\}| \;\; \leq \;\; g(k), \qquad k \in \{1, 2, \ldots\}.$$

The graphical cover $\mathcal{G}_{i,j}^\star$ of $\mathcal{G}$ ensures the following—note that $\mathcal{G}_{\mathcal{A}}$ is the subpopulation graph with set of vertices $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ and edges between vertices $\mathcal{A}_r$ and $\mathcal{A}_l$ if and only if $\mathcal{A}_r \cap \mathcal{A}_l \neq \emptyset$:

- Consider any $j \in \mathcal{V}_{i,1}$ such that $d_{\mathcal{G}_{i,j}^\star}(i,j) = 1$. By the construction of $\mathcal{G}_{i,j}^\star$, we can associate X_i and X_j with a subgraph $\boldsymbol{X}_{\mathcal{A}_r \cup \mathcal{A}_l}$ for some $r \neq l \in \{1, \ldots, K\}$ satisfying $\mathcal{A}_r \cap \mathcal{A}_l \neq \emptyset$. The number $|\mathcal{V}_{i,1}|$ of such vertices $j \in \mathcal{V}_{i,1}$ is bounded above by

$$|\mathcal{V}_{i,1}| \;\; \leq \;\; 2\, g(1) \binom{2\,D}{2} \;\; \leq \;\; 8\, D^2\, g(1),$$

  because the number of subgraphs that overlap with either $\mathcal{A}_r$ or $\mathcal{A}_l$ is bounded above by $g(1)$, hence the total number of overlapping subpopulations is bounded above by $2\, g(1)$; and because the number of edge variables in each such subgraph $\boldsymbol{X}_{\mathcal{A}_r \cup \mathcal{A}_l}$ is bounded above by $\binom{2\,D}{2} \leq 4\, D^2$, using $|\mathcal{A}_r| \leq D + 1 \leq 2\, D$ $(r \in \{1, \ldots, K\})$.

- Consider any $j \in \mathcal{V}_{i,2}$ such that $d_{\mathcal{G}_{i,j}^\star}(i,j) = 2$. Then by the construction of $\mathcal{G}_{i,j}^\star$, there exist distinct $\{r, t, l\} \subseteq \{1, \ldots, K\}$ such that $i \in \mathcal{A}_r \cup \mathcal{A}_t$ and $j \in \mathcal{A}_l \cup \mathcal{A}_t$, where $\mathcal{A}_t$ and $\mathcal{A}_r$ overlap, $\mathcal{A}_t$ and $\mathcal{A}_l$

overlap, but $\mathcal{A}_r$ and $\mathcal{A}_l$ do not overlap. In other words, $d_{\mathcal{G}_\mathcal{A}}(r, l) = 2$. As a result, the number $|\mathcal{V}_{i,2}|$ of such vertices $j \in \mathcal{V}_{i,2}$ is bounded by

$$|\mathcal{V}_{i,2}| \quad \leq \quad 2\, g(2) \binom{2\,D}{2} \quad \leq \quad 8\, D^2\, g(2),$$

because the number of edge variables in each subgraph $\boldsymbol{X}_{\mathcal{A}_l \cup \mathcal{A}_t}$ is bounded above by $\binom{2\,D}{2} \leq 4\, D^2$, as discussed above, and the number of subpopulations $\mathcal{A}_l$ at distance 2 of a given subpopulation $\mathcal{A}_r$ is bounded above by $g(2)$.

- Consider any $j \in \mathcal{V}_{i,k}$ $(k \geq 3)$ such that $d_{\mathcal{G}^\star_{i,j}}(i, j) = k$. If $k = 3$, then there exists a vertex $v \in \mathcal{V}_{i,2} \setminus \{i, j\}$ such that $d_{\mathcal{G}^\star_{i,j}}(i, v) = 2$ and $d_{\mathcal{G}^\star_{i,j}}(v, j) = 1$. Applying the first argument above, there exists $r \neq t \in \{1, \dots, K\}$ such that $d_{\mathcal{G}_\mathcal{A}}(r, t) = 1$, and X_v can be associated with subgraph $\boldsymbol{X}_{\mathcal{A}_r \cup \mathcal{A}_t}$ and X_j can be associated with subgraph $\boldsymbol{X}_{\mathcal{A}_r \cup \mathcal{A}_t}$. The number of edge variables in subgraphs $\boldsymbol{X}_{\mathcal{A}_r \cup \mathcal{A}_t}$ is bounded above by $\binom{2\,D}{2}$. Since $d_{\mathcal{G}^\star_{i,j}}(i, v) = 2$, we can conclude that there exist distinct $\{s, l, r\} \subseteq \{1, \dots, K\}$ such that $i \in \mathcal{A}_s \cup \mathcal{A}_l$ and $j \in \mathcal{A}_l \cup \mathcal{A}_r$, where $\mathcal{A}_l$ and $\mathcal{A}_s$ overlap, $\mathcal{A}_l$ and $\mathcal{A}_r$ overlap, but $\mathcal{A}_s$ and $\mathcal{A}_r$ do not overlap. This means that $d_{\mathcal{G}_\mathcal{A}}(s, r) = 2$ and $d_{\mathcal{G}_\mathcal{A}}(s, t) = 3$. Hence, the number of $\mathcal{A}_t$ satisfying $d_{\mathcal{G}_\mathcal{A}}(s, t) = 3$ is bounded above by $g(3)$, and

$$|\mathcal{V}_{i,3}| \quad \leq \quad 2 \binom{2\,D}{2} g(3) \quad \leq \quad 8\, D^2\, g(3).$$

The case $k \geq 4$ can be proved by using induction, showing that

$$|\mathcal{V}_{i,k}| \quad \leq \quad 8\, D^2\, g(k).$$

We proceed with bounding $\mathcal{D}_{i,j}$ under Assumptions B.1 and B.2.

Bounding $\mathcal{D}_{i,j}$ under Assumption B.1. Define the function $g : \{1, 2, \dots\} \mapsto \mathbb{R}^+$ by

$$g(l) \quad = \quad \omega_1 + \frac{\omega_2}{8\, D^2} \log l, \qquad l \in \{1, 2, \dots\},$$

where $\omega_1 > 0$ and $\omega_2 \geq 0$ satisfy

$$\omega_2 \quad < \quad \frac{1}{|\log(1 - \pi^\star)|},$$

and $\pi^\star < 1 - \zeta \in (0, 1)$ by (D.12). Lemma 13 shows that

$$\pi_v \quad \leq \quad \pi^\star, \qquad v \in \{i + 1, \dots, M\},$$

which implies that

$$\mathbb{B}_{\boldsymbol{\pi}}(v \text{ is open}) \;\; \leq \;\; \pi^\star \;\; < \;\; 1, \qquad v \in \{i+1, \dots, M\}.$$

Assumption B.1 assumes that, for all subpopulations $\mathcal{A}_k$,

$$|\{\mathcal{A}_w \in \{\mathcal{A}_1, \dots, \mathcal{A}_K\} \,:\, d_{\mathcal{G}_{\mathcal{A}}}(\mathcal{A}_k, \mathcal{A}_w) = l\}| \;\; \leq \;\; g(l), \qquad l \in \{1, 2, \dots\},$$

implying that, for all $l \in \{1, 2, \dots\}$,

$$|\mathcal{V}_{i,l}| \;\; \leq \;\; 8\, D^2\, \omega_1 + \omega_2 \log l,$$

using the above definition of $g(l)$. We absorb the constant $8\, D^2$ into ω_1, noting that we are free to choose $\omega_1 > 0$. For there to be a path of disagreement $i \not\leftrightarrow j$ in $\mathcal{G}^\star_{i,j}$, there must be at least one open vertex in each of the sets $\mathcal{V}_{i,1}, \dots, \mathcal{V}_{i,k-1}$ and j must be open; note that i is open with probability 1. The probability that there exists at least one open vertex $v \in \mathcal{V}_{i,l}$ is bounded above by

$$1 - (1 - \pi^\star)^{|\mathcal{V}_{i,l}|} \;\; \leq \;\; 1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, l}, \qquad l \in \{1, 2, \dots\}.$$

Since the events that vertices are open are independent under the Bernoulli product measure $\mathbb{B}_{\boldsymbol{\pi}}$, we obtain

$$\begin{aligned} \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}^\star_{i,j}) \;\; &\leq \;\; \pi^\star \prod_{l=1}^{k-1} \left[1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, l}\right] \\ &\leq \;\; \left[1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, k}\right]^k, \end{aligned}$$

using $\log l \leq \log k$ provided $l \leq k$, along with the fact that $\pi^\star \leq 1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, k}$. We then write

$$\begin{aligned} 1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, k} \;\; &= \;\; \exp\left(\log(1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, k})\right) \\ &\leq \;\; \exp(-(1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, k}) \\ &= \;\; \exp(-\exp((\omega_1 + \omega_2 \,\log(k))\, \log(1 - \pi^\star))) \\ &= \;\; \exp(-\exp(-(\omega_1 + \omega_2 \,\log(k))\, |\log(1 - \pi^\star)|)) \\ &= \;\; \exp(-\exp(-\omega_1\, |\log(1 - \pi^\star)|)\, k^{-\omega_2\, |\log(1 - \pi^\star)|}), \end{aligned}$$

using the inequality $1 - z = \exp(\log(1 - z)) \leq \exp(-z)$ provided $z \in (0, 1)$. Let $A = \exp(-\omega_1 \,|\log(1 - \pi^\star)|) \in (0, 1)$. Then

$$\begin{aligned}\left[1 - (1 - \pi^\star)^{\omega_1 + \omega_2 \,\log\, k}\right]^k &\leq \left[\exp(-A\, k^{-\omega_2 \,|\log(1-\pi^\star)|})\right]^k \\ &= \exp(-A\, k^{1-\omega_2 \,|\log(1-\pi^\star)|}).\end{aligned}$$

The probability of the event that vertex j is open is bounded above by $\pi^\star \in (0, 1)$, so

$$\mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}_{i,j}^\star) \quad \leq \quad \pi^\star \exp(-A\, k^{1-\omega_2 \,|\log(1-\pi^\star)|}).$$

We hence obtain

$$\mathcal{D}_{i,j} \quad \leq \quad \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}_{i,j}^\star) \quad \leq \quad \exp(-A\; k^{1-\omega_2 \,|\log(1-\pi^\star)|}). \tag{D.13}$$

Bounding $\mathcal{D}_{i,j}$ under Assumption B.2. By Assumption B.2, the subpopulation graph $\mathcal{G}_{\mathcal{A}}$ is a tree. In other words, the subpopulation graph $\mathcal{G}_{\mathcal{A}}$ is a connected graph, there exists a path between any two subpopulations in $\mathcal{G}_{\mathcal{A}}$, and the path is unique. The fact that there is a unique path between any two subpopulation simplifies bounds on the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$. To demonstrate, choose any two distinct vertices $i \in \mathcal{V}$ and $j \in \mathcal{V}$ of the conditional independence graph $\mathcal{G}$.

- If $d_{\mathcal{G}^\star}(i, j) = 1$, then

  $$\mathcal{D}_{i,j} \quad \leq \quad \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}^\star) \quad \leq \quad \pi^\star,$$

  because i is open with probability 1 and j is open with at most probability $\pi^\star$.

- If $d_{\mathcal{G}^\star}(i, j) = 2$, then by the construction of $\mathcal{G}^\star$, there exist three subpopulations $\mathcal{A}_v$, $\mathcal{A}_w$, and $\mathcal{A}_z$ such that i corresponds to an edge variable in $\boldsymbol{X}_{\mathcal{A}_v \cup \mathcal{A}_w}$ and j corresponds to an edge variable in $\boldsymbol{X}_{\mathcal{A}_w \cup \mathcal{A}_z}$. Since $d_{\mathcal{G}^\star}(i, j) = 2$, i must be in the subgraph $\boldsymbol{X}_{\mathcal{A}_v \cup \mathcal{A}_w} \setminus \boldsymbol{X}_{\mathcal{A}_w \cup \mathcal{A}_z}$, and j must be in the subgraph $\boldsymbol{X}_{\mathcal{A}_w \cup \mathcal{A}_z} \setminus \boldsymbol{X}_{\mathcal{A}_v \cup \mathcal{A}_w}$. Thus, the subgraph $\boldsymbol{X}_{\mathcal{A}_w}$ of edge variables separates i and j in $\mathcal{G}^\star$, implying

  $$\mathcal{D}_{i,j} \quad \leq \quad \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}^\star) \quad \leq \quad \pi^\star\,(1 - (1 - \pi^\star)^{2\,D^2}),$$

  because each subpopulation has no more than $\max_{1 \leq k \leq K} |\mathcal{A}_k| \leq 1 + D$ nodes and hence each subgraph $\boldsymbol{X}_{\mathcal{A}_w}$ contains no more than

  $$\binom{D+1}{2} \quad \leq \quad \frac{(D+1)^2}{2} \quad \leq \quad \frac{1}{2}\left(\frac{3\,D}{2}\right)^2 \quad \leq \quad 2\,D^2$$

possible edges, using the inequality $D + 1 \leq 3\,D\,/\,2$ $(D \geq 2)$. Here, we have taken advantage of the fact that each vertex in the conditional independence graph $\mathcal{G}^\star$ is open with at most probability $\pi^\star$, and the events that vertices are open are independent under the Bernoulli product measure $\mathbb{B}_{\boldsymbol{\pi}}$.

- If $d_{\mathcal{G}^\star}(i,j) = l \geq 3$, then we apply the same argument as in the case when $d_{\mathcal{G}^\star}(i,j) = 2$ iteratively to conclude that there exist $l - 1 \geq 2$ subgraphs $\boldsymbol{X}_{\mathcal{A}_v}$ of edge variables separating i and j in $\mathcal{G}^\star$, implying

  $$\mathcal{D}_{i,j} \;\;\leq\;\; \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}^\star) \;\;\leq\;\; \pi^\star\,(1 - (1 - \pi^\star)^{2\,D^2})^{l-1},$$

  using the same argument as in the preceding case.

To conclude, we obtain the following bound on $\mathcal{D}_{i,j}$:

$$\begin{aligned}\mathcal{D}_{i,j} \;\;\leq\;\; \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j \text{ in } \mathcal{G}^\star) \;\;&\leq\;\; \pi^\star\,(1 - (1 - \pi^\star)^{2\,D^2})^{d_{\mathcal{G}^\star}(i,j)-1}\\ &\leq\;\; (1 - (1 - \pi^\star)^{2\,D^2})^{d_{\mathcal{G}^\star}(i,j)},\end{aligned} \tag{D.14}$$

using $\pi^\star \leq 1 - (1 - \pi^\star)^{2\,D^2}$.

III. Bounding the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$. To bound the spectral norm $|||\mathcal{D}|||_2$ of the coupling matrix $\mathcal{D}$, we first use Hölder's inequality to obtain

$$|||\mathcal{D}|||_2 \;\;\leq\;\; \sqrt{|||\mathcal{D}|||_1\;|||\mathcal{D}|||_\infty},$$

and then bound the elements $\mathcal{D}_{i,j}$ of $\mathcal{D}$. To facilitate bounds on $|||\mathcal{D}|||_2$, we form a symmetric $M \times M$ matrix $\mathcal{T}$ by defining

$$\mathcal{T} \;\;=\;\; \mathcal{D} + \mathcal{D}^\top - \operatorname{diag}(\mathcal{D}),$$

where $\mathcal{D}^\top$ is the $M \times M$ transpose of $\mathcal{D}$ and $\operatorname{diag}(\mathcal{D})$ is the $M \times M$ diagonal matrix with elements $\mathcal{D}_{1,1}, \dots, \mathcal{D}_{M,M}$ on the main diagonal. By construction of $\mathcal{T}$, the elements $\mathcal{T}_{i,j}$ of $\mathcal{T}$ are given by

$$\mathcal{T}_{i,j} \;\;=\;\; \begin{cases} \mathcal{D}_{i,j} & \text{if } i < j \\ \mathcal{D}_{i,i} & \text{if } i = j \\ \mathcal{D}_{j,i} & \text{if } j < i \end{cases},$$

where $\mathcal{D}_{i,i} = 1$ by definition of $\mathcal{D}_{i,i}$ $(i = 1, \dots, M)$. Using the fact that $\mathcal{T}_{i,j} = \max(\mathcal{D}_{i,j},\, \mathcal{D}_{j,i})$ $(i, j = 1, \dots, M)$, we obtain

$$|||\mathcal{D}|||_1 \;=\; \max_{1 \leq j \leq M} \sum_{i=1}^{M} |\mathcal{D}_{i,j}| \;\leq\; \max_{1 \leq j \leq M} \sum_{i=1}^{M} |\mathcal{T}_{i,j}| \;=\; |||\mathcal{T}|||_1$$

and

$$|||\mathcal{D}|||_\infty \;=\; \max_{1 \leq i \leq M} \sum_{j=1}^{M} |\mathcal{D}_{i,j}| \;\leq\; \max_{1 \leq i \leq M} \sum_{j=1}^{M} |\mathcal{T}_{i,j}| \;=\; |||\mathcal{T}|||_\infty.$$

In addition, we know that $\mathcal{T}_{i,j} = \mathcal{T}_{j,i}$ $(i, j = 1, \dots, M)$, which implies that

$$|||\mathcal{T}|||_1 \;=\; |||\mathcal{T}^\top|||_\infty \;=\; |||\mathcal{T}|||_\infty.$$

As a consequence, we obtain

$$|||\mathcal{D}|||_2 \;\leq\; \sqrt{|||\mathcal{D}|||_1\, |||\mathcal{D}|||_\infty} \;\leq\; \sqrt{|||\mathcal{T}|||_1\, |||\mathcal{T}|||_\infty} \;=\; |||\mathcal{T}|||_\infty,$$

where $|||\mathcal{T}|||_\infty$ can be bounded above by using (D.11):

$$|||\mathcal{T}|||_\infty \;=\; \max_{1 \leq i \leq M} \sum_{j=1}^{M} |\mathcal{T}_{i,j}| \;\leq\; 1 + \max_{1 \leq i \leq M} \sum_{j=1:\, j \neq i}^{M} \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j).$$

We apply the bounds on $\mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j)$ derived in (D.13) and (D.14) under Assumptions B.1 and B.2, respectively.

Bounding $|||\mathcal{D}|||_2$ *under Assumption B.1:* Using (D.13),

$$\begin{aligned} |||\mathcal{D}|||_2 \;&\leq\; 1 + \max_{1 \leq i \leq M} \sum_{j=1:\, j \neq i}^{M} \mathbb{B}_{\boldsymbol{\pi}}(i \not\leftrightarrow j) \\ &\leq\; 1 + \sum_{k=1}^{\infty} 8\, D^2\, g(k)\, \exp(-A\, k^{1-\omega_2\, |\log(1-\pi^\star)|}), \end{aligned}$$

noting that the number of vertices $j \in \mathcal{V}$ at distance k in $\mathcal{G}_{i,j}^\star$ to any given vertex i is bounded above by $8\, D^2\, g(k) = 8\, D^2\, \omega_1 + \omega_2\, \log(k)$, derived above. Without loss, as done above previously, we absorb $8\, D^2$ into ω_1 because we are free to choose $\omega_1 > 0$, and we considering bounding the infinite series

$$\sum_{k=1}^{\infty} (\omega_1 + \omega_2\, \log(k))\, \exp(-A\, k^{1-\omega_2\, |\log(1-\pi^\star)|}).$$

The inequality $\exp(z) \geq z^u \,/\, u!$ holds for any $z > 0$ and any $u \geq 1$, which implies that $\exp(-z) \leq u! \,/\, z^u$. So

$$\begin{aligned}
&\sum_{k=1}^{\infty} (\omega_1 + \omega_2\, \log(k))\, \exp\left(-A\, k^{1-\omega_2\,|\log(1-\pi^\star)|}\right) \\
\leq\;& \sum_{k=1}^{\infty} (\omega_1 + \omega_2\, \log(k))\, \frac{u!}{A^u\, k^{u\,(1-\omega_2\,|\log(1-\pi^\star)|)}} \\
\leq\;& \sum_{k=1}^{\infty} (\omega_1 + \omega_2)\, k\, \frac{u!}{A^u\, k^{u\,(1-\omega_2\,|\log(1-\pi^\star)|)}},
\end{aligned}$$

where the last inequality follows from the assumption $\omega_1 > 0$ and the fact that $\log k \leq k$ provided $k \geq 1$. By assumption, $\omega_2 \geq 0$ satisfies

$$\omega_2 \;<\; \frac{1}{|\log(1-\pi^\star)|}.$$

Since we are free to choose u as long as $u \geq 1$, we choose

$$u \;=\; \left\lceil \frac{3}{1 - \omega_2\,|\log(1-\pi^\star)|} \right\rceil \;\geq\; 1,$$

where $\lceil \cdot \rceil$ is the ceiling function; note that $\omega_2\,|\log(1-\pi^\star)| < 1$ by the choice of ω_2, which implies that $u \geq 1$. Then

$$\begin{aligned}
\sum_{k=1}^{\infty} (\omega_1 + \omega_2)\, k\, \frac{u!}{A^u\, k^{u\,(1-\omega_2\,|\log(1-\pi^\star)|)}} \;&\leq\; \sum_{k=1}^{\infty} (\omega_1 + \omega_2)\, k\, \frac{u!}{A^u\, k^3} \\
&=\; \frac{(\omega_1 + \omega_2)\, u!}{A^u} \sum_{k=1}^{\infty} \frac{1}{k^2} \;<\; \infty.
\end{aligned}$$

Therefore, by the series comparison test,

$$\sum_{k=1}^{\infty} (\omega_1 + \omega_2\, \log(k))\, \exp(-A\, k^{1-\omega_2\,|\log(1-\pi^\star)|}) \;<\; \infty,$$

because the series is dominated by a convergent power series and $\pi^\star < 1 - \zeta \in (0, 1)$ by (D.12). Thus, we have shown that there exists a universal constant $1 \leq C < \infty$ such that $|||\mathcal{D}|||_2 \leq C$ under Assumption B.1.

Bounding $|||\mathcal{D}|||_2$ *under Assumption B.2:* Using (D.14),

$$\begin{aligned}
|||\mathcal{D}|||_2 &\leq 1 + \max_{1 \leq i \leq M} \sum_{j=1:\, j \neq i}^{M} \mathbb{B}_{\boldsymbol{\pi}}(i \not\longleftrightarrow j) \\
&\leq 1 + \max_{1 \leq i \leq M} \sum_{j=1:\, j \neq i}^{M} (1 - (1 - \pi^\star)^{2\,D^2})^{d_{\mathcal{G}^\star}(i,j)} \\
&\leq 1 + \sum_{k=1}^{\infty} 2\,D^2\, g(k)\, (1 - (1 - \pi^\star)^{2\,D^2})^k,
\end{aligned}$$

where the number of vertices $j \in \mathcal{V}$ at distance k in $\mathcal{G}_{i,j}^\star$ to a given vertex i is bounded above by $2\,D^2\,g(k)$, using the above bound which showed $\binom{D+1}{2} \leq 2\,D^2$ under the assumption $D \geq 2$. Thus,

$$2\,D^2 \sum_{k=1}^{\infty} g(k)\,(1 - (1 - \pi^\star)^{2\,D^2})^k \quad < \quad \infty,$$

provided $g(k)$ is subexponential in the sense that, for all $\epsilon > 0$, however small, there exists $k_0(\epsilon) > 0$ such that

$$\frac{\log\, g(k)}{k} \quad \leq \quad \epsilon \quad \text{for all} \quad k \; > \; k_0(\epsilon).$$

This conclusion may be drawn from the root test, which shows that

$$\begin{aligned}
&\lim_{k \longrightarrow \infty} \sqrt[k]{g(k)\,(1 - (1 - \pi^\star)^{2D^2})^k} \\
&= \lim_{k \longrightarrow \infty} \exp\left(\frac{\log\, g(k)}{k} + \log(1 - (1 - \pi^\star)^{2D^2})\right) \\
&= 1 - (1 - \pi^\star)^{2D^2} \quad < \quad 1,
\end{aligned}$$

where $\pi^\star < 1 - \zeta \in (0, 1)$ by (D.12). Thus, there exists a universal constant $1 \leq C < \infty$ such that $|||\mathcal{D}|||_2 \leq C$ under Assumption B.2.

Lemma 11. *Consider Models 1–4 with* $\boldsymbol{\theta}_N \in \mathbb{R}^p$*, where* $p = N$ *under Model 1 and* $p = N + 1$ *under Models 2–4,* $\alpha = 0$ *under Models 1, 2, and 4, and* $\alpha \in [0, 1/2)$ *under Model 3. Then there exist functions* $L_k : \mathbb{R}^p \mapsto (0, 1)$ *and* $U_k : \mathbb{R}^p \mapsto (0, 1)$ $(k = 0, 1)$ *such that, for all* $\{i, j\} \subset \mathcal{N}$ *and* $\boldsymbol{x}_{-\{i,j\}} \in \{0, 1\}^{M-1}$,

$$0 \quad < \quad L_k(\boldsymbol{\theta}_N) \quad \leq \quad \mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = k \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}}) \quad \leq \quad U_k(\boldsymbol{\theta}_N) \quad < \quad 1.$$

Under the sparse β-model (which includes Model 1 as a special case with $\alpha = 0$),

$$\begin{aligned} L_0(\boldsymbol{\theta}_N) &= \frac{1}{1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty)} \\ U_0(\boldsymbol{\theta}_N) &= \frac{1}{1 + \exp(-2\, \|\boldsymbol{\theta}_N\|_\infty)\, N^{-\alpha}} \\ L_1(\boldsymbol{\theta}_N) &= \frac{N^{-\alpha}}{1 + \exp(2\, \|\boldsymbol{\theta}_N\|_\infty)} \\ U_1(\boldsymbol{\theta}_N) &= \frac{1}{1 + \exp(-2\, \|\boldsymbol{\theta}_N\|_\infty)}. \end{aligned}$$

Under Model 3, for all pairs of nodes $\{i, j\}$ satisfying $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$,

$$\begin{aligned} L_0(\boldsymbol{\theta}_N) &= \frac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)} \\ U_0(\boldsymbol{\theta}_N) &= \frac{1}{1 + \exp(-(3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)\, N^{-\alpha}} \\ L_1(\boldsymbol{\theta}_N) &= \frac{N^{-\alpha}}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)} \\ U_1(\boldsymbol{\theta}_N) &= \frac{1}{1 + \exp(-(3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}. \end{aligned}$$

Under Model 3, for all pairs of nodes $\{i, j\}$ satisfying $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$, and under Models 2 and 4,

$$\begin{aligned} L_0(\boldsymbol{\theta}_N) &= L_1(\boldsymbol{\theta}_N) = \frac{1}{1 + \exp((3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)} \\ U_0(\boldsymbol{\theta}_N) &= U_1(\boldsymbol{\theta}_N) = \frac{1}{1 + \exp(-(3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty)}. \end{aligned}$$

Proof of Lemma 11. To prove Lemma 11, we return to the notation used in Section 2 of the manuscript, denoting edge variables by $X_{i,j}$ ($\{i, j\} \subset \mathcal{N}$). Consider any pair of nodes $\{i, j\} \subset \mathcal{N}$ and any $\boldsymbol{x}_{-\{i,j\}} \in \{0, 1\}^{\binom{N}{2} - 1}$. We can express the full conditional probabilities

$$\mathbb{P}_{\boldsymbol{\theta}_N}(X_{i,j} = x_{i,j} \mid \boldsymbol{X}_{-\{i,j\}} = \boldsymbol{x}_{-\{i,j\}})$$

as

$$\frac{\exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j})\rangle)\, N^{-\alpha\, x_{i,j}}}{\exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=0)\rangle) + \exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=1)\rangle)\, N^{-\alpha}}$$

$$= \quad \frac{1}{g(0;\, \boldsymbol{x}_{-\{i,j\}},\, x_{i,j},\, \boldsymbol{\theta}_N)\, N^{\alpha\, x_{i,j}} + g(1;\, \boldsymbol{x}_{-\{i,j\}},\, x_{i,j},\, \boldsymbol{\theta}_N)\, N^{\alpha\, (x_{i,j}-1)}}$$

provided $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset$, and

$$\frac{\exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j})\rangle)}{\exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=0)\rangle) + \exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=1)\rangle)}$$

$$= \quad \frac{1}{g(0;\, \boldsymbol{x}_{-\{i,j\}},\, x_{i,j},\, \boldsymbol{\theta}_N) + g(1;\, \boldsymbol{x}_{-\{i,j\}},\, x_{i,j},\, \boldsymbol{\theta}_N)}$$

provided $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$. Here, $\alpha = 0$ under Models 1, 2, and 4, and $\alpha \in [0,\, 1/2)$ under Model 3, and

$$g(y;\, \boldsymbol{x}_{-\{i,j\}},\, x_{i,j},\, \boldsymbol{\theta}_N) \quad = \quad \exp(\langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, y) - s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j})\rangle),$$

where $y \in \{0, 1\}$. We have

$$\max_{\boldsymbol{x}_{-\{i,j\}} \,\in\, \{0,1\}^{M-1}} \left| s_l(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=0) - s_l(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=1) \right|$$

$$= \quad \begin{cases} 0 & \text{if } l \in \{i, j\} \\ 1 & \text{if } l \notin \{1, \dots, N\} \setminus \{i, j\} \\ 2\, D + 1 & \text{if } l = N + 1. \end{cases}$$

The bound on s_{N+1} follows from the definition of s_{N+1},

$$s_{N+1}(\boldsymbol{x}) \quad = \quad \sum_{\{a,b\} \,\subset\, \mathcal{N}} x_{a,b}\, \mathbb{1}\left(\sum_{h \,\in\, \mathcal{N}_a \,\cap\, \mathcal{N}_b} x_{a,h}\, x_{b,h} > 0 \right),$$

along with the fact that $|\mathcal{N}_a \cap \mathcal{N}_b| \leq |\mathcal{N}_a| \leq D$ for all $\{a, b\} \subset \mathcal{N}$ and the number of summands indexed by $\{a, b\}$ which are a function of $x_{i,j}$ is bounded above by $1 + |\mathcal{N}_i| + |\mathcal{N}_j| \leq 2\, D + 1$. As a result,

$$\begin{aligned} & \left| \langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=0) \rangle - \langle \boldsymbol{\theta}_N,\, s(\boldsymbol{x}_{-\{i,j\}},\, x_{i,j}=0) \rangle \right| \\ \leq \quad & (2 + (2\, D + 1)\, \mathbb{1}(\theta_{N+1} \neq 0))\, \|\boldsymbol{\theta}_N\|_\infty \;\leq\; (3 + 2\, D)\, \|\boldsymbol{\theta}_N\|_\infty. \end{aligned} \tag{D.15}$$

We thus obtain the bounds stated in Lemma 11.

D.2.4. *Auxiliary results.* We prove Lemmas 12 and 13.

Lemma 12. *Choose any $i \in \{1, \ldots, M\}$ and any $\boldsymbol{x}_{1:i-1} \in \{0,1\}^{i-1}$. Then the coupling of the conditional distributions*

$$\mathbb{P}(\,\cdot \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\; X_i = 0)$$

and

$$\mathbb{P}(\,\cdot \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\; X_i = 1)$$

constructed in Lemma 10 is a valid coupling.

Proof of Lemma 12. Denote the coupling distribution generated by the algorithm in Lemma 10 by $\mathbb{Q}^{\text{opt}}_{\boldsymbol{x}_{1:i-1}}$ and let $v_1, \ldots, v_{M-i}$ be the vertices added to the set $\mathfrak{V}$ at iteration $1, \ldots, M-i$ of the algorithm. To reduce the notational burden, define

$$\begin{aligned} & q(\boldsymbol{x}^{\star}_{a:b},\; \boldsymbol{x}^{\star\star}_{a:b} \mid \boldsymbol{x}^{\star}_{c:d},\; \boldsymbol{x}^{\star\star}_{c,d}) \\ = \;\; & \mathbb{Q}^{\text{opt}}_{\boldsymbol{x}_{1:i-1}}(\boldsymbol{X}^{\star}_{a:b} = \boldsymbol{x}^{\star}_{a:b},\; \boldsymbol{X}^{\star\star}_{a:b} = \boldsymbol{x}^{\star\star}_{a:b} \mid \boldsymbol{X}^{\star}_{c:d} = \boldsymbol{x}^{\star}_{c:d},\; \boldsymbol{X}^{\star\star}_{c:d} = \boldsymbol{x}^{\star\star}_{c:d}), \end{aligned}$$

where $a, b, c, d \in \{1, \ldots, M\}$ are distinct integers and $\{a, \ldots, b\} \cap \{c, \ldots, d\} = \emptyset$. By construction,

$$q(\boldsymbol{x}^{\star}_{i+1:M},\; \boldsymbol{x}^{\star\star}_{i+1:M}) \;\; = \;\; q(x^{\star}_{v_1},\; x^{\star\star}_{v_1}) \prod_{l=2}^{M-i} q(x^{\star}_{v_l},\; x^{\star\star}_{v_l} \mid \boldsymbol{x}^{\star}_{v_1,\ldots,v_{l-1}},\; \boldsymbol{x}^{\star\star}_{v_1,\ldots,v_{l-1}}).$$

Observe that

$$\begin{aligned} & \sum_{x^{\star}_{v_{M-i}} \in \{0,1\}} q(x^{\star}_{v_{M-i}},\; x^{\star\star}_{v_{M-i}} \mid \boldsymbol{x}^{\star}_{v_1,\ldots,v_{M-i-1}},\; \boldsymbol{x}^{\star\star}_{v_1,\ldots,v_{M-i-1}}) \\ = \;\; & \mathbb{P}(X_{v_{M-i}} = x^{\star\star}_{v_{M-i}} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\; X_i = 1,\; \boldsymbol{X}_{v_1,\ldots,v_{M-i-1}} = \boldsymbol{x}^{\star\star}_{v_1,\ldots,v_{M-i-1}}) \end{aligned}$$

and

$$\begin{aligned} & \sum_{x^{\star\star}_{v_{M-i}} \in \{0,1\}} q(x^{\star}_{v_{M-i}},\; x^{\star\star}_{v_{M-i}} \mid \boldsymbol{x}^{\star}_{v_1,\ldots,v_{M-i-1}},\; \boldsymbol{x}^{\star\star}_{v_1,\ldots,v_{M-i-1}}) \\ = \;\; & \mathbb{P}(X_{v_{M-i}} = x^{\star}_{v_{M-i}} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\; X_i = 1,\; \boldsymbol{X}_{v_1,\ldots,v_{M-i-1}} = \boldsymbol{x}^{\star}_{v_1,\ldots,v_{M-i-1}}), \end{aligned}$$

owing to the fact that $(X^{\star}_{v_{M-i}},\, X^{\star\star}_{v_{M-i}})$ is distributed according to the optimal coupling of the conditional distributions

$$\mathbb{P}(X_{v_{M-i}} = \,\cdot \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\; X_i = 0,\; \boldsymbol{X}_{v_1,\ldots,v_{M-i-1}} = \boldsymbol{x}^{\star}_{v_1,\ldots,v_{M-i-1}})$$

and

$$\mathbb{P}(X_{v_{M-i}} = \,\cdot \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\ X_i = 1,\ \boldsymbol{X}_{v_1,\dots,v_{M-i-1}} = \boldsymbol{x}^{\star\star}_{v_1,\dots,v_{M-i-1}}).$$

By induction,

$$\sum_{x^{\star}_{v_1} \in \{0,1\}} \cdots \sum_{x^{\star}_{v_{M-i}} \in \{0,1\}} q(\boldsymbol{x}^{\star}_{v_1,\dots,v_{M-i}},\ \boldsymbol{x}^{\star\star}_{i+1:M})$$

$$= \quad \mathbb{P}(\boldsymbol{X}_{i+1:M} = \boldsymbol{x}^{\star\star}_{1+i:M} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\ X_i = 1)$$

and

$$\sum_{x^{\star\star}_{v_1} \in \{0,1\}} \cdots \sum_{x_{v^{\star\star}_{M-i}} \in \{0,1\}} q(\boldsymbol{x}^{\star}_{i+1:M},\ \boldsymbol{x}^{\star\star}_{v_1,\dots,v_{M-i}})$$

$$= \quad \mathbb{P}(\boldsymbol{X}_{i+1:M} = \boldsymbol{x}^{\star}_{1+i:M} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\ X_i = 0),$$

so the coupling is indeed a valid coupling of the conditional distributions

$$\mathbb{P}(\boldsymbol{X}_{i+1:M} = \,\cdot \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\ X_i = 0)$$

and

$$\mathbb{P}(\boldsymbol{X}_{i+1:M} = \,\cdot \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1},\ X_i = 1).$$

Lemma 13. *Consider Models 1–4, any* $v \in \{1, \dots, M\}$, *and any* $(\boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}) \in \{0,1\}^{M-1} \times \{0,1\}^{M-1}$. *Define*

$$\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} \quad = \quad \|\mathbb{P}(\,\cdot \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v}) - \mathbb{P}(\,\cdot \mid \boldsymbol{X}_{-v} = \boldsymbol{x}'_{-v})\|_{TV}$$

and

$$\pi^{\star} \quad = \quad \max_{1 \le v \le M} \ \max_{(\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}) \,\in\, \{0,1\}^{M-1} \times \{0,1\}^{M-1}} \pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}}.$$

Assume that $\max_{1 \le i \le N} |\mathcal{N}_i| < D$ *(*$D \ge 0$*). Then*

$$\pi^{\star} \quad \le \quad \begin{cases} 0 & \textit{under Model 1} \\ \dfrac{1}{1 + \exp(-(3 + 2\,D)\,\|\boldsymbol{\theta}^{\star}_N\|_{\infty})} & \textit{under Models 2–4.} \end{cases}$$

Proof of Lemma 13. Under Model 1, edge variables X_v are independent, which implies that $\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} = 0$ for all $v \in \{1, \dots, M\}$ and all $(\boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}) \in \{0,1\}^{M-1} \times \{0,1\}^{M-1}$, which in turn implies that $\pi^{\star} = 0$. To bound $\pi^{\star}$ under Models 2–4, we distinguish two cases:

(a) If edge variable X_v corresponds to a pair of nodes with non-intersecting neighborhoods, then X_v is independent of all other edge variables by Proposition 1. As a result, $\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} = 0$ for all $(\boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}) \in \{0,1\}^{M-1} \times \{0,1\}^{M-1}$, so $\pi^\star = 0$.

(b) If edge variable X_v corresponds to a pair of nodes with intersecting neighborhoods, then X_v is not independent of all other edges, implying $\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} > 0$ for some or all $(\boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}) \in \{0,1\}^{M-1} \times \{0,1\}^{M-1}$.

We focus henceforth on case (b). Consider any $v \in \{1, \dots, M\}$ such that $\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} > 0$ for some $(\boldsymbol{x}_{-v},\, \boldsymbol{x}'_{-v}) \in \{0,1\}^{M-1} \times \{0,1\}^{M-1}$ and define

$$a_0 \;=\; \mathbb{P}(X_v = 0 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v}) \quad \text{and} \quad a_1 \;=\; \mathbb{P}(X_v = 1 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v})$$

$$b_0 \;=\; \mathbb{P}(X_v = 0 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}'_{-v}) \quad \text{and} \quad b_1 \;=\; \mathbb{P}(X_v = 1 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}'_{-v}).$$

Then

$$\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} = \frac{1}{2}\left(|(1-a_1) - (1-b_1)| + |a_1 - b_1|\right) = |a_1 - b_1| \leq \max\{a_1,\, b_1\}.$$

By symmetry,

$$\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} \;\leq\; \max\{a_0,\, b_0\},$$

which implies that

$$\pi_{v,\,\boldsymbol{x}_{-v},\,\boldsymbol{x}'_{-v}} \;\leq\; \min\left\{\max\{a_0,\, b_0\},\; \max\{a_1,\, b_1\}\right\}.$$

Lemma 11 shows that, under Models 2–4,

$$\mathbb{P}(X_v = 0 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v}) \;\leq\; \frac{1}{1 + \exp(-(3 + 2\,D)\,\|\boldsymbol{\theta}_N^\star\|_\infty)}$$

and

$$\mathbb{P}(X_v = 1 \mid \boldsymbol{X}_{-v} = \boldsymbol{x}_{-v}) \;\leq\; \frac{1}{1 + \exp(-(3 + 2\,D)\,\|\boldsymbol{\theta}_N^\star\|_\infty)}.$$

We may therefore conclude that, under Models 2–4,

$$\pi^\star \;\leq\; \min\left\{\max\{a_0,\, b_0\},\; \max\{a_1,\, b_1\}\right\} \;\leq\; \frac{1}{1 + \exp(-(3 + 2\,D)\,\|\boldsymbol{\theta}_N^\star\|_\infty)}.$$

REFERENCES

[1] Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008), "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014.

[2] Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, 41, 2097–2122.

[3] Anandkumar, A., Tan, V. Y. F., Huang, F., and Willsky, A. S. (2012), "High-dimensional structure estimation in Ising models: Local separation criterion," *The Annals of Statistics*, 40, 1346–1375.

[4] Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, 36, 192–225.

[5] Bhamidi, S., Bresler, G., and Sly, A. (2011), "Mixing time of exponential random graphs," *The Annals of Applied Probability*, 21, 2146–2170.

[6] Bhattacharya, B. B., and Mukherjee, S. (2018), "Inference in Ising models," *Bernoulli*, 24, 493–525.

[7] Bickel, P. J., and Chen, A. (2009), "A nonparametric view of network models and Newman-Girvan and other modularities," in *Proceedings of the National Academy of Sciences*, Vol. 106, pp. 21068–21073.

[8] Bresler, G., and Karzand, M. (2020), "Learning a tree-structured Ising model in order to make predictions," *The Annals of Statistics*, 48, 713–737.

[9] Brown, L. (1986), *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, Hayworth, CA, USA: Institute of Mathematical Statistics.

[10] Burt, R. S. (1992), *Structural Holes: The Social Structure of Competition*, Cambridge, MA: Harvard University Press.

[11] Butts, C. T. (2020), "A dynamic process interpretation of the sparse ERGM reference model," *Journal of Mathematical Sociology.*

[12] Cai, D., Campbell, T., and Broderick, T. (2016), "Edge-exchangeable graphs and sparsity," in *Advances in Neural Information Processing Systems*, eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., pp. 4249–4257.

[13] Caimo, A., and Friel, N. (2011), "Bayesian inference for exponential random graph models," *Social Networks*, 33, 41–55.

[14] Caron, F., and Fox, E. B. (2017), "Sparse graphs using exchangeable random measures," *Journal of the Royal Statistical Society, Series B (with discussion)*, 79, 1–44.

[15] Chatterjee, S. (2007), "Estimation in spin glasses: A first step," *The Annals of Statistics*, 35, 1931–1946.

[16] Chatterjee, S., and Diaconis, P. (2013), "Estimating and understanding exponential random graph models," *The Annals of Statistics*, 41, 2428–2461.

[17] Chatterjee, S., Diaconis, P., and Sly, A. (2011), "Random graphs with a given degree sequence," *The Annals of Applied Probability*, 21, 1400–1435.

[18] Chazottes, J. R., Collet, P., Külske, C., and Redig, F. (2007), "Concentration inequalities for random fields via coupling," *Probability Theory and Related Fields*, 137, 201–225.

[19] Chen, M., Kato, K., and Leng, C. (2019), "Analysis of networks via the sparse β-model," *arXiv preprint arXiv:1908.03152.*

[20] Comets, F. (1992), "On consistency of a class of estimators for exponential families of Markov random fields on the lattice," *The Annals of Statistics*, 20, 455–468.

[21] Crane, H., and Dempsey, W. (2018), "Edge exchangeable models for interaction networks," *Journal of the American Statistical Association*, 113, 1311–1326.

[22] Csiszár, I., and Talata, Z. (2006), "Consistent estimation of the basic neighborhood of Markov random fields," *The Annals of Statistics*, 34, 123–145.

[23] Dawid, A. P. (1979), "Conditional independence in statistical theory," *Journal of the Royal Statistical Society, Series B*, 41, 1–31.

[24] — (1980), "Conditional independence for statistical operations," *The Annals of Statistics*, 8, 598–617.

[25] Erdős, P., and Rényi, A. (1960), "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–61.

[26] Fan, Y., Zhang, H., and Yan, T. (2020), "Asymptotic theory for differentially private generalized β-models with parameters increasing," ArXiv:2002.12733v1.

[27] Fortuin, C. M., Kasteleyn, P. W., and Ginibre, J. (1971), "Correlation inequalities on some partially ordered sets," *Communications in Mathematical Physics*, 22, 89–103.

[28] Frank, O., and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

[29] Gao, C., Lu, Y., and Zhou, H. H. (2015), "Rate-optimal graphon estimation," *The Annals of Statistics*, 43, 2624–2652.

[30] Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017), "Achieving optimal misclassification proportion in stochastic block models," *Journal of Machine Learning Research*, 18, 1980–2024.

[31] Geiger, D., Heckerman, D., King, H., and Meek, C. (2001), "Stratified exponential families: graphical models and model selection," *The Annals of Statistics*, 29, 505–529.

[32] Georgii, H. O., Häggström, O., and Maes, C. (2001), "The random geometry of equilibrium phases," in *Phase transitions and critical phenomena*, Elsevier, Vol. 18, pp. 1–142.

[33] Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society, Series B*, 54, 657–699.

[34] Ghosal, P., and Mukherjee, S. (2020), "Joint estimation of parameters in Ising model," *The Annals of Statistics*, 48, 785–810.

[35] Gidas, B. (1986), "Consistency of maximum likelihood and pseudo-likelihood estimation for Gibbs distributions," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, eds. Fleming, W., and Lions, P. L., New York: Springer-Verlag, pp. 1–17.

[36] Gould, R. V., and Fernandez, R. M. (1989), "Structures of mediation: A formal approach to brokerage in transaction networks," *Sociological Methodology*, 19, 89–126.

[37] Handcock, M. S. (2003), "Statistical Models for Social Networks: Inference and Degeneracy," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, eds. Breiger, R., Carley, K., and Pattison, P., Washington, D.C.: National Academies Press, pp. 1–12.

[38] Hillar, C. J., and Wibisono, A. (2015), "A Hadamard-type lower bound for symmetric diagonally dominant positive matrices," *Linear Algebra and its Applications*, 472, 135–141.

[39] Hoff, P. D. (2021), "Additive and multiplicative effects network models," *Statistical Science*, to appear.

[40] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

[41] Holland, P. W., and Leinhardt, S. (1972), "Some evidence on the transitivity of

positive interpersonal sentiment," *American Journal of Sociology*, 77, 1205–1209.

[42] — (1981), "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76, 33–65.

[43] Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008), "Goodness of fit of social network models," *Journal of the American Statistical Association*, 103, 248–258.

[44] Hunter, D. R., and Handcock, M. S. (2006), "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

[45] Jensen, J. L., and Künsch, H. R. (1994), "On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes," *Annals of the Institute of Mathematical Statistics*, 46, 475–486.

[46] Jensen, J. L., and Møller, J. (1991), "Pseudolikelihood for exponential family models of spatial point processes," *The Annals of Applied Probability*, 1, 445–461.

[47] Jonasson, J. (1999), "The random triangle model," *Journal of Applied Probability*, 36, 852–876.

[48] Karwa, V., and Slavković, A. B. (2016), "Inference using noisy degrees: Differentially private β-model and synthetic graphs," *The Annals of Statistics*, 44, 87–112.

[49] Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, New York: Springer-Verlag.

[50] Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011), "Adjusting for network size and composition effects in exponential-family random graph models," *Statistical Methodology*, 8, 319–339.

[51] Krivitsky, P. N., and Kolaczyk, E. D. (2015), "On the question of effective sample size in network modeling: An asymptotic inquiry," *Statistical Science*, 30, 184–198.

[52] Lauritzen, S. (1996), *Graphical Models*, Oxford, UK: Oxford University Press.

[53] Lauritzen, S., Rinaldo, A., and Sadeghi, K. (2018), "Random networks, graphical models and exchangeability," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 481–508.

[54] Lauritzen, S., and Spiegelhalter, D. (1988), "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society, Series B (with discussion)*, 50, 157–224.

[55] Lehmann, E. L. (1983), *Theory of Point estimation*, New York: John Wiley & Sons.

[56] Lindvall, T. (2002), *Lectures On The Coupling Method*, Courier Corporation.

[57] Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2019), *Handbook of Graphical Models*, Boca Raton, Florida: CRC Press.

[58] Mase, S. (1995), "Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes," *The Annals of Applied Probability*, 5, 603–612.

[59] Meinshausen, N., and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the LASSO," *The Annals of Statistics*, 34, 1436–1462.

[60] Mukherjee, R., Mukherjee, S., and Sen, S. (2018), "Detection thresholds for the β-model on sparse graphs," *The Annals of Statistics*, 46, 1288–1317.

[61] Mukherjee, S. (2020), "Degeneracy in sparse ERGMs with functions of degrees as sufficient statistics," *Bernoulli*, 26, 1016–1043.

[62] Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087.

[63] Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, 38, 1287–1319.

[64] Rinaldo, A., Petrović, S., and Fienberg, S. E. (2013), "Maximum likelihood estimation

in the β-model," *The Annals of Statistics*, 41, 1085–1110.

[65] Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.

[66] Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society, Series B*, 77, 647–676.

[67] Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. (2020), "Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios," *Statistical Science*, 35, 627–662.

[68] Schweinberger, M., and Stewart, J. (2020), "Concentration and consistency results for canonical and curved exponential-family models of random graphs," *The Annals of Statistics*, 48, 374–396.

[69] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153.

[70] Stewart, J. R., and Schweinberger, M. (2020), "Supplement: Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension," Tech. rep., Department of Statistics, Florida State University.

[71] Strauss, D., and Ikeda, M. (1990), "Pseudolikelihood estimation for social networks," *Journal of the American Statistical Association*, 85, 204–212.

[72] Tang, M., Sussman, D. L., and Priebe, C. E. (2013), "Universally consistent vertex classification for latent positions graphs," *The Annals of Statistics*, 41, 1406–1430.

[73] van den Berg, J., and Maes, C. (1994), "Disagreement percolation in the study of Markov fields," *The Annals of Probability*, 22, 749–763.

[74] van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

[75] van Duijn, M. A. J., Gile, K., and Handcock, M. S. (2009), "A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models," *Social Networks*, 31, 52–62.

[76] Wainwright, M. J., and Jordan, M. I. (2008), "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1, 1–305.

[77] Whittle, P. (1963), "Stochastic processes in several dimensions," *Bulletin of the International Statistical Institute*, 40, 974–994.

[78] Xue, L., Zou, H., and Cai, T. (2012), "Nonconcave penalized composite conditional likelihood estimation of sparse Ising models," *The Annals of Statistics*, 40, 1403–1429.

[79] Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2019), "Statistical inference in a directed network model with covariates," *Journal of the American Statistical Association*, 114, 857–868.

[80] Yan, T., Leng, C., and Zhu, J. (2016), "Asymptotics in directed exponential random graph models with an increasing bi-degree sequence," *The Annals of Statistics*, 44, 31–57.

[81] Yan, T., and Xu, J. (2013), "A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices," *Biometrika*, 100, 519–524.

Jonathan R. Stewart
Department of Statistics
Florida State University
117 N Woodward Ave
Tallahassee, FL 32306-4330
E-mail: jrstewart@fsu.edu

Michael Schweinberger
Department of Statistics
Rice University
6100 Main St, MS-138
Houston, TX 77005-1827
E-mail: m.s@rice.edu