Learning cross-layer dependence structure in multilayer networks

Jiaheng Li
Department of Statistics, Florida State University

and

Jonathan R. Stewart[*]
Department of Statistics, Florida State University

July 27, 2023

Abstract

Multilayer networks are a network data structure in which elements in a population of interest have multiple modes of interaction or relation, represented by multiple networks called layers. We propose a novel class of models for cross-layer dependence in multilayer networks, aiming to learn how interactions in one or more layers may influence interactions in other layers of the multilayer network, by developing a class of network separable models which separate the network formation process from the layer formation process. In our framework, we are able to extend existing single layer network models to a multilayer network model with cross-layer dependence. We establish non-asymptotic bounds on the error of estimators and demonstrate rates of convergence for both maximum likelihood estimators and maximum pseudolikelihood estimators in scenarios of increasing parameter dimension. We additionally establish non-asymptotic error bounds on the multivariate normal approximation and elaborate a method for model selection which controls the false discovery rate. We conduct simulation studies which demonstrate that our framework and method work well in realistic settings which might be encountered in applications. Lastly, we illustrate the utility of our method through an application to the Lazega lawyers network.

Keywords: multilayer networks, statistical network analysis, social network analysis, network data, Markov random fields, graphical models

[*] This research was supported by the Test Resource Management Center (TRMC) within the Office of the Secretary of Defense (OSD), contract #FA807518D0002.

1 Introduction

Multilayer networks have become a recent focal point of research in the field of statistical network analysis [e.g., Lei et al., 2020, Caimo and Gollini, 2020, Arroyo et al., 2021, Krivitsky et al., 2020, Chen et al., 2022, Sosa and Betancourt, 2022, Huang et al., 2022], arising in applications where a common set of elements of a population of interest have multiple modes of interaction with or relation to other elements in the population. A prototypical example in the literature might be the Lazega law firm network [Lazega, 2001], in which attorneys within a law firm have multiple modes of linkage, which include advice seeking, friendship, collaboration, etc., each of which would form individual layers of the multilayer network [Krivitsky et al., 2020]. A multilayer network is therefore a composite of multiple individual networks, each defined by a distinct mode of interaction or relation.

Often, edges in one layer may depend on edges in another layer, giving rise to what we call cross-layer dependence. Understanding drivers of edge formation in multilayer networks requires learning dependence structures of the layers of multilayer networks. A key challenge lies in the fact that the cross-layer dependence can be varied and complex. In this work, we present a novel modeling framework for multilayer networks which provides a flexible platform for extending single-layer network models to multilayer networks, with the primary goal of learning cross-layer dependence structures of multilayer networks. A key advantage of our framework is that we are able to account for and separate out the network formation process from the layer formation process, enabling us to create a wide-range of novel classes of multilayer network models by extending popular classes of network models (e.g., exponential-family random graph models, stochastic block models, latent space models), and employing Markov random field specifications to develop flexible and comprehensive models of cross-layer dependence in multilayer networks. As a result, we are able to jointly model both network structure and cross-layer dependence through what we refer to as a network separable framework for modeling multilayer networks.

Our main contributions in this work include:

1. Introducing a novel framework for modeling cross-layer dependence in multilayer networks that synchronizes with current network models in the literature.

2. Deriving non-asymptotic theoretical guarantees in scenarios where the number of parameters tends to infinity, which establishes bounds on the:

    (a) Statistical error of both maximum likelihood and pseudolikelihood estimators.

    (b) Error of the multivariate normal approximation of estimators.

3. Elaborate a model selection algorithm which controls the false discovery rate.

The rest of the paper is organized as follows. Section 2 introduces our modeling framework and includes illustrative examples. Our main consistency results are contained in Section 3, and our multivariate normal approximation theory is presented in Section 4. We provide simulation results in Section 5 together with different testing procedures for model selection which control the false discovery rate. An application of our developed framework and methodology is given in Section 6, with a discussion presented in Section 7.

2 Modeling cross-layer dependence in multilayer networks

A multilayer network can be represented as a sequence of $1 \leq K < \infty$ random graphs $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$ each defined on a common set of $N \geq 3$ nodes, which we take without loss to be the set $\mathcal{N} = \{1, \ldots, N\}$. We call the graphs $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$ the *layers* of the network, and represent the multilayer network as the quantity $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)})$.

Connections between pairs of nodes $\{i, j\} \subset \mathcal{N}$ in each layer $k \in \{1, \ldots, K\}$ are modeled by random variables

$$X_{i,j}^{(k)} = \begin{cases} 1 & \text{nodes } i \text{ and } j \text{ are connected in layer } k \\ 0 & \text{otherwise} \end{cases}.$$

We refer to all connections across the K layers of a pair of nodes $\{i, j\} \subset \mathcal{N}$ as a *dyad* which we denote by $\boldsymbol{X}_{i,j} = (X_{i,j}^{(1)}, \ldots, X_{i,j}^{(K)}) \in \{0, 1\}^K$. A multilayer network can alternatively be represented by a collection of dyads where $\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{\{i,j\} \subset \mathcal{N}}$.

For notational ease, we will consider undirected multilayer networks, which imply that the network layers $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$ are undirected random graphs; extensions to directed multilayer networks or mixed multilayer networks with both directed and undirected layers will typically be straightforward, involving only notational adaptations in subscripts in most cases. We adopt the usual conventions for undirected networks, i.e., we assume that $X_{i,j}^{(k)} = X_{j,i}^{(k)}$ (all $\{i, j\} \subset \mathcal{N}$, $1 \leq k \leq K$) and $X_{i,i}^{(k)} = 0$ (all $i \in \mathcal{N}$, $1 \leq k \leq K$). The sample space of each layer $\boldsymbol{X}^{(k)}$ is therefore the product space $\mathbb{X}^{(k)} \coloneqq \{0, 1\}^{\binom{N}{2}}$ $(k = 1, \ldots, K)$, and the sample space $\mathbb{X}$ of $\boldsymbol{X}$ is the product space of the sample spaces of the individual layers, i.e., $\mathbb{X} \coloneqq \mathbb{X}^{(1)} \times \cdots \times \mathbb{X}^{(K)}$. The sample space of dyad $\{i, j\} \subset \mathcal{N}$ is the product space $\mathbb{X}_{i,j} \coloneqq \{0, 1\}^K$.

A challenge in the statistical modeling of network data lies in the fact that networks have many distinguishing properties, including:

1. **Sparsity.** Many real-world networks are sparse, in the sense that the expected number of edges in the network grows at a rate slower than $\binom{N}{2}$. The phenomena of network sparsity manifests in a variety of different applications, usually due to constraints, such as time or financial constraints, which can limit the number of connections any node can maintain at a given point in time [Krivitsky et al., 2011, Krivitsky and Kolaczyk, 2015, Butts, 2020].

2. **Node heterogeneity.** Different actors in a social network will have different proper-

ties, called node covariates, which can lead to differing propensities to form edges. A key example is assortative and disassortative mixing patterns in networks [McPherson et al., 2001, Krivitsky et al., 2009], as well as differences in structural patterns in the network [Albert and Barabási, 2002, Li et al., 2012].

3. **Edge dependence.** In addition to node-based effects that give rise to heterogeneity in propensities for nodes to form edges, scientific and statistical evidence suggests edges are dependent in many applications [Holland and Leinhardt, 1972, Frank, 1980, Goodreau et al., 2009, Block, 2015], and modeling single system of multiple binary random variables without replication is a challenging statistical problem inherent to many statistical network analysis applications.

Each of the above gives rise to distinct challenges for modeling network data and performing statistical inference in statistical network analysis applications, and it is not straightforward to construct models that due justice to each of these and more. To address these challenges, a plethora of statistical models have been proposed to model network data, which for single-layer networks have included exponential-families of random graph models [e.g., Lusher et al., 2013, Holland and Leinhardt, 1981, Snijders et al., 2006, Schweinberger et al., 2020], stochastic block models [e.g., Holland et al., 1983, Airoldi et al., 2008, Rohe et al., 2011], latent metric space models [e.g., Hoff et al., 2002, Tang et al., 2013, Sewell and Chen, 2015], random dot product graphs [e.g., Athreya et al., 2018, Sussman et al., 2014], exchangeable random graph models [e.g., Caron and Fox, 2017, Crane and Dempsey, 2018, Cai et al., 2016], and more. In this work, we build on the many classes of network data models for single layer networks by establishing a new framework for modeling multilayer networks that is capable of extending existing single layer network models to a multilayer network models which are capable of modeling cross-layer dependence and interactions.

2.1 Network separable models of multilayer networks

Multilayer networks are subject to the same forces and phenomena as single layer networks, as multiple modes of relation or interaction do not remove constraints or properties of nodes which are fundamental to network data applications. In order to develop a novel class of models for cross-layer dependence in multilayer networks, we extend the broad literature of single-layer network models by proposing a class of network separable multilayer networks which separates the network formation process from the layer formation process. We explain this distinction through the introduction of our modeling framework.

We introduce a network separable model for multilayer networks by specifying probability distributions on a double of networks $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{Y}$ will represent the network formation process, which we will call the *basis network*, and $\boldsymbol{X}$ will represent the realized multilayer network. We assume that $\boldsymbol{Y} \in \mathbb{Y} \coloneqq \{0,1\}^{\binom{N}{2}}$ is an undirected, single-layer network defined on the set of nodes $\mathcal{N}$ where

$$Y_{i,j} \;=\; \begin{cases} 1 & \text{nodes } i \text{ and } j \text{ are connected in the basis network} \\ \\ 0 & \text{otherwise} \end{cases},$$

for each $\{i,j\} \subset \mathcal{N}$, making the usual conventions for undirected networks mentioned previously. We consider semi-parametric families of probability distributions $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^p\}$ for $(\boldsymbol{X}, \boldsymbol{Y})$ which are absolutely continuous with respect to a σ-finite measure ν defined on $\mathcal{P}(\mathbb{X} \times \mathbb{Y})$, where $\mathcal{P}(\mathbb{X} \times \mathbb{Y})$ is the power set of $\mathbb{X} \times \mathbb{Y}$. Typically, ν will be the counting measure, however sparsity inducing reference measures are also admissible and have found application in network data applications in order to model sparse networks [Butts, 2020, Stewart and Schweinberger, 2021]. We say the family $\mathcal{F} \coloneqq \{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^p\}$ is *network separable* if each $\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{F}$ admits the form:

$$\mathbb{P}_{\boldsymbol{\theta}}(\{(\boldsymbol{x}, \boldsymbol{y})\}) \;=\; f(\boldsymbol{x}, \boldsymbol{\theta})\; g(\boldsymbol{y})\; h(\boldsymbol{x}, \boldsymbol{y})\; \psi(\boldsymbol{\theta}, \boldsymbol{y}), \qquad (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y}, \tag{1}$$

where

- $h : \mathbb{X} \times \mathbb{Y} \mapsto \{0, 1\}$ is given by

  $$h(\boldsymbol{x}, \boldsymbol{y}) \;=\; \prod_{\{i,j\} \subset \mathcal{N}} \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0)^{y_{i,j}} \; \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 = 0)^{1-y_{i,j}},$$

  where $\boldsymbol{x}_{i,j} = (x_{i,j}^{(1)}, \ldots, x_{i,j}^{(K)}) \in \mathbb{X}_{i,j}$ $(\{i,j\} \subset \mathcal{N})$.

- $f : \mathbb{X} \times \mathbb{R}^p \mapsto (0, 1)$ is given by

  $$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\theta}) \;=\; \prod_{\{i,j\} \subset \mathcal{N}} \exp \;& \left(\sum_{k=1}^{K} \theta_k \, x_{i,j}^{(k)} + \sum_{k<l}^{K} \theta_{k,l} \, x_{i,j}^{(k)} \, x_{i,j}^{(l)} + \ldots \right. \\ & \left. + \sum_{k_1 < \ldots < k_H}^{K} \theta_{k_1, k_2, \ldots, k_H} \, x_{i,j}^{(k_1)} \cdots \, x_{i,j}^{(k_H)} \right), \end{aligned}$$

  where $H \leq K$ is the highest order of cross-layer interactions included in the model. We write $\theta_{k_1, k_2, \ldots, k_h}$ to reference the h-order interaction parameter for the interaction term among layers $\{k_1, \ldots, k_h\} \subseteq \{1, \ldots, K\}$.

- $\psi : \boldsymbol{\Theta} \times \mathbb{Y} \mapsto (0, \infty)$ is defined by

  $$\psi(\boldsymbol{\theta}, \boldsymbol{y}) \;=\; \left[\sum_{\boldsymbol{x} \in \mathbb{X}} f(\boldsymbol{x}, \boldsymbol{\theta}) \, h(\boldsymbol{x}, \boldsymbol{y}) \right]^{-1},$$

  and functions to ensure summation to one so that the specification in (1) will be a valid probability mass function for $(\boldsymbol{X}, \boldsymbol{Y})$.

- $g : \mathbb{Y} \mapsto (0, 1)$ is the marginal probability mass function of $\boldsymbol{Y}$ and is assumed to be strictly positive on $\mathbb{Y}$.

The notation $\mathbb{P}_{\boldsymbol{\theta}}(\{(\boldsymbol{x}, \boldsymbol{y})\})$ is well-defined for each $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y}$, as $\mathbb{P}_{\boldsymbol{\theta}}$ is a probability measure defined on $\mathcal{P}(\mathbb{X} \times \mathbb{Y})$. In an abuse of notation, we will frequently write probability expressions $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})$ for the joint probability of $\{(\boldsymbol{x}, \boldsymbol{y})\}$, and $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ for the conditional probability of the event $\boldsymbol{X} = \boldsymbol{x}$ conditional on the event $\boldsymbol{Y} = \boldsymbol{y}$. We denote the data-generating parameter vector by $\boldsymbol{\theta}^\star \in \mathbb{R}^p$, and the corresponding probability measure and expectation operator by $\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\theta}^\star}$ and $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\theta}^\star}$, respectively.

The terminology *network separable* is motivated by the fact that the specification in (1) separates the network formation process $\boldsymbol{Y}$, specified by $g(\boldsymbol{y})$, from the layer formation process, specified by $f(\boldsymbol{x}, \boldsymbol{\theta})$. The two are joined by the function $h(\boldsymbol{x}, \boldsymbol{y})$, which ensures $\|\boldsymbol{x}_{i,j}\|_1 = 0$ whenever $Y_{i,j} = 0$ and $\|\boldsymbol{x}_{i,j}\|_1 > 0$ whenever $Y_{i,j} = 1$, and by $\psi(\boldsymbol{\theta}, \boldsymbol{y})$ which ensures the resulting product of functions will be a valid probability mass function. The latter has less of a direct role in modeling the cross-layer dependence and interaction between $\boldsymbol{X}$ and $\boldsymbol{Y}$, essentially fulfilling the role of a normalizing constant for the conditional probability distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$, as derived in Proposition 1. We call dyads $\{i, j\} \subset \mathcal{N}$ with $Y_{i,j} = 1$ *activated dyads*, as we allow edges between nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ in $\boldsymbol{X}$ if and only if $\{i, j\}$ is an activated dyad. Such specifications have the advantage of being able to specify the network formation process separately from the process that populates the layers of activated dyads, thus modeling the cross-layer dependence conditional on the network $\boldsymbol{Y}$. A pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y}$ that satisfies $h(\boldsymbol{x}, \boldsymbol{y}) = 1$ is said to be *network concordant*.

To illustrate the flexibility and generality of (1), observe that $g(\boldsymbol{y})$ is allowed to be any probability mass function for a single layer network $\boldsymbol{Y}$ (e.g., exponential-family random graph model, stochastic block model, latent space model), provided $g(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in \mathbb{Y}$. We therefore view our framework as semi-parametric as $g(\boldsymbol{y})$ need not assume a specific parametric form. Moreover, our framework can be viewed as non-parametric within the family of network separable multilayer networks when the maximal possible order interaction terms are included in (1), a point on which we further elaborate later. An important feature of our framework lies in the fact that the choice of the probability distribution for the network formation process does not directly influence inference for the cross-layer dependence structure, i.e., the choice of $g(\boldsymbol{y})$ does not directly influence inference for $\boldsymbol{\theta}^\star$. Proposition 1 demonstrates this point in the case of likelihood-based inference.

Proposition 1 *Let $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^p\}$ satisfy (1). Then the following hold:*

1. *For each $\boldsymbol{x} \in \mathbb{X}$, $\boldsymbol{Y} = \boldsymbol{y}$ ($\mathbb{P}_{\boldsymbol{\theta}}$-a.s.) for one and only one $\boldsymbol{y} \in \mathbb{Y}$.*

2. *$\boldsymbol{Y}$ is predictable via $\boldsymbol{X}$, i.e., for each $\boldsymbol{x} \in \mathbb{X}$, $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{X} = \boldsymbol{x}) = 1$ where*

$$y_{i,j} \;=\; \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0), \qquad \{i, j\} \subset \mathcal{N}.$$

3. *For all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y}$ with $h(\boldsymbol{x}, \boldsymbol{y}) = 1$,*

$$\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) \;=\; \log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}) + \log g(\boldsymbol{y}),$$

where $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ belongs to a minimal exponential family with natural parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ and is given by

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}) \;=\; \exp(\log f(\boldsymbol{x}, \boldsymbol{\theta}) + \log \psi(\boldsymbol{\theta}, \boldsymbol{y})).$$

Proposition 1 establishes a few key facts for inference of cross-layer dependence structures in network separable multilayer networks. First, we are able to observe $\boldsymbol{Y}$ through $\boldsymbol{X}$, as given any observation $\boldsymbol{x} \in \mathbb{X}$ of the multilayer network $\boldsymbol{X}$, $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{X} = \boldsymbol{x}) = 1$ for one, and only one, $\boldsymbol{y} \in \mathbb{Y}$. In other words, through the observation of $\boldsymbol{x}$, we can infer with probability 1 the corresponding $\boldsymbol{y}$ due to the form of (1). The significance of this result is that we do not need to treat the basis network $\boldsymbol{Y}$ as a latent network, which would require additional statistical and computational methodology to handle the latent missing data network. Second, we see that inference for $\boldsymbol{\theta}^\star$ is unaffected by the choice of $g(\boldsymbol{y})$; although, the statistical guarantees for estimators of $\boldsymbol{\theta}^\star$ will be indirectly influenced by the choice of $g(\boldsymbol{y})$, a point which we discuss in later sections. Moreover, the above choice for $f(\boldsymbol{x}, \boldsymbol{\theta})$ and the functional form of $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ derived in Proposition 1 establishes that $\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ corresponds to the log-likelihood of a minimal exponential family, accessing a broad literature of statistical methodology and theory [e.g., Sundberg, 2019]. We note that other specifications for $f(\boldsymbol{x}, \boldsymbol{\theta})$ are possible, but that Markov random field specifications provide a powerful class of models for dependent data [e.g., Wainwright and Jordan, 2008], and in the case of the saturated model with maximal interaction term $H = K$, it completely specifies all possible probabilities of outcomes $\boldsymbol{x}_{i,j} \in \{0,1\}^K$, presenting a non-parametric model class for network separable multilayer networks.

2.2 Example of a multilayer network with pairwise interactions

We illustrate cross-layer dependence among layers in our modeling framework by considering a network separable multilayer network model using the Markov random field specification for $f(\boldsymbol{x}, \boldsymbol{\theta})$ given in the previous section and maximal interaction term $H = 2$, i.e., we consider a Markov random field specification which includes all single-layer effects and all pairwise interaction effects between layers. We can write this model down as

$$f(\boldsymbol{x}, \boldsymbol{\theta}) \quad = \prod_{\{i,j\} \subset \mathcal{N}} \exp\left(\sum_{k=1}^{K} \theta_k \, x_{i,j}^{(k)} + \sum_{k<l}^{K} \theta_{k,l} \, x_{i,j}^{(k)} \, x_{i,j}^{(l)} \right). \tag{2}$$

The dimension of the parameter vector $\boldsymbol{\theta}$ is $\dim(\boldsymbol{\theta}) = K + \binom{K}{2}$, with K parameters governing the single-layer effects for the K layers and $\binom{K}{2}$ combinations of layers to form the pairwise interactions for the cross-layer dependence effects.

Define the (K-1)-dimensional vector $X_{i,j}^{(-k)} := (X_{i,j}^{(l)} : l \in \{1, \ldots, K\} \setminus \{k\})$ to be the vector of edge variables in $\boldsymbol{X}_{i,j}$ which excludes the edge variable $X_{i,j}^{(k)}$, i.e., the edge variable between nodes i and j in layer k. The conditional log-odds of edge $X_{i,j}^{(k)}$ takes the form:

$$\log \frac{\mathbb{P}(X_{i,j}^{(k)} = 1 \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, Y_{i,j} = 1)}{\mathbb{P}(X_{i,j}^{(k)} = 0 \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, Y_{i,j} = 1)} = \begin{cases} \theta_k + \displaystyle\sum_{l \neq k}^{K} \theta_{k,l} \, x_{i,j}^{(l)}, & \|\boldsymbol{x}_{i,j}^{(-k)}\|_1 > 0 \\ +\infty, & \|\boldsymbol{x}_{i,j}^{(-k)}\|_1 = 0 \end{cases}.$$

A primary advantage and motivation of using a parametric Markov random field specification for $f(\boldsymbol{x}, \boldsymbol{\theta})$ lies in the interpretability of the model. An effective approach to analyzing and understanding marginal network effects in such specifications is to study conditional log-odds of edges under different conditioning statements [e.g., Stewart et al., 2019]. By the form of $h(\boldsymbol{x}, \boldsymbol{y})$, when $Y_{i,j} = 1$, we require $\|\boldsymbol{x}_{i,j}\|_1 > 0$, meaning nodes i and j must have at least one connection in $\boldsymbol{X}$. This is seen through the log-odds formula above, where the log-odds of edge $X_{i,j}^{(k)}$ is equal to $+\infty$ when $\|\boldsymbol{x}_{i,j}^{(-k)}\|_1 = 0$. In contrast, when $\|\boldsymbol{x}_{i,j}^{(-k)}\|_1 > 0$, the constraint $\|\boldsymbol{x}_{i,j}\|_1 > 0$ is already satisfied, and the log-odds of edge $X_{i,j}^{(k)}$ depends on the layer specific parameter θ_k, as well as the pairwise interaction effects where edges present in other layers $l \in \{1, \ldots, K\} \setminus \{k\}$ can influence the likelihood of the edge $X_{i,j}^{(k)}$ depending on the signs and magnitudes of the pairwise interaction parameters $\theta_{k,l}$ ($\{k, l\} \subseteq \{1, \ldots K\}$).

3 Estimation of cross-layer dependence structure

Maximum likelihood estimation for network data with dependent edges faces significant computational challenges, as the normalizing constants for such models are often computationally intractable, which makes direct maximization of likelihood functions infeasible in general cases. For network separable multilayer networks satisfying (1), Proposition 1 establishes that the log-likelihood function takes the form

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;\; := \;\; \log \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) \;\; = \;\; \log \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y}) + \log \, g(\boldsymbol{y}). \tag{3}$$

Given an observation $\boldsymbol{x} \in \mathbb{X}$ of the multilayer network $\boldsymbol{X}$, and therefore an observation $\boldsymbol{y} \in \mathbb{Y}$ of $\boldsymbol{Y}$ by Proposition 1, we denote the set of maximum likelihood estimators by

$$\widehat{\boldsymbol{\Theta}} \;\; := \;\; \left\{ \boldsymbol{\theta} \in \mathbb{R}^p \, : \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) = \sup_{\boldsymbol{\theta}' \in \mathbb{R}^p} \, \ell(\boldsymbol{\theta}'; \boldsymbol{x}, \boldsymbol{y}) \right\},$$

and reference individual elements of the set by $\widehat{\boldsymbol{\theta}} \in \widehat{\boldsymbol{\Theta}}$. As Proposition 1 establishes $\log \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ to be a minimal, and by construction regular, exponential family, $|\widehat{\boldsymbol{\Theta}}| \in \{0, 1\}$, i.e., when the maximum likelihood estimator exists, the set $\widehat{\boldsymbol{\Theta}}$ will contain a unique element when non-empty [Proposition 3.11, pp. 32–33, Sundberg, 2019].

Two predominant methods of approximating $\boldsymbol{\theta}^\star$ when the likelihood function is computationally intractable have emerged in the literature. Monte Carlo maximum likelihood estimation (MCMLE) [Geyer and Thompson, 1992], which constructs a simulation-based approximation to the likelihood function in order to approximate the maximum likelihood estimator, is an established method for approximating maximum likelihood estimators in the statistical network analysis literature [Hunter and Handcock, 2006]. While able to provide accurate estimates of maximum likelihood estimators for complex models [e.g., Stewart et al., 2019, Schweinberger et al., 2020], a drawback of MCMLE, and other simulation-based estimation methodology, is the computational burden which can scale with both the complexity of the model and the size of the network [Bhamidi et al., 2011]. In settings where the computation of the MCMLE is impractical, a computationally efficient alternative is provided via the maximum pseudolikelihood estimator (MPLE) [Besag, 1974], whose application to social network analysis and to statistical network analysis dates back to Strauss and Ikeda [1990]. As Proposition 1 establishes that $\boldsymbol{Y}$ is observable through $\boldsymbol{X}$,

$$\mathbb{P}(Y_{i,j} = y_{i,j} \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y}_{-\{i,j\}} = \boldsymbol{y}_{-\{i,j\}}) \;\; = \;\; 1,$$

when $y_{i,j} = \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0)$ and $\boldsymbol{Y}_{-\{i,j\}}$ is defined to be the $(\binom{N}{2}$-1)-dimensional vector of edge variables in $\boldsymbol{Y}$ which excludes $Y_{i,j}$. As a result, if $(\boldsymbol{x}, \boldsymbol{y})$ is network concordant, then

$$\log\, \mathbb{P}(Y_{i,j} = y_{i,j} \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y}_{-\{i,j\}} = \boldsymbol{y}_{-\{i,j\}}) \;\; = \;\; 0, \qquad \text{for all } \{i,j\} \subset \mathcal{N}.$$

The log-pseudolikelihood of (1) can then be written down as

$$\widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;\coloneqq\; \sum_{\{i,j\} \subset \mathcal{N}} \sum_{k=1}^{K} \log \mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}), \tag{4}$$

provided $(\boldsymbol{x}, \boldsymbol{y})$ is network concordant and by exploiting the conditional independence properties implied by (1). We denote the set of maximum pseudolikelihood estimators of the data-generating parameter vector $\boldsymbol{\theta}^\star$ by

$$\widetilde{\boldsymbol{\Theta}} \;\; \coloneqq \;\; \left\{ \boldsymbol{\theta} \in \mathbb{R}^p \,:\, \widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) = \sup_{\boldsymbol{\theta}' \in \mathbb{R}^p} \widetilde{\ell}(\boldsymbol{\theta}'; \boldsymbol{x}, \boldsymbol{y}) \right\}.$$

Individual elements are referenced by $\widetilde{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$. Uniqueness of maximum pseudolikelihood estimators for exponential families is more complicated than for maximum likelihood estimators. However, our theoretical results establish that all elements $\widetilde{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ will all be within the same Euclidean distance to $\boldsymbol{\theta}^\star$. The assumption that $(\boldsymbol{x}, \boldsymbol{y})$ is network concordant comes at no cost since $\boldsymbol{Y}$ is predictable through $\boldsymbol{X}$, as discussed above, meaning that given an observation $\boldsymbol{x}$ of $\boldsymbol{X}$, we can find the unique network concordant pair $(\boldsymbol{x}, \boldsymbol{y})$ with probability one. The advantage of (4) is that the conditional probabilities $\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y})$ of edges in the multilayer network are often computationally tractable since the conditional distribution is a Bernoulli distribution when $Y_{i,j} = 1$, and is a degenerate point mass at 0 when $Y_{i,j} = 0$.

In this work, we consider both maximum likelihood estimators and maximum pseudolikelihood estimators. As seen from the forms of $\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ and $\widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ given above, the gradients and Hessians of the log-likelihood and log-pseudolikelihood equations do not directly depend on $g(\boldsymbol{y})$, echoed by the results in Proposition 1. However, as mentioned in the previous section, theoretical guarantees for estimators of $\boldsymbol{\theta}^\star$ will be indirectly influenced by the choice of $g(\boldsymbol{y})$, a point supported by the following lemma.

Lemma 1 *Consider a network separable family* $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^p\}$ *satisfying* (1) *and an observation* $\boldsymbol{x} \in \mathbb{X}$ *of* $\boldsymbol{X}$ *and let* $(\boldsymbol{x}, \boldsymbol{y})$ *be the network concordant pair where* $\boldsymbol{y}$ *is given by Proposition 1. Define, for each pair of nodes* $\{i, j\} \subset \mathcal{N}$*,*

$$
\begin{aligned}
L_{i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y}) &\;:=\; \log\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}_{i,j} = \boldsymbol{x}_{i,j} \,|\, \boldsymbol{Y} = \boldsymbol{y}) \\
\widetilde{L}_{i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y}) &\;:=\; \sum_{k=1}^{K} \log\, \mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}).
\end{aligned}
$$

Then there exist $p \times p$ matrices $I(\boldsymbol{\theta})$ and $\widetilde{I}(\boldsymbol{\theta})$ such that

$$\mathbb{E}\left[-\nabla^2_{\boldsymbol{\theta}} L_{i,j}(\boldsymbol{\theta}, \boldsymbol{X}_{i,j}, \boldsymbol{Y}) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \;=\; \begin{cases} I(\boldsymbol{\theta}) & Y_{i,j} = 1 \\ \mathbf{0}_{p,p} & Y_{i,j} = 0 \end{cases}$$

$$\mathbb{E}\left[-\nabla^2_{\boldsymbol{\theta}} \widetilde{L}_{i,j}(\boldsymbol{\theta}, \boldsymbol{X}_{i,j}, \boldsymbol{Y}) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \;=\; \begin{cases} \widetilde{I}(\boldsymbol{\theta}) & Y_{i,j} = 1 \\ \mathbf{0}_{p,p} & Y_{i,j} = 0, \end{cases}$$

for all $\{i, j\} \subset \mathcal{N}$, where $\mathbf{0}_{p,p}$ is the $p \times p$ matrix with all 0 entries, and

$$\lambda_{\min}(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})) \;=\; \lambda_{\min}(I(\boldsymbol{\theta}))\, \mathbb{E}\, \|\boldsymbol{Y}\|_1$$

$$\lambda_{\min}(-\mathbb{E}\, \nabla^2_{\boldsymbol{\theta}}\, \widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})) \;=\; \lambda_{\min}(\widetilde{I}(\boldsymbol{\theta}))\, \mathbb{E} \|\boldsymbol{Y}\|_1,$$

where $\lambda_{\min}(\boldsymbol{A})$ is the smallest eigenvalue of matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$.

In classical scenarios with independent and identically distributed observations, the expected negative Hessian of the log-likelihood function is the Fisher information matrix and is expected to scale with the number of observations. In such cases, standard matrix theory shows the smallest eigenvalue of the expected negative Hessian of the log-likelihood function will scale with the number of observations, provided the smallest eigenvalue of the Fisher information matrix of the population from which observations are drawn is bounded below. With regards to network separable multilayer networks, Lemma 1 demonstrates such a scaling with respect to the expected number of activated dyads $\mathbb{E}\, \|\boldsymbol{Y}\|_1$, proxying the effective sample size. In similar fashion, $I(\boldsymbol{\theta})$ is analogous to the Fisher information of the population distribution from which observations would be sampled in classical scenarios with independent and identically distributed observations, and may be regarded as the Fisher information of the population distribution for activated dyads in $\boldsymbol{Y}$. With regards to pseudolikelihood-based estimation, we have a similar interpretation.

We next present our theoretical guarantees for maximum likelihood and maximum pseudolikelihood estimators in Theorem 1. As we will show in Theorem 1, the choice of

$g(\boldsymbol{y})$ influences the estimation error though the expected number of edges in $\boldsymbol{Y}$ and through the covariances of edge variables in $\boldsymbol{Y}$. Define $[D_g]^+ \coloneqq \max\{0,\, D_g\}$, where

$$D_g \;\; \coloneqq \sum_{\{i,j\} \prec \{v,w\} \subset \mathcal{N}} \mathbb{C}(Y_{i,j},\, Y_{v,w}),$$

and where $\{i, j\} \prec \{v, w\}$ implies the sum is taken with respect to the lexicographical ordering of pairs of nodes. Let $\epsilon^\star > 0$ be fixed independent of N and p and define

$$\xi_{\epsilon^\star} \;\coloneqq\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(I(\boldsymbol{\theta})) \qquad \text{and} \qquad \widetilde{\xi}_{\epsilon^\star} \;\coloneqq\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(\widetilde{I}(\boldsymbol{\theta})),$$

where $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star) = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2 \leq \epsilon^\star\}$.

Theorem 1 *Consider a multilayer network model satisfying (1) defined on a set of $N \geq 3$ nodes and $K \geq 1$ layers and assume that $\mathbb{E}\,\|\boldsymbol{Y}\|_1 \geq 1$ and $p \leq N$. Then there exists $N_0 \geq 3$ such that, for all $N \geq N_0$, the following hold with probability at least $1 - 3\,(\mathbb{E}\,\|\boldsymbol{Y}\|_1)^{-1}$:*

1. *(MLE) The set $\widehat{\boldsymbol{\Theta}}$ is non-empty and the unique element $\widehat{\boldsymbol{\theta}} \in \widehat{\boldsymbol{\Theta}}$ satisfies*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;\;\leq\;\; \sqrt{\frac{3\,p\,\log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\;\frac{\sqrt{1 + [D_g]^+}}{\xi_{\epsilon^\star}},$$

provided the right-hand side tends to 0 as $N \to \infty$.

2. *(MPLE) The set $\widetilde{\boldsymbol{\Theta}}$ is non-empty and each $\widetilde{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ satisfies*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;\;\leq\;\; \sqrt{\frac{3\,p\,K^2\,\log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\;\frac{\sqrt{1 + [D_g]^+}}{\widetilde{\xi}_{\epsilon^\star}},$$

provided the right-hand side tends to 0 as $N \to \infty$.

The results of Theorem 1 establish a few key facts concerning statistical estimation of the parameter vector $\boldsymbol{\theta}^\star$. First, we can view the quantity $\xi_{\epsilon^\star}\,\sqrt{\mathbb{E}\,\|\boldsymbol{Y}\|_1}\,/\,\sqrt{1 + [D_g]^+}$ as the effective sample size in order to compare our results to classical settings with independent and identically distributed data. The effective sample size, together with the dimension of the model p, helps to determine the rate of convergence (with respect to the Euclidean distance) of maximum likelihood and pseudolikelihood estimators. As previously mentioned, the quantities $\mathbb{E}\|\boldsymbol{Y}\|_1$ and $[D_g]^+$ are determined by properties of $g(\boldsymbol{y})$, the marginal probability mass function of $\boldsymbol{Y}$. While specification of $g(\boldsymbol{y})$ does not directly influence estimation algorithms, the statistical guarantees of estimators will depend on $g(\boldsymbol{y})$ producing enough activated dyads and not possessing overly strong dependence among edges in the single network $\boldsymbol{Y}$. The requirement that the right-hand side of the bounds in Theorem 1 tend to 0 as $N \to \infty$ ensures that all regularity assumptions remain valid. Namely, key to our approach lies in the ability to control minimum eigenvalues of matrices $I(\boldsymbol{\theta})$ and $\widetilde{I}(\boldsymbol{\theta})$ in a neighborhood of the data-generating parameter vector $\boldsymbol{\theta}^\star$. The condition that the bounds tend to 0 ensures that it is sufficient to control the smallest eigenvalue in a bounded set, i.e., we may let $\epsilon^\star$ be fixed independent of N, and moreover, to ensure consistency in the sense that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \to 0$ and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \to 0$ (as $N \to \infty$) with probability approaching 1.

4 Error of the normal approximation and model selection

In this section, we show that a standardization of the maximum likelihood estimator (MLE) of the data-generating parameter vector $\boldsymbol{\theta}^\star$ of increasing dimension is asymptotically multivariate normal, i.e., we demonstrate a non-asymptotic bound on the error of the multivariate normal approximation and exhibit conditions on the scaling of relevant model quantities—namely the dimension of the model p together with the scaling of the expected number of activated dyads $\mathbb{E}\,\|\boldsymbol{Y}\|_1$—under which the error bound on the multivariate normal approximation vanishes in the limit. Leveraging the consistency result in Theorem 1, we may additionally exhibit the asymptotic normality of maximum pseudolikelihood estimators (MPLE). Based on this result, we present a model selection method using multiple hypothesis testing procedures that control the false discovery rate. The main result is presented in Theorem 2, the proof of which is based on a Taylor expansion of the log-likelihood function and through the application of a Lyapunov type bound presented in Raič [2019].

In the following, $\boldsymbol{Z}$ will denote a standard multivariate normal random vector, i.e., with mean vector equal to the zero vector and covariance matrix equal to the identity matrix (each of appropriate dimension), and Φ will denote the corresponding probability measure.

Theorem 2 *Under the assumptions of Theorem 1, there exists $N_0 \geq 3$ such that, for all $N \geq N_0$ and any measurable convex set $\mathcal{A} \subseteq \mathbb{R}^p$, the error of the multivariate normal approximation*

$$\left|\mathbb{P}((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})\right|$$

is bounded above by

$$\frac{83}{\xi_{\epsilon^\star}^{3/2}} \sqrt{\frac{p^{7/2}}{\mathbb{E}\, \|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\; [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}$$

where $\tilde{\boldsymbol{R}}$ satisfies

$$\mathbb{P}\left(\|\tilde{\boldsymbol{R}}\|_2 \leq \frac{3\sqrt{2}\,(1 + [D_g]^+)}{\xi_{\epsilon^\star}^2}\, \frac{p^{5/2}\, \log N}{\sqrt{\mathbb{E}\, \|\boldsymbol{Y}\|_1}}\right) \;\; \geq \;\; 1 - \frac{7}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} - \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}.$$

Theorem 2 serves as a foundation for establishing the asymptotic normality of maximum likelihood estimators $\widehat{\boldsymbol{\theta}}$ and maximum pseudolikelihood estimators $\widetilde{\boldsymbol{\theta}}$, noting Theorem 1 established conditions under which both $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ are consistent estimators of $\boldsymbol{\theta}^\star$ (with respect to the Euclidean distance metric), assumptions which are met by Theorem 2. If

$$\lim_{N \to \infty} \left[\frac{83}{\xi_{\epsilon^\star}^{3/2}} \sqrt{\frac{p^{7/2}}{\mathbb{E}\, \|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\; [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}\right] \;\; = \;\; 0,$$

Theorem 2 implies $(I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}}$ will converge in distribution to a standard multivariate normal random vector, as error bound on the multivariate normal approximation will vanish in this case. The term $\tilde{\boldsymbol{R}}$ can be viewed as an error term, resulting from the fact that the normal approximation in Theorem 2 is obtained via a multivariate Taylor approximation in order to bridge the distributional gap between key statistics which admit forms amenable to existing theorems for the normal approximation and the parameter vectors of interest, thus introducing an additional source of error in the normal approximation.

The same theory may be exported to the case of maximum pseudolikelihood estimators by exploiting the consistency of both $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ (with respect to the Euclidean distance metric) implied via Theorem 1 as the triangle inequality implies $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_2 \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 + \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2$.

While involved, the above condition for asymptotic multivariate normality essentially places restrictions on the dependence induced through the single-layer network $\boldsymbol{Y}$ measured by $[D_g]^+$, as well as the smallest eigenvalue of the dyad-based information matrix $I(\boldsymbol{\theta})$ in a neighborhood of the data-generating parameter vector $\boldsymbol{\theta}^\star$ as measured by $\xi_{\epsilon^\star}$, and the dimension of the model p. As a result, if the information matrix $I(\boldsymbol{\theta})$ is nearly singular at $\boldsymbol{\theta}^\star$, in which case $\xi_{\epsilon^\star}$ will be small, the error of the normal approximation will be uniformly larger (all else equal). Likewise, if the edge dependence in $\boldsymbol{Y}$ is large as measured by $[D_g]^+$, we may not have sufficient activated dyads to ensure the error bound is small, as $\|\boldsymbol{Y}\|_1$ may not be tightly concentrated around $\mathbb{E}\,\|\boldsymbol{Y}\|_1$. The dependence of the error approximation on the dimension of the random vector is a known challenge in establishing multivariate normality [see, e.g., Raič, 2019]. All quantities which are not explicit constants can increase or decrease with N, with the rates of these increases or decreases having implications for the rate of convergence in distribution. Theorem 2 demonstrates that the allowable scaling for most of quantities is with respect to the expected number of activated dyads $\mathbb{E}\,\|\boldsymbol{Y}\|_1$.

We further examine Theorem 2 through an example where $\boldsymbol{Y}$ is a Bernoulli random graph model, which assumes edge variables are independent Bernoulli random variables with probability $\pi \in (0, 1)$. Under this model, $[D_g]^+ = 0$ owing to the independence of edge variables and $\mathbb{E}\|\boldsymbol{Y}\|_1 = \pi \binom{N}{2}$. Under this scenario, we can show that

$$\left|\mathbb{P}((I(\boldsymbol{\theta}^\star)\,\|\boldsymbol{Y}\|_1)^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})\right| \;\leq\; \frac{166}{\sqrt{\pi\,\xi_{\epsilon^\star}^3}}\,\frac{p^{1.75}}{N} + \frac{16}{\pi N^2},$$

with the additional bound

$$\mathbb{P}\left(\|\tilde{\boldsymbol{R}}\|_2 \;\leq\; \frac{6\sqrt{2}}{\xi_{\epsilon^\star}^2}\,\frac{p^{2.5}\,\log N}{\pi\,N}\right) \;\geq\; 1 - \frac{28}{\pi\,N^2}.$$

If $\xi_{\epsilon^\star}$ and π are both bounded away from 0, then the error of the normal approximation will convergence to 0 provided $(p^{2.5} \log N) / N \to 0$ as $N \to \infty$, which is sufficient to ensure $\|\tilde{\boldsymbol{R}}\|_2$ converges in probability to 0. Under the fully saturated model specification for (1) $(H = K)$, the Binomial theorem shows that $p = 2^K - 1 \leq 2^K$. Hence, the dimension restriction on p in turn implies a restriction on the allowable rate of growth of the number of layers K with N, where a sufficient condition for $(p^{2.5} \log N) / N \to 0$ is for $K \leq .5 \log N$. In other words, the number of layers K can grow at most logarithmically with N in the fully saturated model. In cases when the number of interaction terms included in the cross-layer dependence probability model is fixed, K may admit a sublinear scaling with N.

4.1 Model selection via univariate testing with FDR control

Provided with the consistency results and the multivariate normal approximation of $\widehat{\boldsymbol{\theta}}$ through Theorems 1 and 2, we outline a procedure for model selection that controls the false discovery rate. Hotelling's T-squared statistic can be used to conduct a global test for $H_0 : \boldsymbol{\theta}^\star = \boldsymbol{\mu}$ versus $H_1 : \boldsymbol{\theta}^\star \neq \boldsymbol{\mu}$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the value of $\boldsymbol{\theta}$ we want to test [Chapter 5, Johnson and Wichern, 2002]. We will mostly be interested in the case when $\boldsymbol{\mu} = \mathbf{0}_p$, i.e., the zero vector of dimension p.

If the global test is rejected, or if the global test is not of interest, we can perform model selection by leveraging the multivariate normal approximation to obtain univariate normal approximation results for the components of $\widehat{\boldsymbol{\theta}}$ and proceed to test each component: $H_{i,0} : \theta_i^\star = \mu_i$ versus $H_{i,1} : \theta_i^\star \neq \mu_i$, for $i = 1, \ldots p$ and $\mu_i \in \mathbb{R}$. In general, $\mu_i = 0$ will allow us to test whether the estimated effect $\widehat{\theta}_i$ is present in the model (i.e., whether $\theta_i^\star \neq 0$). One challenge in this approach lies in the fact that the model selection procedure is sensitive to multiple testing error. We propose to control the multiple testing error by appropriate multiple testing adjustments by elaborating a model selection algorithm which will control the false discovery rate in order to accurately learn the cross-layer dependence effects present in the multilayer network, and in effect learning the cross-layer dependence structure of the multilayer network. We provide simulation examples of four different univariate testing procedures including Bonferroni, Benjamini-Hochberg, Hochberg, and Holm procedures in Section 5.2. In simulation studies, all four univariate testing procedures exhibit strong statistical power for detecting non-zero parameters while controlling the false discovery rate at a preset family-wise significance level. As Theorem 1 establishes the consistency of both $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, the above procedure remains justifiable for performing model selection with maximum pseudolikelihood estimators as well, as it is straightforward to prove a corollary to Theorem 1 which establishes that $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_2$ converges in probability to 0 under the assumptions of Theorem 1, further obtaining convergence in distribution.

5 Simulation studies

We conduct simulation studies to investigate the performance of the maximum pseudolikelihood estimation (MPLE) in realistic settings that could be encountered in application in order to study the realized outcomes of the theoretical results established in Sections 3 and 4. In section 5.1, we demonstrate the consistency results of Theorem 1 for pseudolikelihood estimators $\widetilde{\boldsymbol{\theta}}$ in settings of different model-generating parameters and different basis network structures of $\boldsymbol{Y}$. We study the multivariate normal approximation of $\widetilde{\boldsymbol{\theta}}$ established by Theorem 2 (and by additionally leveraging the consistency result of Theorem 1) in the simulation study conducted in Section 5.2. Lastly, we discuss several testing procedures for selecting non-zero effects while controlling the false discovery rate (FDR) at a given family-wise significance level α.

In all simulation studies, we sample network concordant multilayer networks $(\boldsymbol{X}, \boldsymbol{Y})$ from (1) with the number of nodes varying from $N = 200$ to 1000 and $K = 3$ layers. The basis network $\boldsymbol{Y}$ is generated from three different models: the Bernoulli random graph model, the stochastic block model, and the latent space model. The layer mechanism of

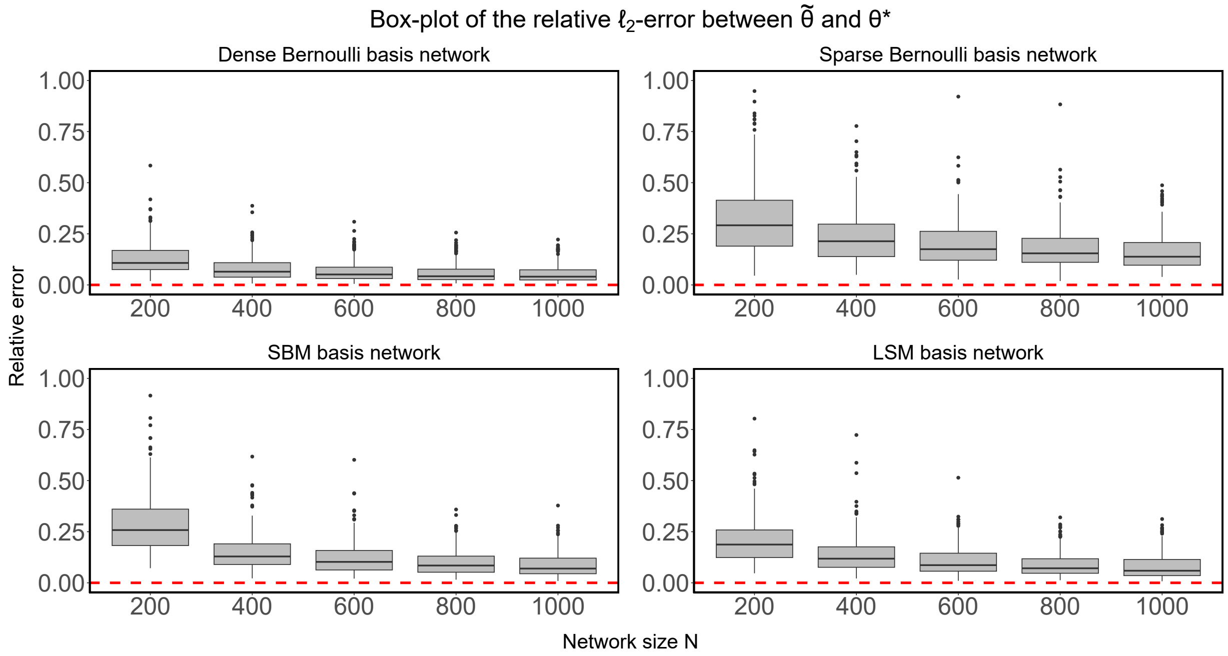

Figure 1: The relative ℓ_2-errors between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$ decrease as the network size increases in four basis network structures.

the multilayer network is given by

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{\{i,j\} \subset \mathcal{N}} \exp\left(\sum_{k=1}^{3} \theta_k \, x_{i,j}^{(k)} + \sum_{k<l}^{3} \theta_{k,l} \, x_{i,j}^{(k)} \, x_{i,j}^{(l)} \right). \tag{5}$$

5.1 Consistency of MPLE

The consistency of the maximum pseudolikelihood estimator $\widetilde{\boldsymbol{\theta}}$ is demonstrated through the decay of the relative ℓ_2-errors between $\widetilde{\boldsymbol{\theta}}$ and the data-generating parameter $\boldsymbol{\theta}^\star$. We generate $M = 250$ multilayer networks by M different model-generating parameters at five network sizes from $N = 200$ to 1000. For each network size N, type of basis network $\boldsymbol{Y}$, and replicate, we sample a network separable multilayer network $\boldsymbol{X}$ from (1) using the specification in (5) with the data-generating parameter vector $\boldsymbol{\theta}^\star$ populated by randomly selecting each component from the uniform distribution on $(-1, 1)$. We make the exception that the third and the sixth components $\theta_3^\star$ and $\theta_{1,3}^\star$ are set to 0. In each replicate, we

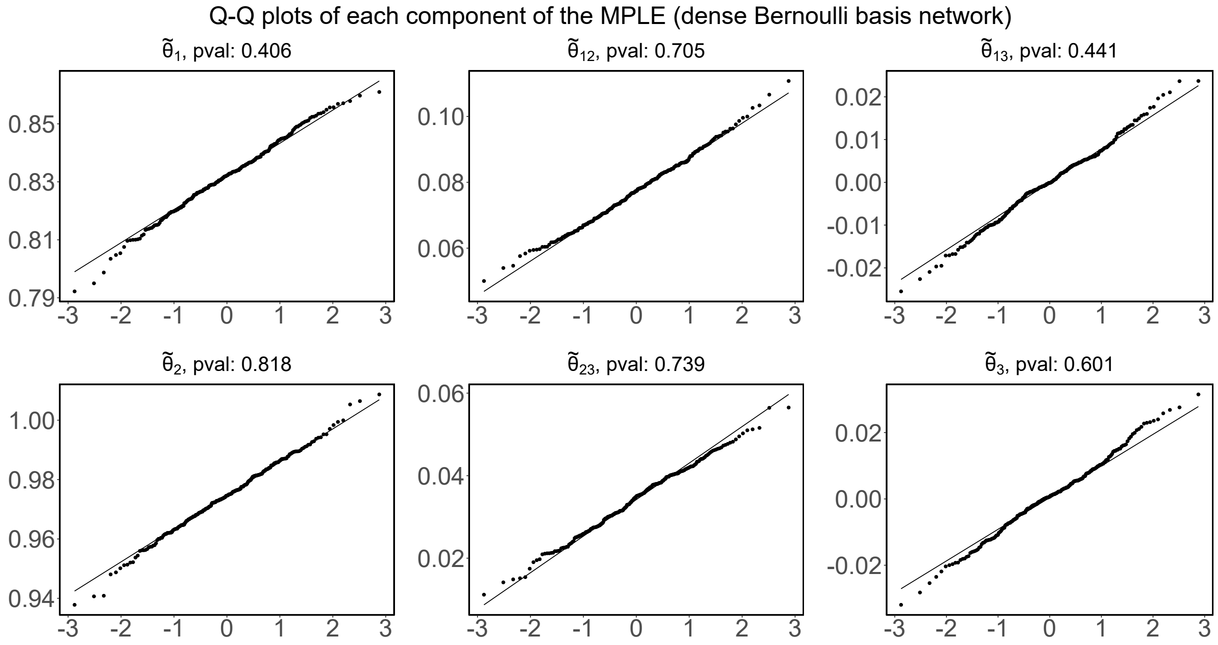

Figure 2: Q-Q plots and p-values of six components of $\widetilde{\boldsymbol{\theta}}$ estimated from 250 multilayer network samples at size 1000 on the dense Bernoulli basis network. The univariate normal test failed to reject the null hypothesis that each component of $\widetilde{\boldsymbol{\theta}}$ is marginally normal at a significance level of .05.

compute the maximum pseudolikelihood estimator. The results of this simulation study are given in Figure 1, which shows the decay of the relative ℓ_2-errors between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$ as the network size increases in four different basis network structures. The broad selection of model-generating parameter values on different basis network structures verifies that Theorem 1 holds in many practical settings.

The top-left subplot of Figure 1 shows a baseline result for a dense Bernoulli basis network with $\mathbb{P}(Y_{i,j} = 1) = 0.8$ for all $\{i, j\} \subset \mathcal{N}$. As the network size increases from 200 to 1000, the relative ℓ_2-errors decrease to 0. The performance of MPLE on different structures of the basis network $\boldsymbol{Y}$ is also studied with results being shown in the rest of the subplots of Figure 1. Basis networks $\boldsymbol{Y}$ are generated by a sparse Bernoulli random graph (top-right), by a stochastic block model (SBM, bottom-left), and by a latent space model (LSM, bottom-right). In contrast to the baseline result of the dense Bernoulli random graph, where the basis network is populated with more dyadic connections owing to the fact that the expected number of activated dyads $\mathbb{E}\|\boldsymbol{Y}\|_1$ is equal to $.8\binom{N}{2}$, the three different basis structures possess fewer connections. A key example is the sparse Bernoulli basis network which has a varying density of $\mathbb{P}(Y_{i,j} = 1) = 20/N$ for all $\{i, j\} \subset \mathcal{N}$, admitting a reciprocal scaling with the network size N which results in a sparse network with bounded average node degree. The SBM generated networks have 5 blocks where the within-block density is .5 and the between-block density is .05 for all network sizes simulated. The LSM generated networks follow the specifications of Hoff et al. [2002] with a fixed density parameter of .6. We simulate node positions on the plane in $\mathbb{R}^2$, where coordinates of the position of each node are randomly generated from the standard normal distribution. In order for SBM and LSM generated basis networks to have a comparable number of effective sample size, the parameters of the SBM and the LSM are chosen so that the expected number of activated dyads $\mathbb{E}\|\boldsymbol{Y}\|_1$ in both basis networks is approximately $.24\binom{N}{2}$.

5.2 Multivariate normality of MPLE and model selection

As stated in Section 4 and Theorem 2, the distribution of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ and the maximum pseudolikelihood estimator $\widetilde{\boldsymbol{\theta}}$ converge in distribution to a multivariate normal distribution asymptotically. In order to study the quality of the normal approximation—especially for univariate testing which would be used for false discovery control and model selection—we randomly select 6 of the 250 data-generating parameter vectors $\boldsymbol{\theta}^\star$ used to study the consistency results of Theorem 1 in the simulation study conducted in Section 5.1. We then generate 250 replicates of multilayer network samples by each of these 6 parameter vectors, using specification (5) on different basis network structures. The multivariate normality of $\widetilde{\boldsymbol{\theta}}$ passed Zhou-Shao's multivariate normal test [Zhou and Shao, 2014], with p-values provided in the Appendix G.1 in the supplement to

Table 1: False discovery rates of four procedures for detecting non-zero effects of 6 model-generating parameters ($\boldsymbol{\theta}_1^\star$, $\boldsymbol{\theta}_2^\star$, $\boldsymbol{\theta}_3^\star$, $\boldsymbol{\theta}_4^\star$, $\boldsymbol{\theta}_5^\star$, $\boldsymbol{\theta}_6^\star$) estimated from 250 multilayer network samples at size 1000 on the dense Bernoulli basis network. All FDRs are smaller than .05.

Procedure	$\boldsymbol{\theta}_1^\star$	$\boldsymbol{\theta}_2^\star$	$\boldsymbol{\theta}_3^\star$	$\boldsymbol{\theta}_4^\star$	$\boldsymbol{\theta}_5^\star$	$\boldsymbol{\theta}_6^\star$
Bonferroni	.004	.002	.001	.002	.001	.005
Benjamini-Hochberg	.014	.014	.014	.011	.017	.020
Hochberg	.012	.008	.009	.008	.011	.016
Holm	.010	.008	.006	.008	.007	.013

this paper. We visualize the marginal normality of individual component in $\widetilde{\boldsymbol{\theta}}$ with a dense Bernoulli basis network in Figure 2, through Q-Q plots of the simulated maximum pseudolikelihood estimators. Univariate tests for normality failed to reject the null hypothesis that each component of $\widetilde{\boldsymbol{\theta}}$ is marginally normal at a significance level of .05. Additional results studying the multivariate normality of $\widetilde{\boldsymbol{\theta}}$ on different basis network structures are provided in Appendix G.1 in the supplement to this paper.

We then implement the multiple testing correction procedures of Bonferroni, Benjamini-Hochberg, Hochberg, and Holm, for the 6 selected model-generating parameter vectors $\boldsymbol{\theta}^\star$ with 250 replicates to detect components that are significantly different from 0 while controlling the false discovery rate (FDR) at a family-wise significance level of $\alpha = .05$—recall the third and the sixth component $\theta_{1,3}^\star$ and $\theta_3^\star$ of $\boldsymbol{\theta}^\star$ are set to 0 in each simulation replicate. We estimate the FDR of the four procedures by averaging the false discovery proportions from 250 replicates of each of the 6 randomly selected model-generating parameters $\boldsymbol{\theta}^\star$. We provide the estimated FDRs for $\boldsymbol{\theta}^\star$ on a dense Bernoulli basis network in Table 1. In addition, we show the receiver operating characteristic (ROC) curves for $\widetilde{\boldsymbol{\theta}}$ estimating the 6 selected model-generating parameters in each of the subplot of Figure 3, on four basis

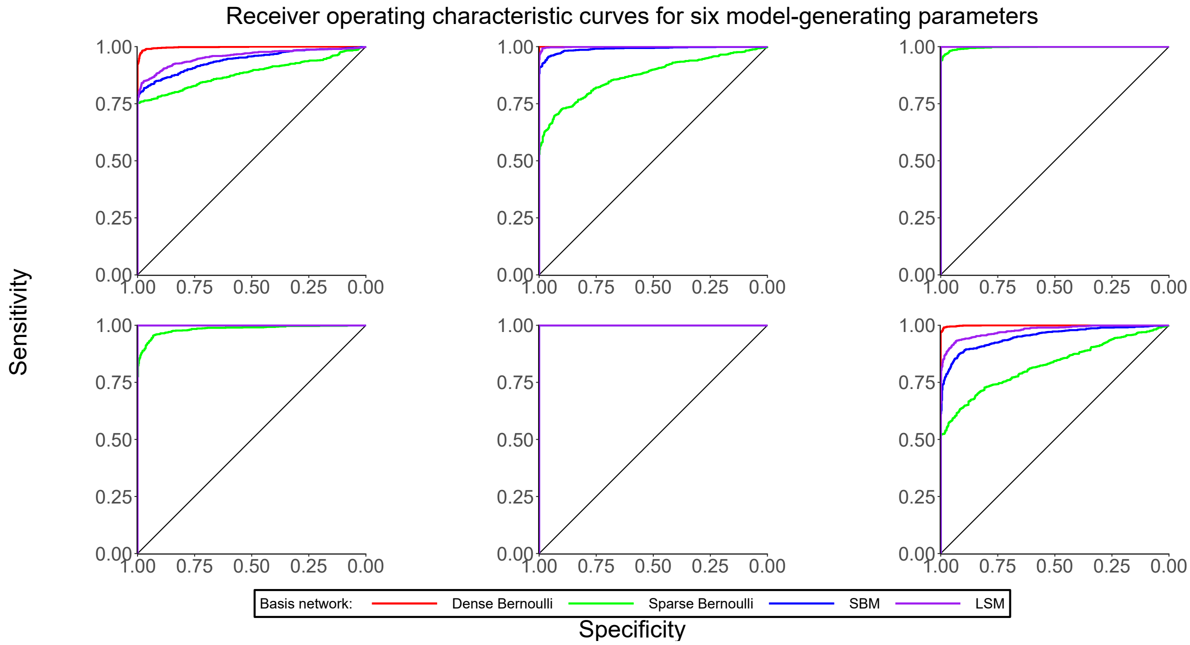

Figure 3: ROC curves for $\widetilde{\boldsymbol{\theta}}$ estimated from 250 multilayer network samples at size 1000 of six model-generating parameters on four different basis networks.

network structures. Simulation results suggest that the false discovery rate is controlled below the preset threshold α. Different model-generating parameter values affect the trade-off between the sensitivity and the specificity of the model selection. In general, multilayer networks with a larger effective sample size lead to a larger area under the ROC curve which offers a tool to choose appropriate correction procedures and thresholds for model selection in different scenarios. Additional results on the false discovery rate with different basis network structures are provided in Appendix G.2 in the supplement to the paper.

6 Application

We present a case study using a dataset on corporate law partnership among a Northeastern US corporate law firm in New England collected by Lazega [2001]. The dataset collected information about three types of cooperation among 71 lawyers in the corporate

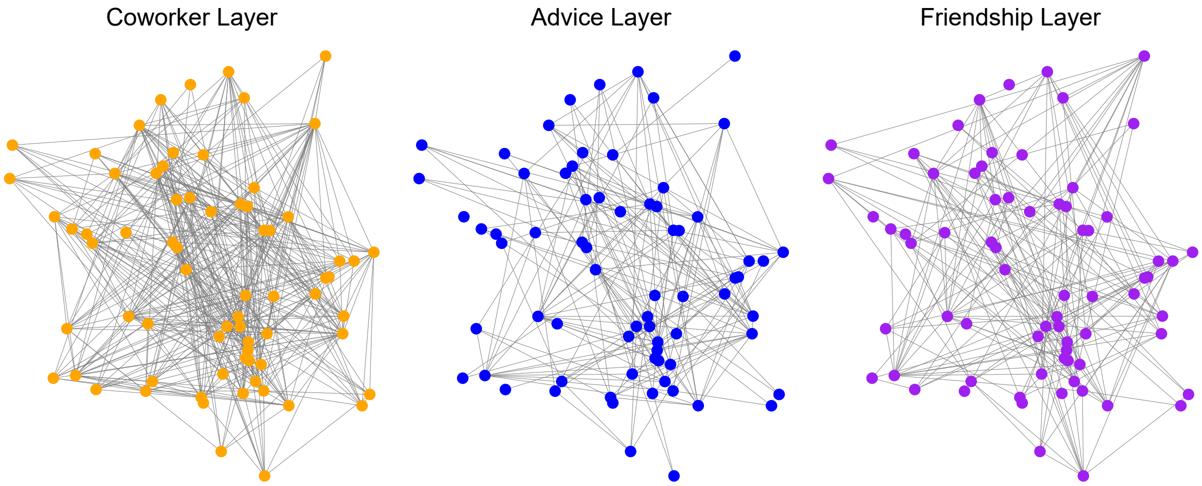

Figure 4: Coworker layer, advice layer and the friendship layer of Lazega's corporate law partnership network.

Table 2: Summary of Lazega's corporate law partnership data with 71 lawyers (nodes).

	Average Node Degree	Number of Edges
Co-Worker Layer	11	378
Advice Layer	5	175
Friendship Layer	5	176

law firm, resulting in three networks including the strong-coworker network, the advice network, and the friendship network. Since the cooperation relationship collected are not symmetric, we only consider a connection to be present when both sides acknowledged their cooperation. We treat these three types of networks as a three-layer multilayer network embedded among the 71 lawyers. A summary of this multilayer network is provided in Table 2. We apply model (1) with up to 2-layer interaction terms to the Lazega dataset, i.e., $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_{1,2}, \theta_{1,3}, \theta_{2,3})$. The maximum pseudolikelihood estimator $\widetilde{\boldsymbol{\theta}}$ is obtained for $\boldsymbol{\theta}$ and the results are provided in Table 3.

As shown in Table 3 of the MPLE of the Lazega network data, θ_1, θ_2, and θ_3 correspond

Table 3: MPLEs for parameters of the Lazega network.

$\widetilde{\theta}_1$	$\widetilde{\theta}_2$	$\widetilde{\theta}_3$	$\widetilde{\theta}_{1,2}$	$\widetilde{\theta}_{1,3}$	$\widetilde{\theta}_{2,3}$	$\mathbb{P}(Y_{i,j}=1)$
-1.450	-3.334	-2.695	1.801	0.218	2.458	0.208
Coworker (C)	Advice (A)	Friendship (F)	C × A	C × F	A × F	Basis network $\boldsymbol{Y}$

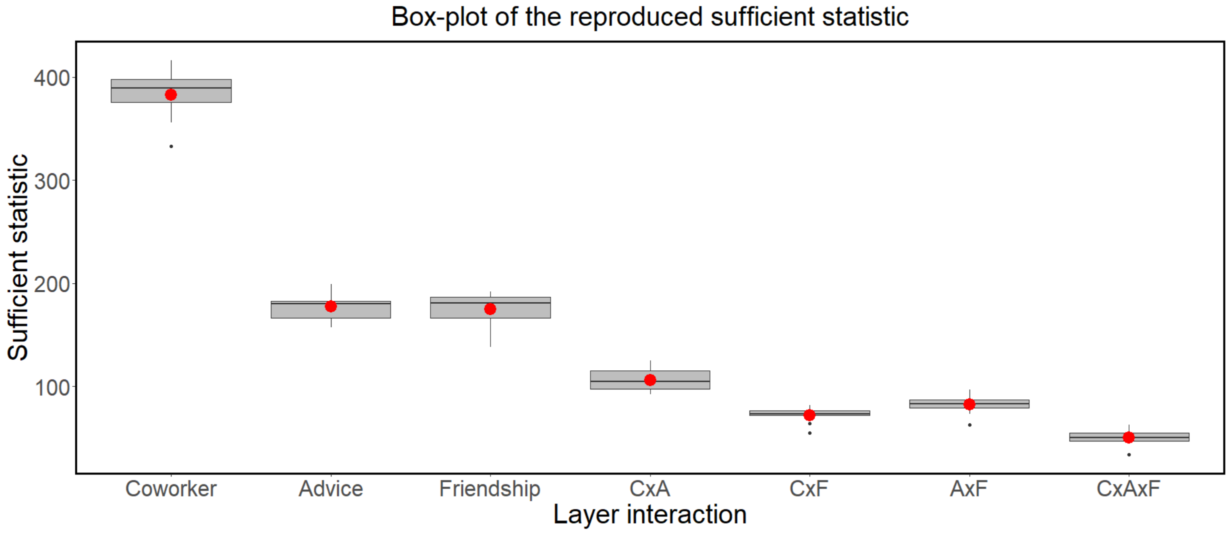

Figure 5: Box-plot of reproduced statistics from 10 replications for each dimension. Red dots are values of the observed sufficient statistics of the Lazega network.

to single-layer effects of the strong-coworker network, the advice network, and the friendship network, respectively, whereas $\theta_{1,2}$, $\theta_{1,3}$, and $\theta_{2,3}$ correspond to the layer interaction effects. We can calculate the conditional log-odds of each edge being present in the multilayer network given the rest of the network. For example, if lawyer i and lawyer j are observed to have an advice relationship and are friends at the same time, the odds of these two lawyers to have a strong-coworker relationship is given by

$$\frac{\mathbb{P}(X_{i,j}^{(C)}=1\,|\,\boldsymbol{X}_{i,j}^{(A)}=1,\ \boldsymbol{X}_{i,j}^{(F)}=1)}{\mathbb{P}(X_{i,j}^{(C)}=0\,|\,\boldsymbol{X}_{i,j}^{(A)}=1,\ \boldsymbol{X}_{i,j}^{(F)}=1)} \;=\; \exp\left(\theta_1+\theta_{1,2}\,x_{i,j}^{(A)}+\theta_{1,3}\,x_{i,j}^{(F)}\right)=1.767,$$

providing interpretation of the interaction and influence among the different layers.

Next, we use the estimated MPLE $\widetilde{\boldsymbol{\theta}}$ to simulate networks of the same size and calculate the sufficient statistics of the simulated networks. Comparisons of the sufficient statistics between the observed Lazega network and the simulated networks are provided in Figure 5. Such comparisons serve two key purposes. First, such comparisons are an established method of diagnosing model fit in the statistical network analysis literature [Hunter et al., 2008], and second, provide a check on the approximate solution to the score equation. Note that MPLEs are not guaranteed to reproduce (on average) observed values of sufficient statistics in exponential families—in contrast to MLEs. The relative ℓ_2-error of the sufficient statistics between the observed and the average of 10 simulated networks is 0.013, suggesting a successful re-construction of the observed network.

7 Discussion

In this work, we introduced a flexible class of statistical models for multilayer networks. Key to our approach lies in the integrative nature by which we establish our framework, extending arbitrary strictly positive probability distributions for single-layer networks to multilayer-network models through a network separable framework with Markov random field specifications. We established the foundations for statistical inference through consistency and multivariate normality results, the results of which have been demonstrated in simulation studies and in an application. The key assumption to our approach lies in the network separability assumption, which necessitates network dyads be conditionally independent given the basis network. This assumption may or may not be valid in practice, which would necessitate the development of generalizations of the framework we established in this work through the relaxation of the conditional independence assumption. Such relaxations would result in more complex dependence structures, requiring new and careful theoretical treatment in order to establish similar statistical foundations of models to the ones we have developed here, representing potential avenues for future research.

References

E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

J. A. Arroyo, A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. *The Journal of Machine Learning Research*, 22(1):6303–6351, 2021.

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, J. T. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, D. L. Sussman, E. Fishkind, and Y. Park. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18:1–92, 2018.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–225, 1974.

S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. *The Annals of Applied Probability*, 21:2146–2170, 2011.

P. Block. Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40: 163–173, 2015.

C. T. Butts. A dynamic process interpretation of the sparse ERGM reference model. *Journal of Mathematical Sociology*, 2020.

D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity. *Advances in Neural Information Processing Systems*, 29, 2016.

A. Caimo and I. Gollini. A multilayer exponential random graph modelling approach for weighted networks. *Computational Statistics & Data Analysis*, 142:106825, 2020.

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

S. Chen, S. Liu, and Z. Ma. Global and individualized community detection in inhomogeneous multilayer networks. *The Annals of Statistics*, 50(5):2664–2693, 2022.

H. Crane and W. Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018.

O. Frank. Transitivity in stochastic graphs and digraphs. *Journal of Mathematical Sociology*, 7:199–213, 1980.

M. Furi and M. Martelli. On the mean value theorem, inequality, and inclusion. *The American Mathematical Monthly*, 98(9):840–846, 1991.

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

S. M. Goodreau, J. A. Kitts, and M. Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103–125, 2009.

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

P. W. Holland and S. Leinhardt. Some evidence on the transitivity of positive interpersonal sentiment. *American Journal of Sociology*, 77:1205–1209, 1972.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76:33–65, 1981.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic block models: some first steps. *Social Networks*, 5:109–137, 1983.

S. Huang, H. Weng, and Y. Feng. Spectral clustering via adaptive layer aggregation for multi-layer networks. *Journal of Computational and Graphical Statistics*, pages 1–15, 2022.

D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15:565–583, 2006.

D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103:248–258, 2008.

R.A. Johnson and D.W. Wichern. *Applied multivariate statistical analysis.* Prentice hall, 2002.

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

P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31:204–213, 2009.

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

P. N. Krivitsky, L. M. Koehly, and C. S. Marcum. Exponential-family random graph models for multi-layer networks. *Psychometrika*, 85(3):630–659, 2020.

E. Lazega. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership.* Oxford University Press, 2001.

J. Lei, K. Chen, and B. Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 2020.

W. Li, Y. Xu, J Yang, and Z. Tang. Finding structural patterns in complex networks. In *2012 IEEE Fifth International Conference on Advanced Computational Intelligence*, pages 23–27, 2012.

D. Lusher, J. Koskinen, and G. Robins. *Exponential Random Graph Models for Social Networks.* Cambridge University Press, Cambridge, UK, 2013.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright. *Handbook of graphical models.* CRC Press, 2018.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* SIAM, 2000.

M. Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A): 2824–2853, 2019.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *The Annals of Statistics*, 39:1878–1915, 2011.

M. Schweinberger, P. N. Krivitsky, C. T. Butts, and Jonathan Stewart. Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios. *Statistical Science*, 35:627–662, 2020.

D. K. Sewell and Y. Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110:1646–1657, 2015.

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

J. Sosa and B. Betancourt. A latent space model for multilayer network data. *Computational Statistics & Data Analysis*, 169:107432, 2022.

J. Stewart, M. Schweinberger, M. Bojanowski, and M. Morris. Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. *Social Networks*, 59:98–119, 2019.

J. R. Stewart and M. Schweinberger. Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension. *arXiv preprint arXiv:2012.07167*, 2021.

D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85:204–212, 1990.

R. Sundberg. *Statistical modelling by exponential families*, volume 12. Cambridge University Press, 2019.

D. Sussman, M. Tang, and C. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:48–57, 2014.

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

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

M. Zhou and Y. Shao. A powerful test for multivariate normality. *Journal of Applied Statistics*, 41(2):351–363, 2014.

Supplement:

Learning cross-layer dependence structure in multilayer networks

By Jiaheng Li and Jonathan R. Stewart

Department of Statistics, Florida State University

A Proof of Proposition 1

We prove Proposition 1 from Section 2.

Proof of Proposition 1. For the first and second results, define the set

$$\mathcal{A}_+ \;\; := \;\; \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X} \times \mathbb{Y} \; : \; h(\boldsymbol{x}, \boldsymbol{y}) = 1\},$$

and the vector-valued map $\boldsymbol{\varphi} : \mathbb{X} \mapsto \mathbb{Y}$ by defining its components to be

$$\varphi_{i,j}(\boldsymbol{x}) \;=\; \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0), \qquad \{i,j\} \subset \mathcal{N},$$

populating the vector $\boldsymbol{\varphi}(\boldsymbol{x})$ in the lexicographical ordering of the dyad indices $\{i,j\} \subset \mathcal{N}$. By the definition of $h : \mathbb{X} \times \mathbb{Y} \mapsto \{0,1\}$ and $\boldsymbol{\varphi} : \mathbb{X} \mapsto \mathbb{Y}$, $\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{y}$ for each pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{A}_+$. Furthermore, the element $\boldsymbol{y}$ is unique for a given $\boldsymbol{x} \in \mathbb{X}$, because if there would exists some $\boldsymbol{y}' \in \mathbb{Y}$ such that $\boldsymbol{y} \neq \boldsymbol{y}'$ with the property that $\{(\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y}')\} \subseteq \mathcal{A}_+$, then there would exist a pair $\{i,j\} \subset \mathcal{N}$ such that $y_{i,j} = 1 - y'_{i,j}$, implying $\mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0) \neq y'_{i,j}$, in which case $h(\boldsymbol{x}, \boldsymbol{y}') = 0$, contradicting the assumption that $\{(\boldsymbol{x}, \boldsymbol{y})\} \in \mathcal{A}_+$. By (1), the functions f and g are assumed to be strictly positive in their respective domains. Hence, $(\mathbb{X} \times \mathbb{Y}) \setminus \mathcal{A}_+$ is the largest null set of $\mathbb{X} \times \mathbb{Y}$, i.e., $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{A}) = 0$ if and only if $\mathcal{A} \subseteq (\mathbb{X} \times \mathbb{Y}) \setminus \mathcal{A}_+$. Thus, the first and second results are established.

For the third result, note that g is assumed to be strictly positive on its domain $\mathbb{Y}$. Hence, $g(\boldsymbol{y}) = \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in \mathbb{Y}$ and $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ is therefore well-defined. By definition,

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y}) \;=\; \frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x},\, \boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})},$$

where $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})$ is the marginal probability of event $\boldsymbol{Y} = \boldsymbol{y}$ and is assumed to be equal to $g(\boldsymbol{y})$. The model form for $\mathbb{P}_{\boldsymbol{\theta}}$ given in (1) implies

$$\frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x},\, \boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})} \;=\; \frac{f(\boldsymbol{x}, \boldsymbol{\theta})\, g(\boldsymbol{y})\, \psi(\boldsymbol{\theta}, \boldsymbol{y})}{g(\boldsymbol{y})} \;=\; \exp(\log f(\boldsymbol{x}, \boldsymbol{\theta}) + \log \psi(\boldsymbol{\theta}, \boldsymbol{y})),$$

under the assumption that $h(\boldsymbol{x}, \boldsymbol{y}) = 1$. Hence,

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) \;=\; \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})$$

so that

$$\log\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) \;=\; \log\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}) + \log\, g(\boldsymbol{y}),$$

as $g(\boldsymbol{y})$ is the marginal probability mass function of $\boldsymbol{Y}$, i.e., $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) = g(\boldsymbol{y})$. Lemma 4 establishes that $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$ belongs to a minimal exponential family, completing the proof of the third and last result of the proposition.

■

B Proof of Lemma 1

We prove Lemma 1 from Section 2.

Proof of Lemma 1. We prove the result in the case of the log-likelihood function. The proof in the case of the log-pseudolikelihood function follows similarly, substituting the appropriate quantities relevant to the log-pseudolikelihood. Using (1),

$$\begin{aligned}
-\mathbb{E}\, \nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{\boldsymbol{y} \in \mathbb{Y}} \sum_{\boldsymbol{x} \in \mathbb{X}} -\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})\, g(\boldsymbol{y}) \\
&= \sum_{\boldsymbol{y} \in \mathbb{Y}} g(\boldsymbol{y}) \sum_{\boldsymbol{x} \in \mathbb{X}} -\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y}) \\
&= \sum_{\boldsymbol{y} \in \mathbb{Y}} g(\boldsymbol{y}) \sum_{\{i,j\} \subset \mathcal{N} \,:\, y_{i,j} = 1} I(\boldsymbol{\theta}) \\
&= I(\boldsymbol{\theta}) \sum_{\boldsymbol{y} \in \mathbb{Y}} g(\boldsymbol{y})\, \|\boldsymbol{y}\|_1 \\
&= I(\boldsymbol{\theta})\, \mathbb{E} \|\boldsymbol{Y}\|_1.
\end{aligned}$$

The above follows by exploiting the conditional independence of vectors $\boldsymbol{x}_{i,j}$ ($\{i, j\} \subset \mathcal{N}$) given $\boldsymbol{Y} = \boldsymbol{y}$ under (1), which implies

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) = \sum_{\{i,j\} \subset \mathcal{N}} \log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}_{i,j} = \boldsymbol{x}_{i,j} \,|\, \boldsymbol{Y} = \boldsymbol{y}),$$

and from the fact that the conditional probability distribution of $\boldsymbol{X}_{i,j}$ given $\boldsymbol{Y}$ is a degenerate point mass at $\boldsymbol{0}$ when $Y_{i,j} = 0$ so that $-\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ is a sum of $\|\boldsymbol{y}\|_1$ matrices each equal to $I(\boldsymbol{\theta})$, i.e., given $\boldsymbol{y} \in \mathbb{Y}$, we have

$$\sum_{\boldsymbol{x} \in \mathbb{X}} -\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$$

$$= \sum_{\{i,j\} \subset \mathcal{N}} \mathbb{E}\left[-\nabla_{\boldsymbol{\theta}}^2\, L_{i,j}(\boldsymbol{\theta}, \boldsymbol{X}_{i,j}, \boldsymbol{Y}) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \;=\; \sum_{\{i,j\} \subset \mathcal{N} \,:\, y_{i,j} = 1} I(\boldsymbol{\theta}).$$

The fact that $I(\boldsymbol{\theta})$ is constant for all pairs $\{i, j\} \subset \mathcal{N}$ satisfying $Y_{i,j} = 1$ follows from the form of (1), which assumes each vector $\boldsymbol{X}_{i,j}$ ($\{i, j\} \subset \mathcal{N}$) is conditionally independent and identically distributed, conditional on $\boldsymbol{Y}$. Hence,

$$\mathbb{E}\left[-\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})\right] \;=\; I(\boldsymbol{\theta})\, \mathbb{E}\, \|\boldsymbol{Y}\|_1,$$

which in turn implies $\lambda_{\min}(-\mathbb{E}\nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})) = \lambda_{\min}(I(\boldsymbol{\theta}))\, \mathbb{E}\, \|\boldsymbol{Y}\|_1$.

■

C Concentration inequalities for multilayer networks

We establish concentration inequalities of gradients of log-likelihoods and log-pseudolikelihoods functions of network separable multilayer networks in Lemma 2 and Lemma 3, respectively. Recall the definition $[D_g]^+ := \max\{0,\, D_g\}$, where

$$D_g \;:=\; \sum_{\{i,j\} \prec \{v,w\} \subset \mathcal{N}} \mathbb{C}(Y_{i,j},\, Y_{v,w}),$$

with $\{i, j\} \prec \{v, w\}$ implying the sum is taken with respect to the lexicographical ordering of pairs of nodes, and where $g : \mathbb{Y} \mapsto (0, 1)$ is the marginal probability mass function of $\boldsymbol{Y}$.

Lemma 2 *Consider multilayer networks satisfying (1) which are defined on a set of $N \geq 3$ nodes and $K \geq 1$ layers. Define $\gamma_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) := -\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$, where $\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ is the log-likelihood function. Then, for all $t > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^p$,*

$$\mathbb{P}\left(\|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty \geq t\right) \;\leq\; 2\, \exp\left(-\frac{t^2}{\mathbb{E}\, \|\boldsymbol{Y}\|_1 + [D_g]^+} + \log\, p\right) + \frac{1}{\mathbb{E}\|\boldsymbol{Y}\|_1}.$$

Proof of Lemma 2. By Proposition 1,

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;=\; \log\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y}) + \log\, g(\boldsymbol{y}).$$

Thus,

$$\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;=\; \nabla_{\boldsymbol{\theta}} \log\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y}) + \nabla_{\boldsymbol{\theta}} \log\, g(\boldsymbol{y}) \;=\; s(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X}), \tag{6}$$

as $g(\boldsymbol{y}) = \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})$ is assumed to not be a function of $\boldsymbol{\theta}$. The last equality in (6) follows from Lemma 4, which showed that $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ is a minimal exponential family with sufficient statistic vector $\boldsymbol{s}(\boldsymbol{x})$ defined in Lemma 4 and natural parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, inserting the familiar form of the score equation of an exponential family with respect to the natural parameter vector [e.g., Proposition 3.10, p. 32, Sundberg, 2019]. Thus,

$$-(\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})) \;=\; s(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X}) - \mathbb{E}\left[s(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X})\right] \;=\; s(\boldsymbol{X}) - \mathbb{E}\, s(\boldsymbol{X}).$$

Let $t > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^p$ be arbitrary and fixed and define $\mathcal{D}_\infty(\boldsymbol{\theta}, t)$ to be the event that $\|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty = \|\boldsymbol{s}(\boldsymbol{X}) - \mathbb{E}\, \boldsymbol{s}(\boldsymbol{X})\|_\infty \geq t$. By a union bound,

$$\begin{aligned}\mathbb{P}(\|\boldsymbol{s}(\boldsymbol{X}) - \mathbb{E}\, \boldsymbol{s}(\boldsymbol{X})\|_\infty \geq t) \;&=\; \mathbb{P}\left(\bigcup_{l=1}^{p} \Big[|s_l(\boldsymbol{X}) - \mathbb{E}\, s_l(\boldsymbol{X})| \geq t\Big]\right) \\ &\leq\; \sum_{l=1}^{p} \mathbb{P}\left(|s_l(\boldsymbol{X}) - \mathbb{E}\, s_l(\boldsymbol{X})| \geq t\right).\end{aligned}$$

For each $l \in \{1, \ldots, p\}$, define $\mathcal{D}_l(\boldsymbol{\theta}, t)$ to be the event $|s_l(\boldsymbol{X}) - \mathbb{E}\, s_l(\boldsymbol{X})| \geq t$. Let $\epsilon > 0$ and define $\mathcal{E}(\epsilon)$ to be the event that $|\|\boldsymbol{Y}\|_1 - \mathbb{E}\|\boldsymbol{Y}\|_1| \leq \epsilon$, i.e.,

$$\mathcal{E}(\epsilon) \;=\; \{\boldsymbol{y} \in \mathbb{Y} \,:\, |\|\boldsymbol{y}\|_1 - \mathbb{E}\|\boldsymbol{Y}\|_1| \leq \epsilon\}.$$

We assume that $\epsilon > 0$ is chosen so that $\mathcal{E}(\epsilon)$ is not empty, which implies $\mathbb{P}(\mathcal{E}(\epsilon)) > 0$ as $g(\boldsymbol{y})$ is assumed to be strictly positive on $\mathbb{Y}$. By the law of total probability,

$$\begin{aligned}\mathbb{P}\left(\mathcal{D}_\infty(\boldsymbol{\theta}, t)\right) \;&=\; \mathbb{P}\left(\mathcal{D}_\infty(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right) \mathbb{P}\left(\mathcal{E}(\epsilon)\right) + \mathbb{P}\left(\mathcal{D}_\infty(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)^c\right) \mathbb{P}\left(\mathcal{E}(\epsilon)^c\right) \\ &\leq\; \mathbb{P}\left(\mathcal{D}_\infty(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right) + \mathbb{P}\left(\mathcal{E}(\epsilon)^c\right) \\ &\leq\; \sum_{l=1}^{p} \mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right) + \mathbb{P}\left(\mathcal{E}(\epsilon)^c\right).\end{aligned} \tag{7}$$

Note that we have not necessarily guaranteed that $\mathbb{P}\left(\mathcal{E}(\epsilon)^c\right) > 0$. However, if $\mathbb{P}\left(\mathcal{E}(\epsilon)^c\right) = 0$ the non-conditional form of the law of total probability would yield the bound

$$\mathbb{P}\left(\mathcal{D}_\infty(\boldsymbol{\theta}, t)\right) \;\leq\; \sum_{l=1}^{p} \mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right),$$

which is strictly sharper than the bound we give in (7). We will use a divide and conquer strategy to bound each probability in (7) in turn. The form of (1) implies, through factorization principles, that the dyad-based vectors $\boldsymbol{X}_{i,j}$ ($\{i,j\} \subset \mathcal{N}$) are conditionally independent given $\boldsymbol{Y}$ [e.g., Maathuis et al., 2018, p. 11–13]. Hence, using Lemma 4, the components of the sufficient statistic vector decompose into the sum

$$s_l(\boldsymbol{X}) \;=\; \sum_{\{i,j\}\subset\mathcal{N}} s_{l,i,j}(\boldsymbol{X}_{i,j}), \qquad l \in \{1, \ldots, p\},$$

so that the components of $\boldsymbol{s}(\boldsymbol{X})$ are sums of bounded conditionally independent random variables given $\boldsymbol{Y}$. Using the forms for $s_l(\boldsymbol{X})$ and $s_{l,i,j}(\boldsymbol{X}_{i,j})$ outlined in Lemma 4, we have $0 \leq s_{l,i,j}(\boldsymbol{X}_{i,j}) \leq Y_{i,j}$ $\mathbb{P}$-almost surely, because $s_{l,i,j}(\boldsymbol{X}_{i,j}) \in \{0, 1\}$ and $s_{l,i,j}(\boldsymbol{X}_{i,j}) = 0$ if $Y_{i,j} = 0$ $\mathbb{P}$-almost surely. We may then apply Hoeffding's inequality to obtain

$$\mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \;\leq\; 2 \exp\left(-\frac{2\, t^2}{\|\boldsymbol{y}\|_1}\right), \tag{8}$$

where the denominator follows because $\sum_{\{i,j\}\subset\mathcal{N}} y_{i,j}^2 = \|\boldsymbol{y}\|_1$. Using the law of total probability, we bound $\mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right)$ as follows:

$$\begin{aligned} \mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right) &= \sum_{\boldsymbol{y}\in\mathbb{Y}} \mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \cap [\boldsymbol{Y} = \boldsymbol{y}] \,|\, \mathcal{E}(\epsilon)\right) \\ &= \sum_{\boldsymbol{y}\in\mathcal{E}(\epsilon)} \mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \cap [\boldsymbol{Y} = \boldsymbol{y}] \,|\, \mathcal{E}(\epsilon)\right) \\ &= \sum_{\boldsymbol{y}\in\mathcal{E}(\epsilon)} \mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, [\boldsymbol{Y} = \boldsymbol{y}] \cap \mathcal{E}(\epsilon)\right) \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \,|\, \mathcal{E}(\epsilon)) \\ &= \sum_{\boldsymbol{y}\in\mathcal{E}(\epsilon)} \mathbb{P}(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \boldsymbol{Y} = \boldsymbol{y})\, \frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\mathcal{E}(\epsilon))}, \end{aligned} \tag{9}$$

noting that $[\boldsymbol{Y} = \boldsymbol{y}] \cap \mathcal{E}(\epsilon) = [\boldsymbol{Y} = \boldsymbol{y}]$ whenever $\boldsymbol{y} \in \mathcal{E}(\epsilon)$ and in the case when $\boldsymbol{y} \notin \mathcal{E}(\epsilon)$, the intersection is empty, implying

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \,|\, \mathcal{E}(\epsilon)) \;=\; \frac{\mathbb{P}([\boldsymbol{Y} = \boldsymbol{y}] \cap \mathcal{E}(\epsilon))}{\mathbb{P}(\mathcal{E}(\epsilon))} \;=\; \begin{cases} \dfrac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\mathcal{E}(\epsilon))} & \boldsymbol{y} \in \mathcal{E}(\epsilon) \\ 0 & \boldsymbol{y} \notin \mathcal{E}(\epsilon) \end{cases}.$$

We now bound (9) using the bound in (8):

$$\begin{aligned} \sum_{\boldsymbol{y} \in \mathcal{E}(\epsilon)} \mathbb{P}(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \boldsymbol{Y} = \boldsymbol{y}) \, \frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\mathcal{E}(\epsilon))} \;&\leq\; \sum_{\boldsymbol{y} \in \mathcal{E}(\epsilon)} 2 \, \exp\left(-\frac{2\,t^2}{\|\boldsymbol{y}\|_1}\right) \frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\mathcal{E}(\epsilon))} \\ &\leq\; 2 \, \exp\left(-\frac{2\,t^2}{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + \epsilon}\right) \sum_{\boldsymbol{y} \in \mathcal{E}(\epsilon)} \frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\mathcal{E}(\epsilon))} \\ &=\; 2 \, \exp\left(-\frac{2\,t^2}{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + \epsilon}\right), \end{aligned}$$

showing

$$\mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right) \;\leq\; 2 \, \exp\left(-\frac{2\,t^2}{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + \epsilon}\right).$$

The replacement of $\|\boldsymbol{y}\|_1$ by $\mathbb{E}\|\boldsymbol{Y}\|_1 + \epsilon$ follows because $\|\boldsymbol{y}\|_1 \leq \mathbb{E}\|\boldsymbol{Y}\|_1 + \epsilon$ for $\boldsymbol{y} \in \mathcal{E}(\epsilon)$, resulting in the upper bound above. We bound the second term in the inequality (7) using Chebyshev's inequality:

$$\mathbb{P}(\mathcal{E}(\epsilon)^c) \;=\; \mathbb{P}(|\|\boldsymbol{Y}\|_1 - \mathbb{E}\,\|\boldsymbol{Y}\|_1| > \epsilon) \;\leq\; \mathbb{P}(|\|\boldsymbol{Y}\|_1 - \mathbb{E}\,\|\boldsymbol{Y}\|_1| \geq \epsilon) \;\leq\; \frac{\mathbb{V}(\|\boldsymbol{Y}\|_1)}{\epsilon^2}.$$

We bound the variance $\mathbb{V}(\|\boldsymbol{Y}\|_1)$ as follows:

$$\begin{aligned} \mathbb{V}(\|\boldsymbol{Y}\|_1) \;&=\; \sum_{\{i,j\} \subset \mathcal{N}} \mathbb{V}\, Y_{i,j} + 2 \sum_{\{i,j\} \prec \{v,w\} \subset \mathcal{N}} \mathbb{C}(Y_{i,j},\, Y_{v,w}) \\ &\leq\; \mathbb{E}\,\|\boldsymbol{Y}\|_1 + 2 \sum_{\{i,j\} \prec \{v,w\} \subset \mathcal{N}} \mathbb{C}(Y_{i,j},\, Y_{v,w}), \end{aligned}$$

noting $Y_{i,j} \in \{0, 1\}$ so that $\mathbb{V}\, Y_{i,j} = \mathbb{P}(Y_{i,j} = 1)\, \mathbb{P}(Y_{i,j} = 0) \leq \mathbb{E}\, Y_{i,j}$. Hence,

$$\mathbb{P}(\mathcal{E}(\epsilon)^c) \;\leq\; \frac{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + 2 \sum_{\{i,j\} \prec \{v,w\} \subset \mathcal{N}} \mathbb{C}(Y_{i,j},\, Y_{v,w})}{\epsilon^2} \;=\; \frac{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + 2\;[D_g]^+}{\epsilon^2}. \tag{10}$$

Taking $\epsilon = \mathbb{E}\,\|\boldsymbol{Y}\|_1 + 2\;[D_g]^+ > 0$ shows that $\mathbb{P}(\mathcal{E}(\epsilon)^c) \leq (\mathbb{E}\,\|\boldsymbol{Y}\|_1)^{-1}$ and

$$\mathbb{P}\left(\mathcal{D}_l(\boldsymbol{\theta}, t) \,|\, \mathcal{E}(\epsilon)\right) \;\leq\; 2 \, \exp\left(-\frac{2\,t^2}{2(\mathbb{E}\,\|\boldsymbol{Y}\|_1 + [D_g]^+)}\right) \;=\; 2 \, \exp\left(-\frac{t^2}{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + [D_g]^+}\right).$$

Combining all results shows that

$$\mathbb{P}\left(\|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty \geq t\right) \;\leq\; 2\, \exp\left(-\frac{t^2}{\mathbb{E}\,\|\boldsymbol{Y}\|_1 + [D_g]^+} + \log\, p\right) + \frac{1}{\mathbb{E}\|\boldsymbol{Y}\|_1}.$$

As a final matter, note that this choice of $\epsilon > 0$ ensures $\mathcal{E}(\epsilon)$ contains all $\boldsymbol{y} \in \mathbb{Y}$ with $\|\boldsymbol{y}\|_1 \in [0,\, 2(\mathbb{E}\,\|\boldsymbol{Y}\| + [D_g]^+)]$ as the empty graph is an element of $\mathbb{Y}$ with 0 edges.

■

We next prove a related result for gradients of log-pseudolikelihood functions of network separable multilayer networks in Lemma 3. The proof of Lemma 3 essentially follows the same proof of Lemma 2, and as a result we do not repeat key arguments, instead opting to only outline the changes in the proof.

Lemma 3 *Consider multilayer networks satisfying* (1) *which are defined on a set of* $N \geq 3$ *nodes and* $K \geq 1$ *layers. Define* $\widetilde{\gamma}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq -\nabla_{\boldsymbol{\theta}}\, \widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$, *where* $\widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ *is the log-pseudolikelihood function. Then, for all* $t > 0$ *and* $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$\mathbb{P}\left(\|\widetilde{\gamma}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \widetilde{\gamma}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty \geq t\right) \;\;\leq\;\; 2\, \exp\left(-\frac{t^2}{K^2\left(\mathbb{E}\,\|\boldsymbol{Y}\|_1 + [D_g]^+\right)} + \log p\right) + \frac{1}{\mathbb{E}\|\boldsymbol{Y}\|_1}.$$

Proof of Lemma 3. From (4), the log-pseudolikelihood function is given by

$$\widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) = \sum_{\{i,j\} \subseteq \mathcal{N}} \sum_{k=1}^{K} \log\, \mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}).$$

By Lemma 5, $\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y})$ is an exponential family with sufficient statistic vector $s : \mathbb{X} \mapsto \mathbb{R}^p$ defined in Lemma 4 and natural parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$. Hence,

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}\, \widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;&=\; \sum_{\{i,j\} \subseteq \mathcal{N}} \sum_{k=1}^{K} \nabla_{\boldsymbol{\theta}} \log\, \mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}) \\ &=\; \sum_{\{i,j\} \subseteq \mathcal{N}} \sum_{k=1}^{K} \left[\, \boldsymbol{s}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}} \left[\boldsymbol{s}(\boldsymbol{X}) \,|\, \boldsymbol{X}_{i,j}^{-(k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}\right] \right], \end{aligned}$$

where $\boldsymbol{X}_{i,j}^{-(k)}$ denotes the $(K\text{-}1)$-dimensional vector of edge variables of dyad $\{i, j\}$ in $\boldsymbol{X}_{i,j}$ which excludes the single edge variable $X_{i,j}^{(k)}$, and by inserting the familiar form of the score equation of an exponential family with respect to the natural parameter vector [e.g., Proposition 3.10, p. 32, Sundberg, 2019]. Note that $\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y})$ may not belong to a minimal exponential family. This presents no issues as we do not require the conditional probability distributions of individual edge variables belong to a minimal exponential family. Under the assumption that $(\boldsymbol{X}, \boldsymbol{Y})$ follow (1), the vectors $\boldsymbol{X}_{i,j}$ $(\{i,j\} \subset \mathcal{N})$ are conditionally independent given $\boldsymbol{Y}$ (as discussed in the proof of Lemma 2). Therefore, the l^{th} component $s_l(\boldsymbol{X})$ decomposes into the sum of conditionally independent Bernoulli random variables:

$$s_l(\boldsymbol{X}) \;=\; \sum_{\{i,j\} \subset \mathcal{N}} s_{l,i,j}(\boldsymbol{X}_{i,j}), \qquad l \in \{1, \dots, p\},$$

so that the components of $\boldsymbol{s}(\boldsymbol{X})$ are sums of bounded conditionally independent random variables given $\boldsymbol{Y}$. Thus,

$$\nabla_{\boldsymbol{\theta}}\, \widetilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;=\; \sum_{\{i,j\} \subseteq \mathcal{N}} \sum_{k=1}^{K} \left(s_{l,i,j}(\boldsymbol{X}_{i,j}) - E_{l,i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y})\right),$$

where

$$E_{l,i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y}) \;\coloneqq\; \mathbb{E}_{\boldsymbol{\theta}}\left[s_{l,i,j}(\boldsymbol{X}_{i,j}) \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}\right].$$

Using the form of $s_l(\boldsymbol{X})$ and $s_{l,i,j}(\boldsymbol{X}_{i,j})$ outlined in Lemma 4, we have $0 \leq s_{l,i,j}(\boldsymbol{X}_{i,j}) \leq Y_{i,j}$ $\mathbb{P}$-almost surely, because $s_{l,i,j}(\boldsymbol{X}_{i,j}) \in \{0, 1\}$ and $s_{l,i,j}(\boldsymbol{X}_{i,j}) = 0$ if $Y_{i,j} = 0$ $\mathbb{P}$-almost surely. This also implies $0 \leq E_{l,i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y}) \leq Y_{i,j}$ $\mathbb{P}$-almost surely. Taken together,

$$0 \;\leq\; \left|\sum_{k=1}^{K} \left(s_{l,i,j}(\boldsymbol{X}_{i,j}) - E_{l,i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y})\right)\right| \;\leq\; \sum_{k=1}^{K} \left|s_{l,i,j}(\boldsymbol{X}_{i,j}) - E_{l,i,j}(\boldsymbol{\theta}, \boldsymbol{x}_{i,j}, \boldsymbol{y})\right| \;\leq\; K\, Y_{i,j},$$

$\mathbb{P}$-almost surely. From here, the remainder of the proof follows the proof of Lemma 2, with the sole exception using the bound $K\, Y_{i,j}$ in the application of Hoeffding's inequality. Reiterating the proof of Lemma 2 with this change will yield

$$\mathbb{P}\left(\|\widetilde{\gamma}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\,\widetilde{\gamma}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_{\infty} \geq t\right) \;\leq\; 2\,\exp\left(-\frac{t^2}{K^2\left(\mathbb{E}\,\|\boldsymbol{Y}\|_1 + [D_g]^{+}\right)} + \log p\right) + \frac{1}{\mathbb{E}\|\boldsymbol{Y}\|_1}.$$

■

C.1 Auxiliary results

Lemma 4 *Consider multilayer networks satisfying* (1) *with maximum interaction term* $H \leq K$ *and defined on a set of* $N \geq 3$ *nodes and* $K \geq 1$ *layers. Then the following hold:*

1. *The conditional probability mass function of* $\boldsymbol{X}$ *given* $\boldsymbol{Y}$ *is an exponential family:*

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}) \;\; \propto \;\; h(\boldsymbol{x},\, \boldsymbol{y}) \; \exp\left(\langle \boldsymbol{\theta},\, \boldsymbol{s}(\boldsymbol{x})\rangle\right),$$

where

$$h(\boldsymbol{x},\, \boldsymbol{y}) \;\; = \;\; \prod_{\{i,j\} \subset \mathcal{N}} \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0)^{y_{i,j}} \; \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 = 0)^{1-y_{i,j}},$$

sufficient statistic vector $s : \mathbb{X} \mapsto \mathbb{R}^p$ *and natural parameter vector* $\boldsymbol{\theta} \in \mathbb{R}^p$*.*

2. *For each* $l \in \{1, \ldots, p\}$*, there exists* $h \in \{1, \ldots, H\}$ *and* $\{k_1, \ldots, k_h\} \subseteq \{1, \ldots, K\}$ *such that the* l^{th} *component of the sufficient statistic vector* $s(\boldsymbol{x})$ *can be written as*

$$s_l(\boldsymbol{x}) \;\; = \sum_{\{i,j\} \subset \mathcal{N}} s_{l,i,j}(\boldsymbol{x}) \;\; = \;\; \prod_{r=1}^{h} x_{i,j}^{(k_r)}. \tag{11}$$

3. *The exponential family outlined above is both minimal, full, and regular.*

Proof of Lemma 4. First, the form of the conditional probability distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$ derived in Proposition 1 is given by

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y}) \;\; = \;\; \exp\left(\log\, f(\boldsymbol{x}, \boldsymbol{\theta}) + \log\, \psi(\boldsymbol{\theta}, \boldsymbol{y})\right), \tag{12}$$

provided $h(\boldsymbol{x}, \boldsymbol{y}) = 1$. The form of (1) suggests that (12) will be a minimal exponential family in canonical form due to the form of the Markov random field specification for $f(\boldsymbol{\theta}, \boldsymbol{x})$ and the definition of $\psi(\boldsymbol{\theta}, \boldsymbol{y})$. From the form of $f(\boldsymbol{x}, \boldsymbol{\theta})$ in (1),

$$\log f(\boldsymbol{x}, \boldsymbol{\theta}) \;=\; \sum_{\{i,j\} \subset \mathcal{N}} \left(\sum_{k=1}^{K} \theta_k x_{i,j}^{(k)} + \sum_{k<l}^{K} \theta_{k,l} x_{i,j}^{(k)} x_{i,j}^{(l)} + \ldots + \sum_{k_1 < \ldots < k_H}^{K} \theta_{k_1,k_2,\ldots,k_H} x_{i,j}^{(k_1)} \cdots x_{i,j}^{(k_H)} \right),$$

where $H \leq K$ is the highest order of cross-layer interactions included in the model. We write $\theta_{k_1,k_2,\ldots,k_h}$ to reference the h-order interaction parameter for the interaction term among layers $\{k_1, \ldots, k_h\} \subseteq \{1, \ldots, K\}$. As specified, $\psi(\boldsymbol{\theta}, \boldsymbol{y})$ is the normalizing constant for the exponential family. As such, the natural parameter space of the exponential family is $\mathbb{R}^p$ as the support of $\mathbb{X}$ is finite, which implies $\psi(\boldsymbol{\theta}, \boldsymbol{y}) < \infty$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{Y}$. We establish minimality by noting that the components of the parameter vector $\boldsymbol{\theta}$ satisfy no linear or affine constraints. Attached to each parameter $\theta_{k_1,\ldots,k_h}$ ($\{k_1, \ldots, k_h\} \subset \{1, \ldots K\}$, $h \in \{1, \ldots, H\}$) is the sufficient statistic

$$s_{k_1,\ldots,k_h}(\boldsymbol{x}) \;=\; \sum_{\{i,j\} \subset \mathcal{N}} x_{i,j}^{(k_1)} \cdots x_{i,j}^{(k_h)}.$$

Each statistic $s_{k_1,\ldots,k_h}$ is a function of distinct, non-degenerate random variables, provided $\|\boldsymbol{y}\|_1 > 0$, and so none of the statistics $s_{k_1,\ldots,k_h}$ satisfy any linear or affine constraints. Hence, (1) specifies a minimal and full exponential family with natural parameter space $\mathbb{R}^p$ of dimension $p = \sum_{h=1}^{H} \binom{K}{h}$ and sufficient statistic vector $s(\boldsymbol{x})$ with components $s_{k_1,\ldots,k_h}(\boldsymbol{x})$ ($\{k_1, \ldots, k_h\} \subseteq \{1, \ldots, K\}, h = 1, \ldots, H$). Regularity follows trivially [e.g., Proposition 3.7, pp. 28, Sundberg, 2019]. The form of (11) outlines this for a linear indexing of the components of the sufficient statistic vector.

■

Lemma 5 *Consider multilayer networks satisfying* (1) *with maximum interaction term* $H \leq K$ *and defined on a set of* $N \geq 3$ *nodes and* $K \geq 1$ *layers. Then the conditional probability mass function of* $X_{i,j}^{(k)}$ *given* $\boldsymbol{Y} = \boldsymbol{y}$ *and* $\boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}$ *is an exponential family*

$$\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}) \;\propto\; h(\boldsymbol{x},\, \boldsymbol{y})\, \exp\left(\langle \boldsymbol{\theta},\, \boldsymbol{s}(\boldsymbol{x})\rangle\right),$$

with sufficient statistic vector $\boldsymbol{s} : \mathbb{X} \mapsto \mathbb{R}^p$ *defined in Lemma 4, natural parameter vector* $\boldsymbol{\theta} \in \mathbb{R}^p$*, and*

$$h(\boldsymbol{x},\, \boldsymbol{y}) \;=\; \prod_{\{i,j\} \subset \mathcal{N}} \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 > 0)^{y_{i,j}}\, \mathbb{1}(\|\boldsymbol{x}_{i,j}\|_1 = 0)^{1-y_{i,j}}.$$

Proof of Lemma 5. First, note that the form of (1) and Proposition 1 suggests that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}(X_{i,j}^{(k)} = x_{i,j}^{(k)} \,|\, \boldsymbol{X}_{i,j}^{(-k)} = \boldsymbol{x}_{i,j}^{(-k)}, \boldsymbol{Y} = \boldsymbol{y}) \\
&\quad = \frac{h(x_{i,j}^{(k)},\, \boldsymbol{x}_{i,j}^{(-k)},\, \boldsymbol{y})\ \exp\left(\langle \boldsymbol{\theta},\, \boldsymbol{s}(x_{i,j}^{(k)},\, \boldsymbol{x}_{i,j}^{(-k)})\rangle\right)}{\displaystyle\sum_{x_{i,j}^{(k)} \in \{0,1\}} h(x_{i,j}^{(k)},\, \boldsymbol{x}_{i,j}^{(-k)},\, \boldsymbol{y})\ \exp\left(\langle \boldsymbol{\theta},\, \boldsymbol{s}(x_{i,j}^{(k)},\, \boldsymbol{x}_{i,j}^{(-k)})\rangle\right)}
\end{aligned} \tag{13}$$

is an exponential family in canonical form using the Markov random field specification for $f(\boldsymbol{\theta}, \boldsymbol{x})$ and the form of the conditional probability distribution of $X_{i,j}^{(k)}$ given $\boldsymbol{Y}$ and $\boldsymbol{X}_{i,j}^{(-k)}$. However, this exponential family may not be full rank due to possible 0 values of components of the given (K-1)-dimensional vector $\boldsymbol{x}_{i,j}^{(-k)}$ and thus may not be minimal.

■

D Proof of Theorem 1

Proof of Theorem 1. We first prove the theorem for maximum likelihood estimators, and then discuss extensions and changes necessary to prove the result for maximum pseudolikelihood estimators. By Proposition 1, observing $\boldsymbol{X} = \boldsymbol{x}$ implies we observe $\boldsymbol{Y} = \boldsymbol{y}$, as for each given $\boldsymbol{x} \in \mathbb{X}$, $\boldsymbol{Y} = \boldsymbol{y}$ ($\mathbb{P}$-a.s.) for one and only one $\boldsymbol{y} \in \mathbb{Y}$ given by

$$y_{i,j} \;=\; \mathbb{1}\left(\|\boldsymbol{x}_{i,j}\|_1 \,>\, 0\right), \qquad \{i,j\} \subset \mathcal{N}.$$

Denote the gradient of $-\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ by

$$\gamma_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) \;\coloneqq\; -\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$$

and the expected Hessian matrix of the negative log-likelihood by

$$\boldsymbol{H}(\boldsymbol{\theta}) \;\coloneqq\; -\mathbb{E}\, \nabla_{\boldsymbol{\theta}}^2\, \ell(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y}).$$

Theorem 6.3.4 of Ortega and Rheinboldt [2000] states that if

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^{\star})^{\top}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) \;\geq\; 0$$

for all $\boldsymbol{\theta} \in \partial\, \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$, where $\partial\, \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ is the boundary of the set

$$\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon) \;\; = \;\; \{\boldsymbol{\theta} \in \mathbb{R}^p \,:\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 < \epsilon\},$$

then $\gamma_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$ has a root in $\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon) \cup \partial\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$, i.e., $\widehat{\boldsymbol{\theta}}$ exists and satisfies $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \leq \epsilon$. Note that a root of $\gamma_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$ is also a root of $-\gamma_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$; in what follows, we consider finding a maximizer of $\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ by finding a minimizer of $-\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$. The classification of roots as maximizers/minimizers is justified from the fact that that $\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ is concave in $\boldsymbol{\theta}$, a fact which follows from Proposition 1, as $g(\boldsymbol{y})$ is constant in $\boldsymbol{\theta}$ and $\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ is the log-likelihood of a minimal, full, and regular exponential family with natural parameter vector $\boldsymbol{\theta}$ and thus is strictly concave in $\boldsymbol{\theta}$ [Proposition 3.10, p. 32, Sundberg, 2019]. By the multivariate mean-value theorem [Furi and Martelli, 1991, Theorem 5],

$$\begin{aligned}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) \;\; &= \;\; (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \mathbb{E}\, \gamma_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{Y}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{H}(\dot{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\ &= \;\; (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{H}(\dot{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^\star),\end{aligned}$$

where $\dot{\boldsymbol{\theta}} = t\, \boldsymbol{\theta} + (1 - t)\, \boldsymbol{\theta}^\star$ (some $t \in [0, 1]$) and by invoking Lemma 2 of Stewart and Schweinberger [2021], which shows that both the expected log-likelihood and log-pseudolikelihood of a minimal exponential family is uniquely maximized at the data-generating parameter vector $\boldsymbol{\theta}^\star$, implying $\mathbb{E}\, \gamma_{\boldsymbol{\theta}^\star}(\boldsymbol{X}, \boldsymbol{Y}) = 0$. Let $\epsilon \in (0, \epsilon^\star)$ and arbitrarily take $\boldsymbol{\theta} \in \partial\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$. Then

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{H}(\dot{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{H}(\dot{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)}{(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \;\geq\; \epsilon^2\, \lambda_{\min}(\boldsymbol{H}(\dot{\boldsymbol{\theta}})),$$

since $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 = \epsilon$ as $\boldsymbol{\theta} \in \partial\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ and because the Rayleigh quotient of a matrix is bounded below by the smallest eigenvalue of that matrix so that

$$\frac{(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \boldsymbol{H}(\dot{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)}{(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)} \;\; \geq \;\; \lambda_{\min}(\boldsymbol{H}(\dot{\boldsymbol{\theta}})) \;\; \geq \;\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta})),$$

where $\lambda_{\min}(\boldsymbol{H}(\dot{\boldsymbol{\theta}}))$ is the smallest eigenvalue of $\boldsymbol{H}(\dot{\boldsymbol{\theta}})$, noting that

$$\|\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;\; = \;\; \|t\, \boldsymbol{\theta} + (1 - t)\, \boldsymbol{\theta}^\star - \boldsymbol{\theta}^\star\|_2 \;\; = \;\; t\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 \;\; \leq \;\; \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 \;\; \leq \;\; \epsilon^\star,$$

since $t \in [0,1]$. Lemma 1 showed that

$$\lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta})) \;=\; \lambda_{\min}(I(\boldsymbol{\theta}))\, \mathbb{E}\, \|\boldsymbol{Y}\|_1,$$

which in turn implies

$$\inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(\boldsymbol{H}(\boldsymbol{\theta})) \;=\; \xi_{\epsilon^\star}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1,$$

where

$$\xi_{\epsilon^\star} \;\coloneqq\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(I(\boldsymbol{\theta})),$$

with $I(\boldsymbol{\theta})$ defined in Lemma 1. Hence, for $\boldsymbol{\theta} \in \partial\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$ $(\epsilon \in (0, \epsilon^\star))$,

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) \;\geq\; \epsilon^2\, \xi_{\epsilon^\star}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1.$$

We next turn to showing

$$\mathbb{P}\left(\inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) \;\geq\; 0 \right) \;\geq\; 1 - 2\, (\mathbb{E}\|\boldsymbol{Y}\|_1)^{-1},$$

by showing that the event

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)} |(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top (\mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}))| \;<\; \epsilon^2\, \xi_{\epsilon^\star}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1$$

occurs with probability at least $1 - 2\, (\mathbb{E}\|\boldsymbol{Y}\|_1)^{-1}$, in turn implying that the event $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \leq \epsilon$ happens with probability at least $1 - 2\, (\mathbb{E}\|\boldsymbol{Y}\|_1)^{-1}$. Applying the Cauchy-Schwarz inequality and utilizing standard vector norm inequalities,

$$\begin{aligned} |(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top (\mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}))| &\leq \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2\, \|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_2 \\ &\leq \epsilon\, \sqrt{p}\, \|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty, \end{aligned}$$

noting $\boldsymbol{\theta} \in \partial\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$. It suffices to demonstrate, for all $\boldsymbol{\theta} \in \partial\mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon)$, that

$$\mathbb{P}\left(\|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty \;<\; \epsilon\, p^{-\frac{1}{2}}\, \xi_{\epsilon^\star}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1 \right) \;\geq\; 1 - 2\, (\mathbb{E}\|\boldsymbol{Y}\|_1)^{-1}.$$

For ease of presentation, we define $\mathcal{D}_{N,\epsilon,p}$ to be the event

$$\|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y}) - \mathbb{E}\, \gamma_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\|_\infty \;\geq\; \epsilon\, p^{-\frac{1}{2}}\, \xi_{\epsilon^\star}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1.$$

Applying Lemma 2,

$$\mathbb{P}\,(\mathcal{D}_{N,\epsilon,p}) \;\leq\; 2\,\exp\left(-\frac{(\epsilon\,\xi_{\epsilon^\star}\,\mathbb{E}\,\|\boldsymbol{Y}\|_1)^2}{p\,(\mathbb{E}\,\|\boldsymbol{Y}\|_1+[D_g]^+)}+\log\,p\right)+\frac{1}{\mathbb{E}\|\boldsymbol{Y}\|_1},$$

where $[D_g]^+ \coloneqq \max\{0,\,D_g\}$, recalling

$$D_g \;\coloneqq\; \sum_{\{i,j\}\prec\{v,w\}\subset\mathcal{N}} \mathbb{C}(Y_{i,j},\,Y_{v,w}),$$

where $\{i,j\}\prec\{v,w\}$ implies the sum is taken with respect to the lexicographical ordering of pairs of nodes. Under the assumption that $\mathbb{E}\|\boldsymbol{Y}\|_1 \geq 1$,

$$2\,\exp\left(-\frac{\epsilon^2\,\xi_{\epsilon^\star}^2\,(\mathbb{E}\|\boldsymbol{Y}\|_1)^2}{p\,(\mathbb{E}\,\|\boldsymbol{Y}\|_1+[D_g]^+)}+\log\,p\right) \;\leq\; 2\,\exp\left(-\frac{\epsilon^2\,\xi_{\epsilon^\star}^2\,\mathbb{E}\|\boldsymbol{Y}\|_1}{p\,(1+[D_g]^+)}+\log\,p\right).$$

Take

$$\epsilon \;=\; \sqrt{\frac{3\,p\,\log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\;\frac{\sqrt{1+[D_g]^+}}{\xi_{\epsilon^\star}}.$$

If

$$\lim_{N\to\infty}\sqrt{\frac{3\,p\,\log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\;\frac{\sqrt{1+[D_g]^+}}{\xi_{\epsilon^\star}} \;=\; 0,$$

then for N sufficiently large, we will have $\epsilon < \epsilon^\star$, which ensures $\epsilon^\star$ may be chosen independent of N and p. While $\epsilon^\star$ can be chosen independent of N and p, note that p is expected to be a function of N and thus $\xi_{\epsilon^\star}$ will not (in general) be independent of N, possibly holding implications for how fast p may grow with N for certain $\boldsymbol{\theta}^\star$ and $\epsilon^\star$. This choice of ϵ in turn implies

$$2\,\exp\left(-\frac{\epsilon^2\,\xi_{\epsilon^\star}^2\,\mathbb{E}\|\boldsymbol{Y}\|_1}{p\,(1+[D_g]^+)}+\log\,p\right) \;=\; 2\,\exp\left(-3\log N+\log p\right) \;\leq\; 2\,N^{-2},$$

under the assumption that $p \leq N$, which ensures $-3\log N + \log p \leq -2\,\log N$. Note that $\mathbb{E}\|\boldsymbol{Y}\|_1 \leq \binom{N}{2} \leq N^2$. We have thus shown, for all $\boldsymbol{\theta}\in\partial\mathcal{B}_2(\boldsymbol{\theta}^\star,\epsilon)$, that

$$\mathbb{P}\left(\|\gamma_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{Y})-\mathbb{E}\,\gamma_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{Y})\|_\infty \;\leq\; \epsilon\,p^{-\frac{1}{2}}\,\xi_{\epsilon^\star}\,\mathbb{E}\,\|\boldsymbol{Y}\|_1\right) \;\geq\; 1-3\,(\mathbb{E}\,\|\boldsymbol{Y}\|_1)^{-1},$$

under the above conditions. As a result, there exists $N_0 \geq 3$ such that, for all $N \geq N_0$ and with probability at least $1-3\,(\mathbb{E}\,\|\boldsymbol{Y}\|_1)^{-1}$, the set $\widehat{\boldsymbol{\Theta}}$ is non-empty and the unique element of the set $\widehat{\boldsymbol{\theta}} \in \widehat{\boldsymbol{\Theta}}$ satisfies (uniqueness following from minimality, as discussed in Section 3)

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;\; \leq \;\; \sqrt{\dfrac{3\, p\, \log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\; \dfrac{\sqrt{1 + [D_g]^+}}{\xi_{\epsilon^\star}}.$$

The above proof can be extended to maximum pseudolikelihood estimators by substituting the relevant quantities (e.g., $\widetilde{\xi}_{\epsilon^\star}$ for $\xi_{\epsilon^\star}$, etc.). The one change of note is that instead of applying the concentration inequality in Lemma 2, we apply the concentration inequality in Lemma 3, which includes an additional factor of K^2. Following these steps and repeating the above proof will show that there exists $N_0 \geq 3$ such that, for all $N \geq N_0$ and with probability at least $1 - 3\,(\mathbb{E}\,\|\boldsymbol{Y}\|_1)^{-1}$, the set $\widetilde{\boldsymbol{\Theta}}$ is non-empty and each $\widetilde{\boldsymbol{\theta}} \in \widetilde{\boldsymbol{\Theta}}$ satisfies

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;\; \leq \;\; \sqrt{\dfrac{3\, p\, K^2 \log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\; \dfrac{\sqrt{1 + [D_g]^+}}{\widetilde{\xi}_{\epsilon^\star}},$$

where

$$\widetilde{\xi}_{\epsilon^\star} \;\; := \;\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(\widetilde{I}(\boldsymbol{\theta})).$$

■

E Proposition 2 and proof

In order to establish a bound on the error of the multivariate normal approximation for estimators of data-generating parameters, we first establish an error bound on the multivariate normal approximation of a standardization of the sufficient statistic vector $\boldsymbol{s}(\boldsymbol{X})$ of the exponential family distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$, derived in Lemma 4, in Proposition 2 using a Lyapunov type bound presented in Raič [2019]. Proposition 2 provides the basis for our normality proof for estimators which we presented in Theorem 2.

Proposition 2 *Consider multilayer networks satisfying* (1) *defined on a set of* $N \geq 3$ *nodes and* $K \geq 1$ *layers. Denote by* $\boldsymbol{s}(\boldsymbol{X}) \in \mathbb{R}^p$ *the sufficient statistic vector of the exponential family* $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \,|\, \boldsymbol{Y} = \boldsymbol{y})$ *as defined in Lemma 4. Let* $\mathbb{E}^{\boldsymbol{Y}}$ *be the random* *conditional expectation operator for the distribution of* $\boldsymbol{X}$ *conditional on* $\boldsymbol{Y}$*, and define*

$$\begin{aligned}\boldsymbol{S}_{\mathcal{N}} &\coloneqq (I(\boldsymbol{\theta}^\star)\,\|\boldsymbol{Y}\|_1)^{-1/2}\,(\boldsymbol{s}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}} \boldsymbol{s}(\boldsymbol{X}))\\ &= \sum_{\{i,j\}\subset\mathcal{N}} (I(\boldsymbol{\theta}^\star)\,\|\boldsymbol{Y}\|_1)^{-1/2}\,(\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}} \boldsymbol{s}_{i,j}(\boldsymbol{X})).\end{aligned}$$

For any measurable convex set $\mathcal{A} \subset \mathbb{R}^p$*,*

$$|\,\mathbb{P}(\boldsymbol{S}_{\mathcal{N}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})\,| \;\leq\; \frac{83}{\xi_{\epsilon^\star}^{3/2}} \sqrt{\frac{p^{7/2}}{\mathbb{E}\,\|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\,\|\boldsymbol{Y}\|_1} + \frac{8\,[D_g]^+}{(\mathbb{E}\,\|\boldsymbol{Y}\|_1)^2},$$

where Φ *is the standard multivariate normal measure and* $\boldsymbol{Z} \sim \mathit{MvtNorm}(\boldsymbol{0}_p, \boldsymbol{I}_p)$*, where* $\boldsymbol{0}_p$ *is the* p*-dimensional vector of zeros and* $\boldsymbol{I}_p$ *is the* $p \times p$ *identity matrix.*

Before we prove Proposition 2, we introduce a Lyapunov type bound in Lemma 6 provided by Theorem 1 of Raic [Raič, 2019].

Lemma 6 *Consider a sequence of* $n \geq 1$ *independent random vectors* $\boldsymbol{W}_i \in \mathbb{R}^p$*. Assume that* $\mathbb{E}\,\boldsymbol{W}_i = \boldsymbol{0}_p$ *and* $\sum_{i=1}^n \mathbb{V}\,\boldsymbol{W}_i = \boldsymbol{I}_p$ *where* $\boldsymbol{0}_p$ *is the* p*-dimensional vector of zeros and* $\boldsymbol{I}_p$ *is the* $p \times p$ *identity matrix. Define*

$$\boldsymbol{S}_n \;=\; \sum_{i=1}^n \boldsymbol{W}_i$$

and let $\boldsymbol{Z}$ *be the standard multivariate normal variable, i.e.,* $\boldsymbol{Z} \sim \mathit{MvtNorm}(\boldsymbol{0}_p, \boldsymbol{I}_p)$*. Then, for all measurable convex sets* $\mathcal{A} \subset \mathbb{R}^p$*,*

$$|\mathbb{P}(\boldsymbol{S}_n \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})| \;\leq\; (42\,p^{1/4} + 16) \sum_{i=1}^n \mathbb{E}\,\|\boldsymbol{W}_i\|_2^3,$$

where Φ *is the standard multivariate normal measure.*

We now turn to proving Proposition 2.

Proof of Proposition 2. By Proposition 1 and Lemma 4, the conditional distribution of the multilayer network $\boldsymbol{X}$ given $\boldsymbol{Y}$ follows an exponential family with sufficient statistic vector that can be decomposed into the sum of conditionally independent dyad-based statistics:

$$\boldsymbol{s}(\boldsymbol{X}) \;=\; \sum_{\{i,j\}\subset\mathcal{N}} \boldsymbol{s}_{i,j}(\boldsymbol{X}),$$

with the precise formula for $\boldsymbol{s}_{i,j}(\boldsymbol{X})$ given in Lemma 4. Define

$$\begin{aligned}\boldsymbol{S}_{\mathcal{N}} \;&\coloneqq\; (I(\boldsymbol{\theta}^\star)\,\|\boldsymbol{Y}\|_1)^{-1/2}\,(\boldsymbol{s}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\,\boldsymbol{s}(\boldsymbol{X}))\\ &=\; \sum_{\{i,j\}\subset\mathcal{N}} (I(\boldsymbol{\theta}^\star)\,\|\boldsymbol{Y}\|_1)^{-1/2}\,(\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\,\boldsymbol{s}_{i,j}(\boldsymbol{X})),\end{aligned}$$

where $I(\boldsymbol{\theta}^\star)$ is the Fisher information matrix of a single dyad $X_{i,j}$ for $\{i,j\}\subset\mathcal{N}$ satisfying $Y_{i,j} = 1$ (i.e., the subset of activated dyads) evaluated at $\boldsymbol{\theta}^\star$ per Lemma 1 and where $\mathbb{E}^{\boldsymbol{Y}}$ is the random conditional expectation operator with respect to the distribution of $\boldsymbol{X}$ conditional on $\boldsymbol{Y}$. For $\epsilon > 0$ satisfying $\epsilon < \mathbb{E}\,\|\boldsymbol{Y}\|_1$, define the event $\mathcal{E}(\epsilon)$ by

$$\mathcal{E}(\epsilon) \;\coloneqq\; \left\{\boldsymbol{y}\in\mathbb{Y} \,:\, \|\boldsymbol{y}\|_1 \geq \mathbb{E}\|\boldsymbol{Y}\|_1 - \epsilon\right\}.$$

In words, $\mathcal{E}(\epsilon)$ is the subset of configurations of the single-layer network $\boldsymbol{Y}$ which have number of edges equal to at least the expected number of activated dyads $\mathbb{E}\,\|\boldsymbol{Y}\|_1$ minus $\epsilon > 0$. The restrictions placed on ϵ ensure that $\mathbb{E}\,\|\boldsymbol{Y}\|_1 - \epsilon > 0$ which implies that $\mathcal{E}(\epsilon)$ will not contain the empty graph which has no edges and that $\mathcal{E}(\epsilon)$ will contain the complete graph with $\binom{N}{2}$ edges as $\mathbb{E}\,\|\boldsymbol{Y}\|_1 < \binom{N}{2}$ (strict inequality following from the fact that $g(\boldsymbol{y})$, the marginal probability mass function of $\boldsymbol{Y}$, is assumed to be strictly positive on $\mathbb{Y}$). Hence, $\mathbb{P}(\mathcal{E}(\epsilon)) > 0$ and $\mathbb{P}(\mathcal{E}(\epsilon)^c) > 0$. Let $\mathcal{A}\subset\mathbb{R}^p$ be a measurable convex set. By the law of total probability and the triangle inequality, we have

$$\begin{aligned}|\,\mathbb{P}(\boldsymbol{S}_{\mathcal{N}}\in\mathcal{A}) - \Phi(\boldsymbol{Z}\in\mathcal{A})\,| \;&\leq\; |\mathbb{P}(\boldsymbol{S}_n\in\mathcal{A}\,|\,\mathcal{E}(\epsilon)) - \Phi(\boldsymbol{Z}\in\mathcal{A})|\;\mathbb{P}(\mathcal{E}(\epsilon))\\ &\quad+\; |\mathbb{P}(\boldsymbol{S}_n\in\mathcal{A}\,|\,\mathcal{E}^c(\epsilon)) - \Phi(\boldsymbol{Z}\in\mathcal{A})|\;\mathbb{P}(\mathcal{E}^c(\epsilon))\\ &\leq\; \sup_{\boldsymbol{y}\in\mathcal{E}(\epsilon)}\,|\,\mathbb{P}(\boldsymbol{S}_{\mathcal{N}}\in\mathcal{A}\,|\,\boldsymbol{Y}=\boldsymbol{y}) - \Phi(\boldsymbol{Z}\in\mathcal{A})\,| \;+\; \mathbb{P}(\mathcal{E}^c(\epsilon)),\end{aligned} \tag{14}$$

noting $|\mathbb{P}(\boldsymbol{S}_n \in \mathcal{A} \,|\, \mathcal{E}^c(\epsilon)) - \Phi(\boldsymbol{Z} \in \mathcal{A})| \leq 1$ and $\mathbb{P}(\mathcal{E}(\epsilon)) \leq 1$. Taking

$$\boldsymbol{W}_{i,j} \;\; = \;\; (I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, (\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\, \boldsymbol{s}_{i,j}(\boldsymbol{X})),$$

we have

$$\mathbb{E}\, [\boldsymbol{W}_{i,j} \,|\, \boldsymbol{Y} = \boldsymbol{y}] \;\; = \;\; 0,$$

a result of the tower property of conditional expectation, and

$$\mathbb{V}\, \left[\textstyle\sum_{\{i,j\} \subset \mathcal{N}} \boldsymbol{W}_{i,j} \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \;\; = \;\; \boldsymbol{I}_p,$$

which follows from Lemma 1 which establishes that $\mathbb{V}[s_{i,j}(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}] = I(\boldsymbol{\theta}^\star)$ when $Y_{i,j} = 1$, recalling the form of the Fisher information matrix of exponential families to be the covariance matrix of the sufficient statistic vector [e.g., Proposition 3.10, pp. 32, Sundberg, 2019], and due to the fact that $\mathbb{V}[s_{i,j}(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}] = \boldsymbol{0}_{p,p}$ when $Y_{i,j} = 0$. Applying Lemma 6 to the first term of the summation of (14), for any measurable convex set $\mathcal{A} \subset \mathbb{R}^p$,

$$|\mathbb{P}(\boldsymbol{S}_{\mathcal{N}} \in \mathcal{A} \,|\, \boldsymbol{Y} = \boldsymbol{y}) - \Phi(\boldsymbol{Z} \in \mathcal{A})| \;\; \leq \;\; (42\, p^{1/4} + 16) \sum_{\{i,j\} \subset \mathcal{N}} \mathbb{E}\, \left[\|\boldsymbol{W}_{i,j}\|_2^3 \,|\, \boldsymbol{Y} = \boldsymbol{y}\right].$$

Given $\boldsymbol{Y} = \boldsymbol{y}$, using standard matrix and vector norm inequalities,

$$\begin{aligned} \|\boldsymbol{W}_{i,j}\|_2 \;\; &= \;\; \|(I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{y}\|_1)^{-1/2}\, (\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}\, \boldsymbol{s}_{i,j}(\boldsymbol{X}))\|_2 \\ &\leq \;\; \|\boldsymbol{y}\|_1^{-1/2}\, \|I(\boldsymbol{\theta}^\star)^{-1/2}\|_2\, \|\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}\, \boldsymbol{s}_{i,j}(\boldsymbol{X})\|_2 \\ &\leq \;\; (\|\boldsymbol{y}\|_1\, \xi_{\epsilon^\star})^{-1/2}\, p^{1/2}\, y_{i,j}, \end{aligned}$$

where $\|\cdot\|_2$ denotes the spectral norm of a $p \times p$ matrix and

$$\xi_{\epsilon^\star} \;\; := \;\; \inf_{\boldsymbol{\theta} \in \mathcal{B}_2(\boldsymbol{\theta}^\star, \epsilon^\star)} \lambda_{\min}(I(\boldsymbol{\theta})),$$

for a given and fixed $\epsilon^\star > 0$, as defined in Section 3. From proofs of Lemma 2 and 3,

$$0 \;\; \leq \;\; s_{l,i,j}(\boldsymbol{x}) \;\; \leq \;\; 1, \qquad \text{all } l = 1, \ldots, p, \; \{i,j\} \subset \mathcal{N},$$

$\mathbb{P}$-almost surely. Hence,

$$\mathbb{P}(\|\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}} \boldsymbol{s}_{i,j}(\boldsymbol{X})\|_\infty \leq y_{i,j} \,|\, \boldsymbol{Y} = \boldsymbol{y}) \;\; = \;\; 1,$$

implying (conditional on $\boldsymbol{Y} = \boldsymbol{y}$)

$$\|\boldsymbol{s}_{i,j}(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\, \boldsymbol{s}_{i,j}(\boldsymbol{X})\|_2 \;\; \leq \;\; (p\, y_{i,j})^{1/2} \;\; = \;\; p^{1/2}\, y_{i,j},$$

$\mathbb{P}$-almost surely. As a result,

$$\mathbb{E}\; [\|\boldsymbol{W}_{i,j}\|_2^3 \,|\, \boldsymbol{Y} = \boldsymbol{y}] \;\; \leq \;\; (\|\boldsymbol{y}\|_1\, \xi_{\epsilon^\star})^{-3/2}\, p^{3/2}\, y_{i,j},$$

noting that $y_{i,j}^3 = y_{i,j} \in \{0, 1\}$. Using the fact that $42\, p^{1/4} + 16 \leq 58\, p^{1/4}$ ($p \geq 1$), we have

$$\begin{aligned}(42\, p^{1/4} + 16) \sum_{\{i,j\} \subset \mathcal{N}} \mathbb{E}\; \left[\|\boldsymbol{W}_{i,j}\|_2^3 \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \;\; &\leq \;\; 58\, p^{7/4} \sum_{\{i,j\} \subset \mathcal{N}} y_{i,j}\, (\|\boldsymbol{y}\|_1\, \xi_{\epsilon^\star})^{-3/2} \\ &= \;\; 58\, p^{7/4}\, \|\boldsymbol{y}\|_1^{-1/2}\, \xi_{\epsilon_\star}^{-3/2} \\ &\leq \;\; 58\, p^{7/4}\, (\mathbb{E}\, \|\boldsymbol{Y}\|_1 - \epsilon)^{-1/2}\, \xi_{\epsilon_\star}^{-3/2},\end{aligned}$$

as the conditioning event $\mathcal{E}(\epsilon)$ and choice of ϵ ensure that $\|\boldsymbol{y}\|_1 \geq \mathbb{E}\|\boldsymbol{Y}\|_1 - \epsilon > 0$. We bound the second term in (14) by Chebyshev's inequality using equation (10) in the proof of Lemma 2:

$$\mathbb{P}(\mathcal{E}^c(\epsilon)) \;\; \leq \;\; \frac{\mathbb{E}\, \|\boldsymbol{Y}\|_1 + 2\; [D_g]^+}{\epsilon^2}.$$

Taking $\epsilon = 2^{-1}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1 > 0$, we have

$$\mathbb{P}(\mathcal{E}^c(\epsilon)) \;\; \leq \;\; \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\; [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}.$$

Combining terms, we obtain the bound

$$|\mathbb{P}(\boldsymbol{S}_{\mathcal{N}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})| \;\; \leq \;\; \frac{83}{\xi_{\epsilon^\star}^{3/2}} \sqrt{\frac{p^{7/2}}{\mathbb{E}\, \|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\; [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}.$$

■

Note that the asymptotic multivariate normality can be established provided

$$\lim_{N\to\infty}\left[\frac{83}{\xi_{\epsilon^\star}^{3/2}}\sqrt{\frac{p^{7/2}}{\mathbb{E}\,\|\boldsymbol{Y}\|_1}}+\frac{4}{\mathbb{E}\,\|\boldsymbol{Y}\|_1}+\frac{8\,[D_g]^+}{(\mathbb{E}\,\|\boldsymbol{Y}\|_1)^2}\right] \;=\; 0,$$

resulting in following the asymptotic convergence in distribution:

$$\boldsymbol{S}_{\mathcal{N}} \;\xrightarrow{D}\; \boldsymbol{Z} \;\sim\; \mathrm{MvtNorm}\,(\boldsymbol{0},\,\boldsymbol{I}_p)\,.$$

F Proof of Theorem 2

Proof of Theorem 2. In order to demonstrate the feasibility of the normal approximation for maximum likelihood estimators $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$, we first start with a standard Taylor expansion of the negative score equation:

$$-\nabla_{\boldsymbol{\theta}}\,\ell(\widehat{\boldsymbol{\theta}};\boldsymbol{x},\boldsymbol{y}) \;=\; -\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta}^\star;\boldsymbol{x},\boldsymbol{y})-\nabla_{\boldsymbol{\theta}}^2\,\ell(\boldsymbol{\theta}^\star;\boldsymbol{x},\boldsymbol{y})\,(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star)-\boldsymbol{R}, \tag{15}$$

where $\boldsymbol{R}\in\mathbb{R}^p$ is the vector of remainders given in the Lagrange form. Denoting by R_i, $(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star)_i$, and $(\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}))_i$ the i^{th} component of $\boldsymbol{R}$, $(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star)$, and the score function $\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y})$, respectively. The remainder term R_i $(i=1,\ldots,p)$ is given by

$$R_i \;=\; \sum_{j=1}^{p}\frac{1}{2}\,\frac{\partial^2\,(\nabla_{\boldsymbol{\theta}}\,\ell(\dot{\boldsymbol{\theta}}_i;\boldsymbol{x},\boldsymbol{y}))_i}{\partial\,\theta_j^2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star)_j^2+\sum_{1\le j<k\le p}\frac{\partial^2\,(\nabla_{\boldsymbol{\theta}}\ell(\dot{\boldsymbol{\theta}}_i;\boldsymbol{x},\boldsymbol{y}))_i}{\partial\,\theta_j\,\partial\,\theta_k}\,(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star)_j\,(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star)_k,$$

where $\dot{\boldsymbol{\theta}}_i = t_i\,\widehat{\boldsymbol{\theta}}+(1-t_i)\,\boldsymbol{\theta}^\star$ (for some $t_i\in[0,1]$). By Proposition 1,

$$\ell(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) \;=\; \log\,\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}=\boldsymbol{x}\,|\,\boldsymbol{Y}=\boldsymbol{y})+\log\,g(\boldsymbol{y}).$$

By Lemma 4, the probability mass function $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X}=\boldsymbol{x}\,|\,\boldsymbol{Y}=\boldsymbol{y})$ belongs to a minimal exponential family with sufficient statistic vector $s(\boldsymbol{x})=(s_1(\boldsymbol{x}),\ldots,s_p(\boldsymbol{x}))$ given by equation (11) in Lemma 4. We then have,

$$-\nabla_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) \;=\; -(s(\boldsymbol{x})-\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{y}}\,s(\boldsymbol{X}))$$

$$-\nabla_{\boldsymbol{\theta}}^2\,\ell(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) \;=\; \mathbb{V}_{\boldsymbol{\theta}}^{\boldsymbol{y}}\,s(\boldsymbol{X}) \;=\; I(\boldsymbol{\theta}^\star)\,\|\boldsymbol{y}\|_1,$$

where $\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{y}}$ and $\mathbb{V}_{\boldsymbol{\theta}}^{\boldsymbol{y}}$ are the conditional expectation and variance operators, respectively, of the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$, and by using standard formulas for exponential families [e.g., Proposition 3.8, pp. 29, Sundberg, 2019] and the results of Lemma 1. Note $\nabla_{\boldsymbol{\theta}}\, \ell(\widehat{\boldsymbol{\theta}}; \boldsymbol{x}, \boldsymbol{y}) = 0$, as the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ solves the score equation by definition. We re-arrange (15) and multiply both sides by $(I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}$ to obtain

$$\begin{aligned} & (I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \ (I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, \boldsymbol{R} \\ = \ & (I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, (s(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\, s(\boldsymbol{X})). \end{aligned} \tag{16}$$

Define $\tilde{\boldsymbol{R}} \coloneqq (I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, \boldsymbol{R}$. Let $\mathcal{A} \subset \mathbb{R}^p$ be any measurable convex subset of $\mathbb{R}^p$ and $\boldsymbol{Z} \sim \mathrm{MvtNorm}(\boldsymbol{0}_p, \boldsymbol{I}_p)$. We are interested in bounding the quantity

$$\left|\mathbb{P}((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})\right|.$$

Then from (16),

$$\mathbb{P}\left((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}} \in \mathcal{A}\right) \ = \ \mathbb{P}\big((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, (s(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\, s(\boldsymbol{X})) \in \mathcal{A}\big).$$

Applying Proposition 2, for all measurable convex sets $\mathcal{A} \subseteq \mathbb{R}^p$,

$$\begin{aligned} & \left|\mathbb{P}\left((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, (s(\boldsymbol{X}) - \mathbb{E}^{\boldsymbol{Y}}\, s(\boldsymbol{X})) \in \mathcal{A}\right) - \Phi(\boldsymbol{Z} \in \mathcal{A})\right| \\ \leq \ & \frac{83}{\xi_{\epsilon^\star}^{3/2}} \sqrt{\frac{p^{7/2}}{\mathbb{E}\, \|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}. \end{aligned}$$

Hence,

$$\begin{aligned} & \left|\mathbb{P}((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})\right| \\ \leq \ & \frac{83}{\xi_{\epsilon^\star}^{3/2}} \sqrt{\frac{p^{7/2}}{\mathbb{E}\, \|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}. \end{aligned}$$

We lastly handle the term $\tilde{\boldsymbol{R}}$ by showing that $\|\tilde{\boldsymbol{R}}\|_2$ is small with high probability. We first use standard vector/matrix norm inequalities to bound

$$\|\widetilde{\boldsymbol{R}}\|_2 \ = \ \|(I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{-1/2}\, \boldsymbol{R}\|_2 \ \leq \ \frac{\|I(\boldsymbol{\theta}^\star)^{-1/2}\|_2}{\sqrt{\|\boldsymbol{Y}\|_1}}\, \|\boldsymbol{R}\|_2 \ \leq \ \frac{\|\boldsymbol{R}\|_2}{\sqrt{\xi_{\epsilon^\star} \|\boldsymbol{Y}\|_1}},$$

noting that the spectral norm $\|I(\boldsymbol{\theta}^\star)^{-1/2}\|_2$ is equal to the largest eigenvalue of $I(\boldsymbol{\theta}^\star)^{-1/2}$ which will be the reciprocal of the smallest eigenvalue of $I(\boldsymbol{\theta}^\star)^{1/2}$, which is bounded below by $\sqrt{\xi_{\epsilon^\star}}$. Using a standard result from the Taylor theorem for functions with multiple variables, if for each $i = 1, \dots, p$, there exists constants $M_i > 0$ such that

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^p \,:\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_1 \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1} \left| \frac{\partial^2 (\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial\, \theta_j\, \partial\, \theta_k} \right| \;\leq\; M_i, \qquad 1 \leq j \leq k \leq p,$$

then the Lagrange remainder is bounded above by

$$R_i \;\leq\; \frac{M_i}{2}\, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1^2$$

on the set $\{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_1 \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2\}$. By Lemma 7, conditional on $\boldsymbol{Y} = \boldsymbol{y}$, we have, for all $i = 1, \dots, p$, the bound $M_i \leq 2\, \|\boldsymbol{y}\|_1$. Hence,

$$\begin{aligned}
\|\tilde{\boldsymbol{R}}\|_2 \;&\leq\; \frac{1}{\sqrt{\xi_{\epsilon^\star}}\, \|\boldsymbol{y}\|_1} \sqrt{\sum_{i=1}^p R_i^2} &&\leq\; \frac{1}{\sqrt{\xi_{\epsilon^\star}}\, \|\boldsymbol{y}\|_1} \sqrt{\sum_{i=1}^p \|\boldsymbol{y}\|_1^2\, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1^4} \\
&\leq\; \frac{1}{\sqrt{\xi_{\epsilon^\star}}\, \|\boldsymbol{y}\|_1} \sqrt{p\, \|\boldsymbol{y}\|_1^2\, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1^4} &&\leq\; \frac{\sqrt{p}\, \|\boldsymbol{y}\|_1\, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_1^2}{\sqrt{\xi_{\epsilon^\star}}\, \|\boldsymbol{y}\|_1} \qquad (17) \\
&\leq\; \frac{\sqrt{p}\, \sqrt{\|\boldsymbol{y}\|_1}\, p\, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2}{\sqrt{\xi_{\epsilon^\star}}} &&\leq\; \frac{p^{3/2}\, \sqrt{\|\boldsymbol{y}\|_1}\, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2}{\sqrt{\xi_{\epsilon^\star}}}.
\end{aligned}$$

By Chebyshev's inequality—as in the proof of Lemma 2—we can establish that

$$\mathbb{P}\left(|\|\boldsymbol{Y}\|_1 - \mathbb{E}\, \|\boldsymbol{Y}\|_1| > \frac{1}{2}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1 \right) \;\leq\; \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}, \qquad (18)$$

whereas Theorem 1 established there exists $N_0 \geq 3$ such that, for all $N \geq N_0$, the event

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;\leq\; \sqrt{\frac{3\, p\, \log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}}\; \frac{\sqrt{1 + [D_g]^+}}{\xi_{\epsilon^\star}}, \qquad (19)$$

occurs with probability at least $1 - 3\, (\mathbb{E}\, \|\boldsymbol{Y}\|_1)^{-1}$. Define $\mathcal{E}_1$ to be the event

$$|\|\boldsymbol{Y}\|_1 - \mathbb{E}\, \|\boldsymbol{Y}\|_1| \;\leq\; \frac{1}{2}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1$$

and $\mathcal{E}_2$ to be the event in (19), and define $\mathcal{R}$ to be the corresponding values of $\tilde{\boldsymbol{R}}$ in the event $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{E}_1 \cap \mathcal{E}_2$, under which we have the bound

$$\begin{aligned}
\|\tilde{\boldsymbol{R}}\|_2 \;\; &\leq \;\; \frac{p^{3/2}\, \sqrt{\|\boldsymbol{y}\|_1}}{\sqrt{\xi_{\epsilon^\star}}}\, \frac{3\, p\, \log N}{\mathbb{E}\|\boldsymbol{Y}\|_1}\, \frac{1 + [D_g]^+}{\xi_{\epsilon^\star}^2} \\
&= \;\; \frac{3\, p^{5/2}\, \sqrt{2\, \mathbb{E}\, \|\boldsymbol{Y}\|_1}\, \log\, N}{\mathbb{E}\|\boldsymbol{Y}\|_1}\, \frac{1 + [D_g]^+}{\xi_{\epsilon^\star}^2} \\
&= \;\; \frac{3\, \sqrt{2}\, (1 + [D_g]^+)}{\xi_{\epsilon^\star}^2}\, \frac{p^{5/2}\, \log N}{\sqrt{\mathbb{E}\, \|\boldsymbol{Y}\|_1}}.
\end{aligned} \tag{20}$$

combining the bounds in (17), (18), and (19) and using the fact that

$$\|\boldsymbol{y}\|_1 \;\; \leq \;\; \mathbb{E}\, \|\boldsymbol{Y}\|_1 + \frac{1}{2}\, \mathbb{E}\, \|\boldsymbol{Y}\|_1 \;\; \leq \;\; 2\, \mathbb{E}\, \|\boldsymbol{Y}\|_1$$

in the event $\boldsymbol{y} \in \mathcal{E}_1$. Moreover, a union bound shows that

$$\mathbb{P}(\tilde{\boldsymbol{R}} \notin \mathcal{R}) \;\; \leq \;\; \frac{7}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}.$$

Hence,

$$\mathbb{P}\left(\|\tilde{\boldsymbol{R}}\|_2 \leq \frac{3\, \sqrt{2}\, (1 + [D_g]^+)}{\xi_{\epsilon^\star}^2}\, \frac{p^{5/2}\, \log N}{\sqrt{\mathbb{E}\, \|\boldsymbol{Y}\|_1}}\right) \;\; \geq \;\; 1 - \frac{7}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} - \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}. \tag{21}$$

Taken together, we have shown under the assumptions of Theorem 1 that there exists $N_0 \geq 3$ such that, for all $N \geq N_0$, the error of the multivariate normal approximation

$$\left|\mathbb{P}((I(\boldsymbol{\theta}^\star)\, \|\boldsymbol{Y}\|_1)^{1/2}\, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) - \tilde{\boldsymbol{R}} \in \mathcal{A}) - \Phi(\boldsymbol{Z} \in \mathcal{A})\right|$$

is bounded above by

$$\frac{83}{\xi_{\epsilon^\star}^{3/2}}\, \sqrt{\frac{p^{7/2}}{\mathbb{E}\, \|\boldsymbol{Y}\|_1}} + \frac{4}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} + \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}$$

where $\tilde{\boldsymbol{R}}$ satisfies

$$\mathbb{P}\left(\|\tilde{\boldsymbol{R}}\|_2 \leq \frac{3\, \sqrt{2}\, (1 + [D_g]^+)}{\xi_{\epsilon^\star}^2}\, \frac{p^{5/2}\, \log N}{\sqrt{\mathbb{E}\, \|\boldsymbol{Y}\|_1}}\right) \;\; \geq \;\; 1 - \frac{7}{\mathbb{E}\, \|\boldsymbol{Y}\|_1} - \frac{8\, [D_g]^+}{(\mathbb{E}\, \|\boldsymbol{Y}\|_1)^2}.$$

■

F.1 Auxiliary results for proof of Theorem 2

Lemma 7 *Consider multilayer networks satisfying (1) defined on a set of $N \geq 3$ nodes and $K \geq 1$ layers and denote by the log-likelihood function by $\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$. Then, for each $i = 1, \dots, p$ and $\epsilon > 0$,*

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^p \,:\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 \leq \epsilon} \left| \frac{\partial^2 \, (\nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial \, \theta_j \, \partial \, \theta_k} \right| \;\; \leq \;\; 2 \, \|\boldsymbol{y}\|_1,$$

where $(\nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i$ is the i^{th} component of the score function $\nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$.

Proof of Lemma 7. By Proposition 1, given the observation $\boldsymbol{x}$ of $\boldsymbol{X}$ (i.e., observation of the event $\boldsymbol{X} = \boldsymbol{x}$), $\boldsymbol{Y}$ is predictable with unique value $\boldsymbol{y} \in \mathbb{Y}$ given by the formula in Proposition 1, and $(\boldsymbol{x}, \boldsymbol{y})$ is network concordant. Further, by Proposition 1

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) \;\; = \;\; \log \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}) + \log \, g(\boldsymbol{y}),$$

where $\log \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{Y} = \boldsymbol{y})$ is the log-likelihood of a minimal, full, and regular exponential family. Thus, the second order derivative of $\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ with respect to the i^{th} and j^{th} components of $\boldsymbol{\theta}$ correspond to the variance (in the case $i = j$) or covariance (in the case of $i \neq j$) of corresponding sufficient statistic(s) of the exponential family [e.g., Proposition 3.8, p. 29, Sundberg, 2019], with sufficient statistics given in Lemma 4. For $\{i, j\} \subseteq \{1, \dots, p\}$,

$$\frac{\partial \, (\nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial \, \theta_j} \;\; = \;\; \frac{\partial^2 \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})}{\partial \, \theta_i \, \partial \, \theta_j} \;\; = \;\; \mathbb{C}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}), s_j(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}),$$

and when $i = j \in \{1, \dots, p\}$,

$$\frac{\partial \, (\nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial \, \theta_i} \;\; = \;\; \frac{\partial^2 \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})}{\partial \, \theta_i^2} \;\; = \;\; \mathbb{V}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}).$$

As a result, for $\{i, j\} \subseteq \{1, \dots, p\}$ and $k \in \{1, \dots, p\}$,

$$\left| \frac{\partial^2 \, (\nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial \, \theta_j \, \partial \, \theta_k} \right| \;\; = \;\; \left| \frac{\partial \, \mathbb{C}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}), s_j(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial \, \theta_k} \right|,$$

and when $i = j \in \{1, \dots, p\}$ and $k \in \{1, \dots, p\}$,

$$\left| \dfrac{\partial^2 \left(\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})\right)_i}{\partial\, \theta_i\, \partial\, \theta_k} \right| \;\; = \;\; \left| \dfrac{\partial\, \mathbb{V}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right|.$$

By Lemma 4 equation (11), conditional on $\boldsymbol{Y} = \boldsymbol{y}$, each sufficient statistic $s_i(\boldsymbol{X})$ ($i \in \{1, \dots, p\}$) can be decomposed into the sum of conditionally independent statistics of each dyad $\boldsymbol{X}_{v,w}$, for $\{v, w\} \subseteq \mathcal{N}$. We can then write

$$\mathbb{C}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}), s_j(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}) \;\; = \;\; \sum_{\{v,w\} \subset \mathcal{N}} \mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}),\, s_{j,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y}),$$

noting that by conditional independence $\mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}), s_{j,r,t}(\boldsymbol{X}_{r,t}) \,|\, \boldsymbol{Y} = \boldsymbol{y}) = 0$ whenever $\{r, t\} \neq \{v, w\}$, and when $i = j$, we can write

$$\mathbb{V}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}) \;\; = \;\; \sum_{\{v,w\} \subset \mathcal{N}} \mathbb{V}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y}),$$

again appealing to the conditional independence given $\boldsymbol{Y}$ of the random variables $s_{i,v,w}(\boldsymbol{X}_{v,w})$ ($\{v, w\} \subset \mathcal{N}$). As a result, for $k \in \{1, \dots, p\}$, it suffices to show that,

$$\left| \dfrac{\partial\, \mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}),\, s_{j,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \;\; \leq \;\; 2,$$

and

$$\left| \dfrac{\partial\, \mathbb{V}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \;\; \leq \;\; 1.$$

Recall that the sufficient statistic $s_{i,v,w}(\boldsymbol{X})$ ($i = 1, \dots, p$) is defined in Lemma 4 by

$$s_{i,v,w}(\boldsymbol{X}_{v,w}) \;\; = \;\; \prod_{t=1}^{h} X_{v,w}^{(k_t)}, \qquad \{v, w\} \subset \mathcal{N},$$

for some $h \in \{1, \dots, H\}$ and $\{k_1, \dots, k_h\} \subseteq \{1, \dots, K\}$. Define the set $S_{i,v,w}$ of components of the sufficient statistic vector $\boldsymbol{s}_{v,w}(\boldsymbol{X})$ for $\{v, w\} \subset \mathcal{N}$ and $i = 1, \dots, p$ by

$$S_{i,v,w} \;\; \coloneqq \;\; \left\{ \prod_{t=1}^{h'} X_{v,w}^{(l_t)} \;:\; \{l_1, \dots, l_{h'}\} \subset \{k_1, \dots, k_h\},\; h' < h \right\},$$

where $h \in \{1, \dots, H\}$ and $\{k_1, \dots, k_h\} \subseteq \{1, \dots, K\}$. The set $S_{i,v,w}$ is the set of components of the sufficient statistic vector $\boldsymbol{s}_{v,w}(\boldsymbol{X})$ of dyad $\{v, w\} \subset \mathcal{N}$ that have a value of 1 when

$s_{i,v,w}(\boldsymbol{X}) = 1$. For the ease of notation, let $I_{S_{i,v,w}}$ be the set of dimension indices whose corresponding components of the sufficient statistic vector $\boldsymbol{s}_{v,w}(\boldsymbol{X})$ belong to the set $S_{i,v,w}$:

$$I_{S_{i,v,w}} \;\coloneqq\; \{j \in \{1, \dots, p\} \;:\; s_{j,v,w}(\boldsymbol{X}) \in S_{i,v,w}\}.$$

Define the conditional expectation of $s_{i,v,w}(\boldsymbol{X})$ given $\boldsymbol{Y} = \boldsymbol{y}$ for any $i \in \{1, \dots, p\}$ and $\{v, w\} \subset \mathcal{N}$ by

$$P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) \;\coloneqq\; \mathbb{P}_{\boldsymbol{\theta}}\left(s_{i,v,w}(\boldsymbol{X}) = 1 \,|\, \boldsymbol{Y} = \boldsymbol{y}\right).$$

For further notation simplicity, denote by L_i the set of layer indices $\{k_1, \dots, k_h\} \subseteq \{1, \dots, K\}$ that define the i^{th} component $s_{i,v,w}(\boldsymbol{X}_{v,w})$ of the sufficient statistic vector $\boldsymbol{s}_{v,w}(\boldsymbol{X})$ for any $\{v, w\} \subset \mathcal{N}$, $j \in \{1, \dots, p\}$, and some $h \in \{1, \dots, H\}$. We then define

$$\boldsymbol{X}_{v,w}^{(L_i)} \;\coloneqq\; \left\{X_{v,w}^{(k_1)}, \dots, X_{v,w}^{(k_h)}\right\}, \qquad \boldsymbol{X}_{v,w}^{(-L_i)} \;\coloneqq\; \boldsymbol{X}_{v,w} \setminus \boldsymbol{X}_{v,w}^{(L_i)},$$

and the corresponding sample space

$$\mathbb{X}_{v,w}^{(L_i)} \;\coloneqq\; \{0,1\}^h, \qquad \mathbb{X}_{v,w}^{(-L_i)} \;\coloneqq\; \{0,1\}^{H-h},$$

for some $h \in \{1, \dots, H\}$. Then we can write

$$\begin{aligned} P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) \;&=\; \mathbb{P}_{\boldsymbol{\theta}}\left(\prod_{l \in L_i} X_{v,w}^{(l)} = 1 \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) \\ &=\; \frac{\displaystyle\sum_{\mathbb{X}_{v,w}^{(-L_i)}} \exp\left(\sum_{j \in I_{S_{i,v,w}}} \theta_j + \sum_{j \in I_{S_{i,v,w}}^c} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right)}{\displaystyle\sum_{\mathbb{X}_{v,w}} \exp\left(\sum_{j=1}^{p} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right)}. \end{aligned}$$

Let

$$Z(\boldsymbol{\theta}) \;\coloneqq\; \sum_{\mathbb{X}_{v,w}} \exp\left(\sum_{j=1}^{p} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right),$$

and take the derivative of $P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$ with respect to θ_k for $k = 1, \ldots, p$. We have

$$\begin{aligned}
& \frac{\partial\, P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})}{\partial\, \theta_k} \\
\leq\ & \frac{\displaystyle\sum_{\mathbb{X}_{v,w}^{(-L_i)}} \exp\left(\sum_{j \in I_{S_{i,v,w}}} \theta_j + \sum_{j \in I_{S_{i,v,w}}^c} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right) \left(Z(\boldsymbol{\theta}) - \frac{\partial\, Z(\boldsymbol{\theta})}{\partial\, \theta_k}\right)}{Z(\boldsymbol{\theta})^2} \\
=\ & \frac{\displaystyle\sum_{\mathbb{X}_{v,w}^{(-L_i)}} \exp\left(\sum_{j \in I_{S_{i,v,w}}} \theta_j + \sum_{j \in I_{S_{i,v,w}}^c} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right) \left(\sum_{\mathbb{X}_{v,w}} \exp\left(\sum_{j=1}^{p} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right) (1 - s_{k,v,w}(\boldsymbol{x}))\right)}{Z(\boldsymbol{\theta})^2} \\
\leq\ & \frac{\displaystyle\sum_{\mathbb{X}_{v,w}^{(-L_i)}} \exp\left(\sum_{j \in I_{S_{i,v,w}}} \theta_j + \sum_{j \in I_{S_{i,v,w}}^c} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right) Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})^2} \\
\leq\ & 1.
\end{aligned} \tag{22}$$

The first inequality is obtained because $s_{k,v,w}(\boldsymbol{x}) \leq 1$, and the last inequality is due to the fact that

$$\sum_{\mathbb{X}_{v,w}^{(-L_i)}} \exp\left(\sum_{j \in I_{S_{i,v,w}}} \theta_j + \sum_{j \in I_{S_{i,v,w}}^c} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right) \ \leq\ \sum_{\mathbb{X}_{v,w}} \exp\left(\sum_{j=1}^{p} \theta_j\, s_{j,v,w}(\boldsymbol{x})\right).$$

Now we turn to show the derivative of the conditional variance and covariance of the sufficient statistics of each dyad are bounded. Given $\boldsymbol{Y} = \boldsymbol{y}$, for all $\{i\} \subset \{1, \ldots, p\}$, $s_{i,v,w}(\boldsymbol{X})$ are conditionally independent across $\{v, w\} \subseteq \mathcal{N}$. Then we have

$$\begin{aligned}
& \mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}),\ s_{j,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y}) \\
=\ & \mathbb{E}\left[s_{i,v,w}(\boldsymbol{X})\, s_{j,v,w}(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] - \mathbb{E}\left[s_{i,v,w}(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \mathbb{E}\left[s_{j,v,w}(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y}\right] \\
=\ & \mathbb{P}_{\boldsymbol{\theta}}\left(s_{i,v,w}(\boldsymbol{X}) = 1, s_{j,v,w}(\boldsymbol{X}) = 1 \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) - P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})\, P_{j,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) \\
=\ & \mathbb{P}_{\boldsymbol{\theta}}\left(\prod_{l \in L_i \cup L_j} \boldsymbol{X}_{v,w}^{(l)} = 1 \,|\, \boldsymbol{Y} = \boldsymbol{y}\right) - P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})\, P_{j,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}).
\end{aligned}$$

Using the inequality derived in (22) and suppressing the notation of $\{v, w\}$ and $(\boldsymbol{X}, \boldsymbol{y})$ in $P_{i,v,w}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})$, the derivative of the covariance with respect to θ_k, $k = 1, \dots, p$ is given by

$$\begin{aligned} & \left| \frac{\partial\, \mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}),\, s_{j,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \\ = \; & \frac{\partial\, \mathbb{P}_{\boldsymbol{\theta}} \left(\displaystyle\prod_{l \in L_i \cup L_j} \boldsymbol{X}_{v,w}^{(l)} = 1 \,|\, \boldsymbol{Y} = \boldsymbol{y} \right)}{\partial\, \theta_k} - \frac{\partial\, P_i(\boldsymbol{\theta})}{\partial\, \theta_k}\, P_j(\boldsymbol{\theta}) - P_i(\boldsymbol{\theta})\, \frac{\partial\, P_j(\boldsymbol{\theta})}{\partial\, \theta_k} \\ \leq \; & 2. \end{aligned}$$

Using the same inequality and notation in (22), the derivative of the variance of a Bernoulli random variable $s_{i,v,w}(\boldsymbol{X})$ is given by

$$\left| \frac{\partial\, \mathbb{V}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \;=\; \left| (1 - 2\, P_i(\boldsymbol{\theta}))\, \frac{\partial\, P_i(\boldsymbol{\theta})}{\partial\, \theta_k} \right| \;\leq\; 1.$$

Finally, for $\{i, j\} \subseteq \{1, \dots, p\}$ and $k \in \{1, \dots, p\}$, we obtain

$$\begin{aligned} \left| \frac{\partial^2\, (\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial\, \theta_j\, \partial\, \theta_k} \right| \;&=\; \left| \frac{\partial\, \mathbb{C}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}), s_j(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \\ &\leq\; \sum_{\{v,w\} \subset \mathcal{N}} \left| \frac{\partial\, \mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}),\, s_{j,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \\ &\leq\; 2\, \|\boldsymbol{y}\|_1 \end{aligned}$$

due to the fact that $\mathbb{C}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{v,w}),\, s_{j,v,w}(\boldsymbol{X}_{v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y}) = 0$ when $Y_{v,w} = 0$ for $\{v, w\} \subset \mathcal{N}$. Similarly, $\mathbb{V}_{\boldsymbol{\theta}}(s_{i,v,w}(\boldsymbol{X}_{i,v,w}) \,|\, \boldsymbol{Y} = \boldsymbol{y}) = 0$ when $Y_{v,w} = 0$ for $\{v, w\} \subset \mathcal{N}$, and when $i = j \in \{1, \dots, p\}$ and $k \in \{1, \dots, p\}$, we have

$$\begin{aligned} \left| \frac{\partial^2\, (\nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}))_i}{\partial\, \theta_i\, \partial\, \theta_k} \right| \;&=\; \left| \frac{\partial\, \mathbb{V}_{\boldsymbol{\theta}}(s_i(\boldsymbol{X}) \,|\, \boldsymbol{Y} = \boldsymbol{y})}{\partial\, \theta_k} \right| \\ &\leq\; \|\boldsymbol{y}\|_1. \end{aligned}$$

■

G Additional simulation results

We present additional simulation results that enhance those contained in Section 5.

Table 4: P-values of the Zhou-Shao's test for multivariate normality of $\widetilde{\boldsymbol{\theta}}$ for 6 model-generating parameters ($\boldsymbol{\theta}_1^\star$, $\boldsymbol{\theta}_2^\star$, $\boldsymbol{\theta}_3^\star$, $\boldsymbol{\theta}_4^\star$, $\boldsymbol{\theta}_5^\star$, $\boldsymbol{\theta}_6^\star$) estimated from 250 network samples at size 1000 on four basis network structures. All p-values are larger than .1.

Basis network model	$\boldsymbol{\theta}_1^\star$	$\boldsymbol{\theta}_2^\star$	$\boldsymbol{\theta}_3^\star$	$\boldsymbol{\theta}_4^\star$	$\boldsymbol{\theta}_5^\star$	$\boldsymbol{\theta}_6^\star$
Dense Bernoulli	.138	.473	.053	.699	.587	.983
Sparse Bernoulli	.554	.132	.232	.634	.904	.373
SBM	.65	.891	.982	.975	.871	.674
LSM	.859	.831	.5	.227	.613	.409

G.1 Normal approximation with different basis networks

The multivariate normality of $\widetilde{\boldsymbol{\theta}}$ is tested by Zhou-Shao's multivariate normal test [Zhou and Shao, 2014], and the p-values are provided in tabel 4. Q-Q plots of $\widetilde{\boldsymbol{\theta}}$ estimated from 6 different model-generating parameters with a dense Bernoulli basis network, a sparse Bernoulli basis network, a stochastic block model (SBM) generated basis network, and a latent space model (LSM) generated basis network are shown in Figure 6, 7, 8 and 9, respectively.

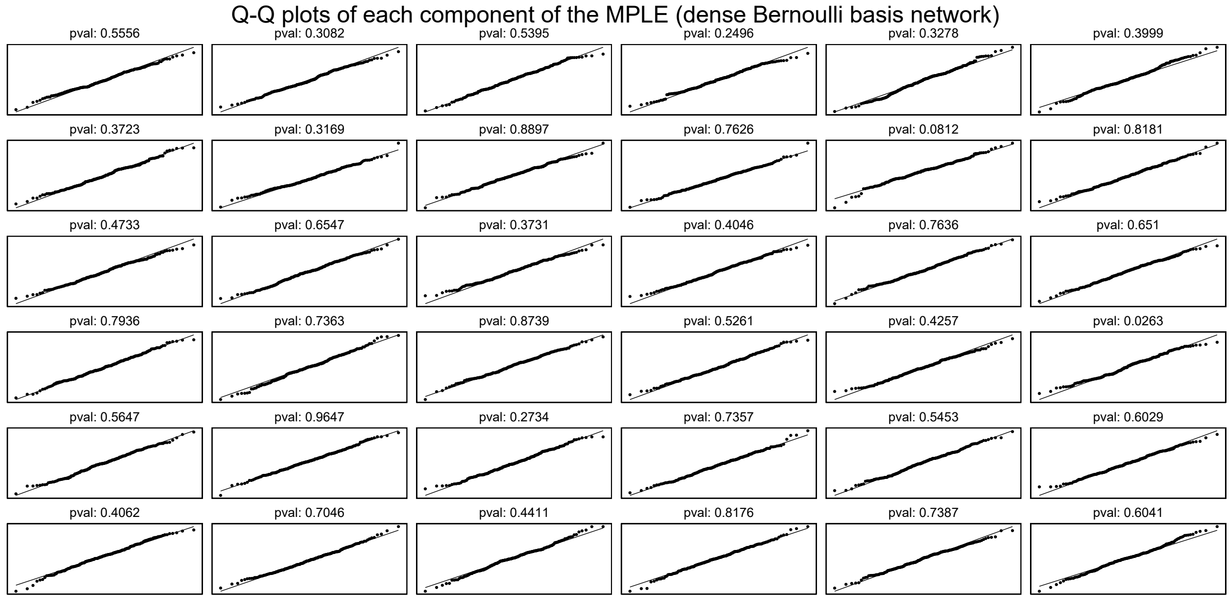

Figure 6: Q-Q plots and p-values of six components of $\widetilde{\boldsymbol{\theta}}$ estimated from 250 multilayer network samples at size 1000 on the dense Bernoulli basis network for 6 model-generating parameters on each row.

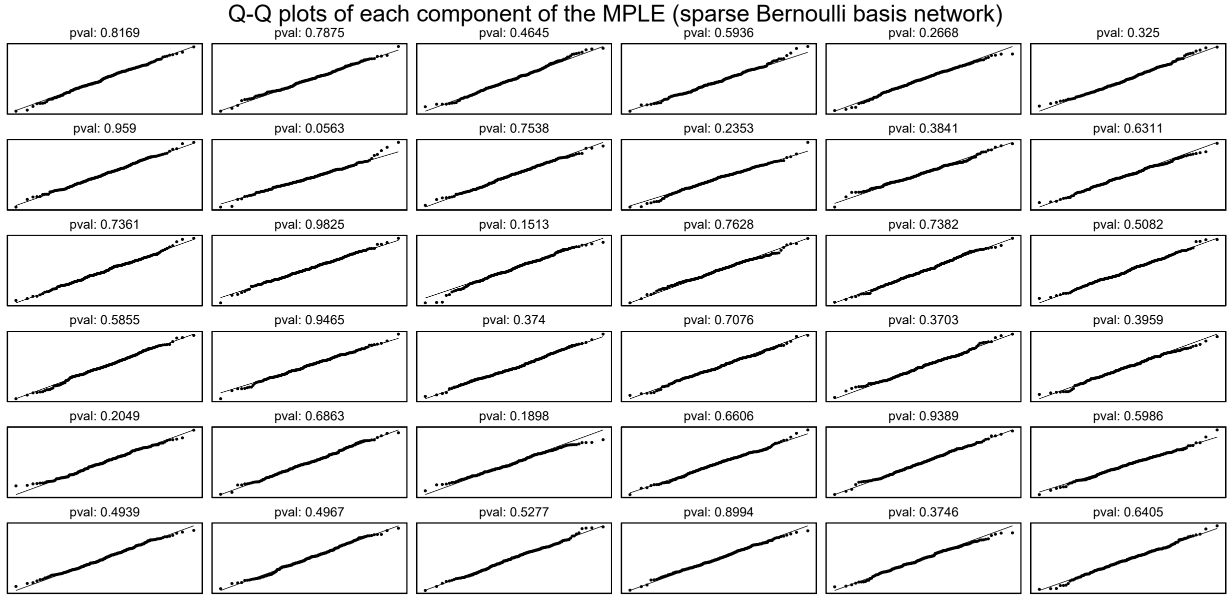

Figure 7: Q-Q plots and p-values of six components of $\widetilde{\boldsymbol{\theta}}$ estimated from 250 multilayer network samples at size 1000 on the sparse Bernoulli basis network for 6 model-generating parameters on each row.

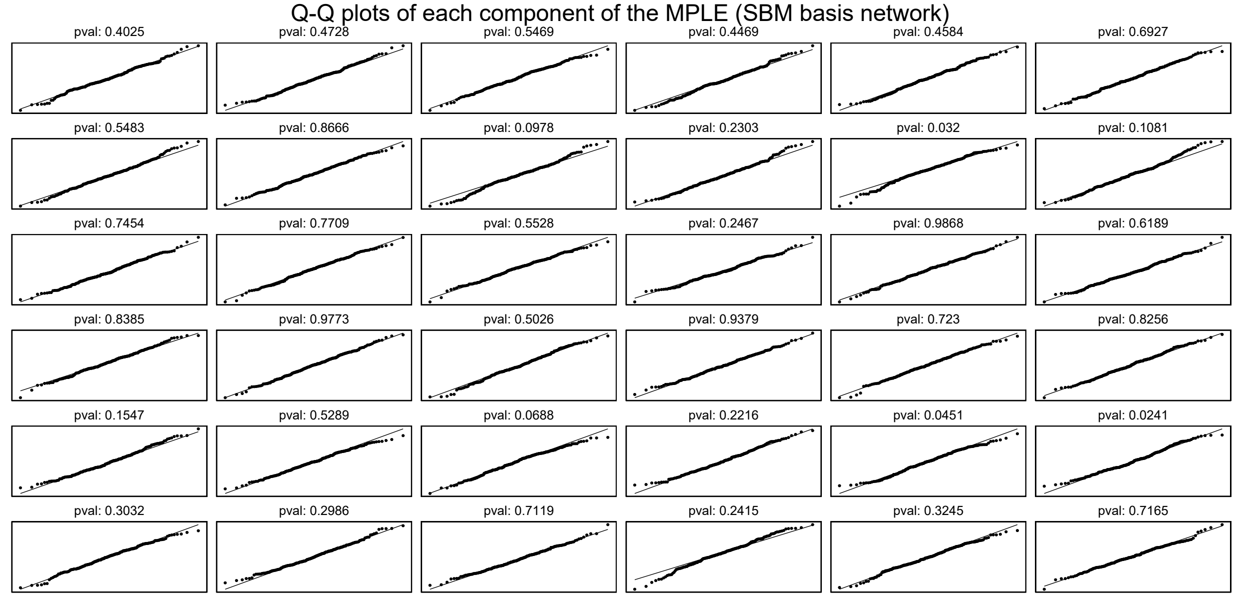

Figure 8: Q-Q plots and p-values of six components of $\widetilde{\boldsymbol{\theta}}$ estimated from 250 multilayer network samples at size 1000 on the SBM generated basis network for 6 model-generating parameters on each row.

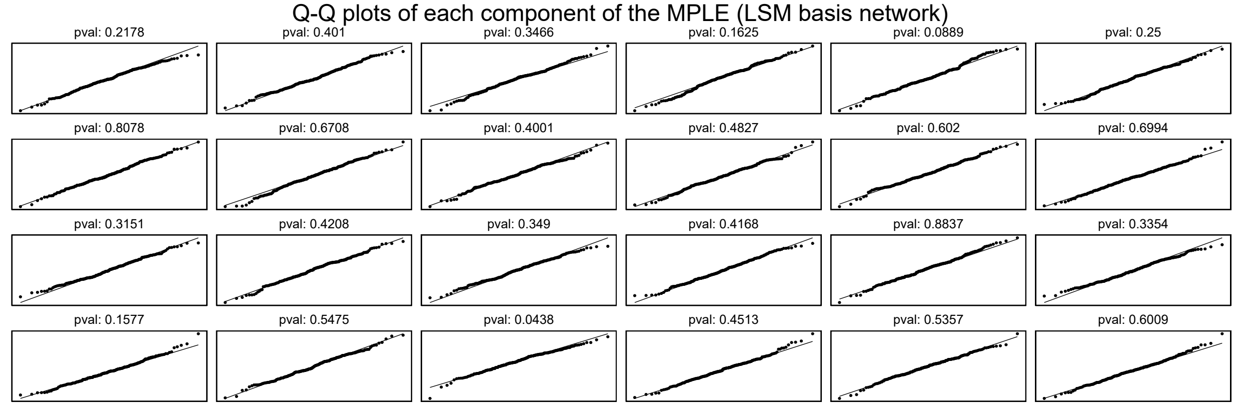

Figure 9: Q-Q plots and p-values of six components of $\widetilde{\boldsymbol{\theta}}$ estimated from 250 multilayer network samples at size 1000 on the LSM generated basis network for 6 model-generating parameters on each row.

Table 5: False discovery rates of four procedures for detecting non-zero effects of six model-generating parameters ($\boldsymbol{\theta}_1^\star$, $\boldsymbol{\theta}_2^\star$, $\boldsymbol{\theta}_3^\star$, $\boldsymbol{\theta}_4^\star$, $\boldsymbol{\theta}_5^\star$, $\boldsymbol{\theta}_6^\star$) estimated from 250 multilayer network samples at size 1000 on the sparse Bernoulli basis network. All FDRs are smaller than 0.05.

Procedure	$\boldsymbol{\theta}_1^\star$	$\boldsymbol{\theta}_2^\star$	$\boldsymbol{\theta}_3^\star$	$\boldsymbol{\theta}_4^\star$	$\boldsymbol{\theta}_5^\star$	$\boldsymbol{\theta}_6^\star$
Bonferroni	.002	.003	.003	.003	.003	.011
Benjamini-Hochberg	.020	.011	.022	.022	.014	.017
Hochberg's	.009	.008	.012	.010	.010	.014
Holm's	.007	.008	.011	.009	.006	.014

G.2 Additional results on the false discovery rate

The false discovery rate (FDR) of the multiple testing correction procedures of Bonferroni, Benjamini-Hochberg, Hochberg, and Holm to detect non-zero components of $\boldsymbol{\theta}^\star$ at a family-wise significance level of $\alpha = 0.05$ with a sparse Bernoulli basis network, an SBM generated basis network and an LSM generated basis network are provided in Table 5, 6 and 7, respectively (recall that the third and the sixth component $\theta_{1,3}^\star$ and $\theta_3^\star$ of $\boldsymbol{\theta}^\star$ are set to 0).

Table 6: False discovery rates of four procedures for detecting non-zero effects of six model-generating parameters ($\boldsymbol{\theta}_1^\star$, $\boldsymbol{\theta}_2^\star$, $\boldsymbol{\theta}_3^\star$, $\boldsymbol{\theta}_4^\star$, $\boldsymbol{\theta}_5^\star$, $\boldsymbol{\theta}_6^\star$) estimated from 250 multilayer network samples at size 1000 on the SBM generated basis network. All FDRs are smaller than 0.05.

Procedure	$\boldsymbol{\theta}_1^\star$	$\boldsymbol{\theta}_2^\star$	$\boldsymbol{\theta}_3^\star$	$\boldsymbol{\theta}_4^\star$	$\boldsymbol{\theta}_5^\star$	$\boldsymbol{\theta}_6^\star$
Bonferroni	.002	.002	.003	.001	.001	.004
Benjamini-Hochberg	.022	.013	.014	.015	.015	.018
Hochberg's	.009	.014	.01	.008	.011	.014
Holm's	.009	.013	.005	.009	.009	.011

Table 7: False discovery rates of four procedures for detecting non-zero effects of six model-generating parameters ($\boldsymbol{\theta}_1^\star$, $\boldsymbol{\theta}_2^\star$, $\boldsymbol{\theta}_3^\star$, $\boldsymbol{\theta}_4^\star$, $\boldsymbol{\theta}_5^\star$, $\boldsymbol{\theta}_6^\star$) estimated from 250 multilayer network samples at size 1000 on the LSM generated basis network. All FDRs are smaller than 0.05.

Procedure	$\boldsymbol{\theta}_1^\star$	$\boldsymbol{\theta}_2^\star$	$\boldsymbol{\theta}_3^\star$	$\boldsymbol{\theta}_4^\star$	$\boldsymbol{\theta}_5^\star$	$\boldsymbol{\theta}_6^\star$
Bonferroni	.004	.006	.000	.005	.003	.004
Benjamini-Hochberg	.016	.013	.011	.015	.016	.017
Hochberg's	.009	.014	.009	.011	.010	.011
Holm's	.008	.014	.009	.011	.007	.010

References

E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

J. A. Arroyo, A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. *The Journal of Machine Learning Research*, 22(1):6303–6351, 2021.

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, J. T. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, D. L. Sussman, E. Fishkind, and Y. Park. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18:1–92, 2018.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–225, 1974.

S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. *The Annals of Applied Probability*, 21:2146–2170, 2011.

P. Block. Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40: 163–173, 2015.

C. T. Butts. A dynamic process interpretation of the sparse ERGM reference model. *Journal of Mathematical Sociology*, 2020.

D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity. *Advances in Neural Information Processing Systems*, 29, 2016.

A. Caimo and I. Gollini. A multilayer exponential random graph modelling approach for weighted networks. *Computational Statistics & Data Analysis*, 142:106825, 2020.

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

S. Chen, S. Liu, and Z. Ma. Global and individualized community detection in inhomogeneous multilayer networks. *The Annals of Statistics*, 50(5):2664–2693, 2022.

H. Crane and W. Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018.

O. Frank. Transitivity in stochastic graphs and digraphs. *Journal of Mathematical Sociology*, 7:199–213, 1980.

M. Furi and M. Martelli. On the mean value theorem, inequality, and inclusion. *The American Mathematical Monthly*, 98(9):840–846, 1991.

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

S. M. Goodreau, J. A. Kitts, and M. Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103–125, 2009.

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

P. W. Holland and S. Leinhardt. Some evidence on the transitivity of positive interpersonal sentiment. *American Journal of Sociology*, 77:1205–1209, 1972.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76:33–65, 1981.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic block models: some first steps. *Social Networks*, 5:109–137, 1983.

S. Huang, H. Weng, and Y. Feng. Spectral clustering via adaptive layer aggregation for multi-layer networks. *Journal of Computational and Graphical Statistics*, pages 1–15, 2022.

D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15:565–583, 2006.

D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103:248–258, 2008.

R.A. Johnson and D.W. Wichern. *Applied multivariate statistical analysis.* Prentice hall, 2002.

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

P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31:204–213, 2009.

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

P. N. Krivitsky, L. M. Koehly, and C. S. Marcum. Exponential-family random graph models for multi-layer networks. *Psychometrika*, 85(3):630–659, 2020.

E. Lazega. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership.* Oxford University Press, 2001.

J. Lei, K. Chen, and B. Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 2020.

W. Li, Y. Xu, J Yang, and Z. Tang. Finding structural patterns in complex networks. In *2012 IEEE Fifth International Conference on Advanced Computational Intelligence*, pages 23–27, 2012.

D. Lusher, J. Koskinen, and G. Robins. *Exponential Random Graph Models for Social Networks*. Cambridge University Press, Cambridge, UK, 2013.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright. *Handbook of graphical models*. CRC Press, 2018.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

M. Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A): 2824–2853, 2019.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *The Annals of Statistics*, 39:1878–1915, 2011.

M. Schweinberger, P. N. Krivitsky, C. T. Butts, and Jonathan Stewart. Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios. *Statistical Science*, 35:627–662, 2020.

D. K. Sewell and Y. Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110:1646–1657, 2015.

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

J. Sosa and B. Betancourt. A latent space model for multilayer network data. *Computational Statistics & Data Analysis*, 169:107432, 2022.

J. Stewart, M. Schweinberger, M. Bojanowski, and M. Morris. Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. *Social Networks*, 59:98–119, 2019.

J. R. Stewart and M. Schweinberger. Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension. *arXiv preprint arXiv:2012.07167*, 2021.

D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85:204–212, 1990.

R. Sundberg. *Statistical modelling by exponential families*, volume 12. Cambridge University Press, 2019.

D. Sussman, M. Tang, and C. Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:48–57, 2014.

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

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

M. Zhou and Y. Shao. A powerful test for multivariate normality. *Journal of Applied Statistics*, 41(2):351–363, 2014.