CONCENTRATION AND CONSISTENCY RESULTS FOR CANONICAL AND CURVED EXPONENTIAL-FAMILY MODELS OF RANDOM GRAPHS

BY MICHAEL SCHWEINBERGER AND JONATHAN STEWART

Rice University

Statistical inference for exponential-family models of random graphs with dependent edges is challenging. We stress the importance of additional structure and show that additional structure facilitates statistical inference. A simple example of a random graph with additional structure is a random graph with neighborhoods and local dependence within neighborhoods. We develop the first concentration and consistency results for maximum likelihood and M-estimators of a wide range of canonical and curved exponential-family models of random graphs with local dependence. All results are non-asymptotic and applicable to random graphs with finite populations of nodes, although asymptotic consistency results can be obtained as well. In addition, we show that additional structure can facilitate subgraph-to-graph estimation, and present concentration results for subgraph-to-graph estimators. As an application, we consider popular curved exponential-family models of random graphs, with local dependence induced by transitivity and parameter vectors whose dimensions depend on the number of nodes.

1. Introduction. Models of network data have witnessed a surge of interest in statistics and related areas [e.g., 31]. Such data arise in the study of, e.g., social networks, epidemics, insurgencies, and terrorist networks.

Since the work of Holland and Leinhardt in the 1970s [e.g., 21], it is known that network data exhibit a wide range of dependencies induced by transitivity and other interesting network phenomena [e.g., 39]. Transitivity is a form of triadic closure in the sense that, when a node k is connected to two distinct nodes i and j, then i and j are likely to be connected as well, which suggests that edges are dependent [e.g., 39]. A large statistical framework for modeling dependencies among edges is given by discrete exponential-family models of random graphs, called exponential-family random graphs [e.g., 15, 57, 18, 53, 24, 39, 20, 36]. Such models are popular among network scientists for the same reason Ising models are popular among physicists: Both classes of models enable scientists to model a wide range of dependencies of scientific interest [e.g., 39].

Despite the appeal of the discrete exponential-family framework and its rela-

*Supported by NSF awards DMS-1513644 and DMS-1812119.

MSC 2010 subject classifications: curved exponential families, exponential families, exponential-family random graph models, M-estimators, multilevel networks, social networks.

tionship to other discrete exponential-family models for dependent random variables [e.g., Ising models and discrete Markov random fields, 9, 42, 3], statistical inference for exponential-family random graphs is challenging. One reason is that some exponential-family random graphs are ill-behaved [e.g., the so-called triangle model, 28, 18, 47, 2, 10, 6], though well-behaved alternatives have been developed, among them curved exponential-family random graphs [53, 24]. A second reason is that in most applications of exponential-family random graphs statistical inference is based on a single observation of a random graph with dependent edges. Establishing desirable properties of estimators, such as consistency, is non-trivial when no more than one observation of a random graph with dependent edges is available. While some consistency results have been obtained under independence assumptions [12, 58, 44, 51, 40, 35, 60, 59] and restrictive dependence assumptions [58, 51, 40]—as discussed in Section 5—the existing consistency results do not cover the models most widely used in practice [e.g., 39]: canonical and curved exponential-family random graphs with dependence among edges induced by transitivity and other interesting network phenomena [39].

We stress the importance of additional structure and show that additional structure facilitates statistical inference. We consider here a simple and common form of additional structure, called multilevel structure. Network data with multilevel structure are popular in network science, as the recent monograph of Lazega and Snijders [37] and a growing number of applications demonstrate [e.g., 56, 62, 38, 52, 22, 54]. A simple form of multilevel structure is given by a partition of a population of nodes into subsets of nodes, called neighborhoods. In applications, neighborhoods may correspond to school classes within schools, departments within companies, and units of armed forces. It is worth noting that in multilevel networks the partition of the population of nodes is observed and models of multilevel networks attempt to capture dependencies within and between neighborhoods [e.g., 56, 62, 38, 52, 22, 54], whereas the well-known class of stochastic block models [41] assumes that the partition is unobserved and that edges are independent conditional on the partition.

Additional structure in the form of multilevel structure offers opportunities in terms of statistical inference. We take advantage of these opportunities to develop the first statistical theory which shows that statistical inference for many canonical and curved exponential-family random graphs with dependent edges is possible. The main idea is based on a simple and general exponential-family argument that may be of independent interest. It helps establish non-asymptotic probability statements about estimators of canonical and curved exponential families for dependent random variables under weak conditions, as long as additional structure helps control the amount of dependence induced by the model and the sufficient statistics are sufficiently smooth functions of the random variables. We exploit the

main idea to develop the first concentration and consistency results for maximum likelihood and M-estimators of canonical and curved exponential-family random graphs with dependent edges, under correct and incorrect model specifications. All results are non-asymptotic and applicable to random graphs with finite populations of nodes, although asymptotic consistency results can be obtained as well. In addition, we show that multilevel structure facilitates subgraph-to-graph estimation, and present concentration results for subgraph-to-graph estimators. As an application, we consider popular curved exponential-family random graphs [53, 24], with local dependence induced by transitivity and parameter vectors whose dimensions depend on the number of nodes.

These concentration and consistency results have important implications, both in statistical theory and practice:

- The most important implication is that statistical inference for transitivity and other network phenomena of great interest to network scientists is possible. To date, it has been widely believed that statistical inference for transitivity based on exponential-family random graphs is challenging [e.g., 51, 10], but additional structure in the form of multilevel structure facilitates it.
- Network scientists can benefit from collecting network data with multilevel structure, because multilevel structure can facilitate statistical inference for exponential-family random graphs with dependent edges.

Last, but not least, it is worth noting that these concentration and consistency results cover two broad inference scenarios:

- *Inference scenarios with finite populations of nodes.* In many applications of exponential-family random graphs, there is a finite population of nodes and a population graph is assumed to have been generated by an exponential-family random graph model. A common goal of statistical inference, then, is to estimate the parameters of the data-generating exponential-family random graph model based on a complete or incomplete observation of the population graph. Our concentration results cover inference scenarios with finite populations of nodes, when the whole population graph is observed or when neighborhoods are sampled and the subgraphs induced by the sampled neighborhoods are observed.
- *Inference scenarios with populations of nodes growing without bound.* In addition, our concentration results can be used to obtain asymptotic consistency results by allowing the number of neighborhoods to grow without bound. The resulting asymptotic consistency results resemble asymptotic consistency results in other areas of statistics, albeit with two notable differences: first, the units of statistical analysis are subsets of nodes (neighborhoods) rather than nodes or edges; and, second, the sizes of the units need not be

identical, but are similar in a well-defined sense.

Since the first application is more interesting than the second one, we state all results with the first application in mind, i.e., all results focus on random graphs with finite populations of nodes, although we do mention some asymptotic consistency results along the way.

The remainder of our paper is structured as follows. Section 2 introduces models. Section 3 describes concentration and consistency results for maximum likelihood and M-estimators, under correct and incorrect model specifications. Section 4 shows that multilevel structure facilitates subgraph-to-graph estimation. A comparison with existing consistency results can be found in Section 5. Section 6 presents simulation results.

2. Exponential-family random graphs with multilevel structure. We introduce exponential-family random graphs with multilevel structure.

A simple and common form of multilevel structure is a partition of a population of nodes into $K \geq 2$ non-empty subsets of nodes $\mathcal{A}_1, \ldots, \mathcal{A}_K$, called neighborhoods. We note that in multilevel networks the partition of the population of nodes is observed [e.g., 56, 62, 38, 52, 22, 54] and that some neighborhoods may be larger than others. We consider random graphs with undirected edges that may be either absent or present or may have weights, where the weights are elements of a countable set. Extensions to random graphs with directed edges are straightforward. Let $\boldsymbol{X} = (\boldsymbol{X}_k)_{k=1}^K$ and $\boldsymbol{Y} = (\boldsymbol{Y}_{k,l})_{k<l}^K$ be sequences of within- and between-neighborhood edge variables based on a sequence of neighborhoods $\mathcal{A}_1, \ldots, \mathcal{A}_K$, where $\boldsymbol{X}_k = (X_{i,j})_{i\in\mathcal{A}_k<j\in\mathcal{A}_k}$ and $\boldsymbol{Y}_{k,l} = (Y_{i,j})_{i\in\mathcal{A}_k, j\in\mathcal{A}_l}$ $(k < l)$ correspond to within- and between-neighborhood edge variables $X_{i,j} \in \mathbb{X}_{i,j}$ and $Y_{i,j} \in \mathbb{Y}_{i,j}$, taking on values in countable sets $\mathbb{X}_{i,j}$ and $\mathbb{Y}_{i,j}$, respectively. We exclude self-edges, assuming that $X_{i,i} = 0$ holds with probability 1 ($i \in \mathcal{A}_k$, $k = 1, \ldots, K$), and write $\mathbb{X}_k = \prod_{i\in\mathcal{A}_k<j\in\mathcal{A}_k} \mathbb{X}_{i,j}$, $\mathbb{X} = \prod_{k=1}^K \prod_{i\in\mathcal{A}_k<j\in\mathcal{A}_k} \mathbb{X}_{i,j}$, and $\mathbb{Y} = \prod_{k<l}^K \prod_{i\in\mathcal{A}_k, j\in\mathcal{A}_l} \mathbb{Y}_{i,j}$.

We assume that within-neighborhood edges $\boldsymbol{X}$ are independent of between-neighborhood edges $\boldsymbol{Y}$, i.e.,

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{X}, \boldsymbol{Y} \in \mathcal{Y}) \quad = \quad \mathbb{P}(\boldsymbol{X} \in \mathcal{X})\,\mathbb{P}(\boldsymbol{Y} \in \mathcal{Y}) \quad \text{for all} \quad \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{X} \times \mathbb{Y},$$

where $\mathbb{P}$ denotes a probability distribution with support $\mathbb{X} \times \mathbb{Y}$. We do not assume that edges are independent, neither within nor between neighborhoods.

While in principle both within-neighborhood edge variables $\boldsymbol{X}$ and between-neighborhood edge variables $\boldsymbol{Y}$ may be of interest, we focus on within-neighborhood edge variables, which are of primary interest in applications [e.g., 37, 56, 62, 38, 52, 22, 54]. We therefore restrict attention to the probability law

of $\boldsymbol{X}$ and do not make assumptions about the probability law of $\boldsymbol{Y}$. We assume that the parameter vectors of the probability laws of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are variation-independent, i.e., the parameter space is a product space, so that statistical inference concerning the parameter vector of the probability law of $\boldsymbol{X}$ can be based on $\boldsymbol{X}$ without requiring knowledge of $\boldsymbol{Y}$.

The distribution of within-neighborhood edge variables $\boldsymbol{X}$ is presumed to belong to an exponential family with local dependence, defined as follows.

Definition. Exponential family with local dependence. An exponential family with local dependence is an exponential family of distributions with countable support $\mathbb{X}$, having densities with respect to counting measure of the form

$$\begin{aligned} p_{\boldsymbol{\eta}}(\boldsymbol{x}) &= \exp\left(\langle \boldsymbol{\eta},\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\eta})\right)\, \nu(\boldsymbol{x}) \\ &= \exp\left(\sum_{k=1}^{K} \langle \boldsymbol{\eta}_k,\, s_k(\boldsymbol{x}_k)\rangle - \psi(\boldsymbol{\eta})\right)\, \nu(\boldsymbol{x}), \end{aligned} \tag{2.1}$$

where

$$\psi(\boldsymbol{\eta}) = \log \sum_{\boldsymbol{x}_1 \in \mathbb{X}_1} \cdots \sum_{\boldsymbol{x}_K \in \mathbb{X}_K} \exp\left(\sum_{k=1}^{K} \langle \boldsymbol{\eta}_k,\, s_k(\boldsymbol{x}_k)\rangle\right)\, \nu(\boldsymbol{x})$$

and $\nu(\boldsymbol{x}) = \prod_{k=1}^{K} \nu_k(\boldsymbol{x}_k)$.

In other words, edges may depend on other edges in the same neighborhood, but do not depend on edges in other neighborhoods [48]. Here, $\langle \boldsymbol{\eta},\, s(\boldsymbol{x})\rangle = \sum_{k=1}^{K} \langle \boldsymbol{\eta}_k,\, s_k(\boldsymbol{x}_k)\rangle$ is the inner product of a natural parameter vector $\boldsymbol{\eta} \in \mathbb{R}^m$ and a sufficient statistic vector $s : \mathbb{X} \mapsto \mathbb{R}^m$ while $\boldsymbol{\eta}_k \in \mathbb{R}^{m_k}$ and $s_k : \mathbb{X}_k \mapsto \mathbb{R}^{m_k}$ denote the natural parameter vector and sufficient statistic vector of neighborhood $\mathcal{A}_k$, respectively ($k = 1, \ldots, K$). The functions $\nu : \mathbb{X} \mapsto \mathbb{R}^+ \cup \{0\}$ and $\nu_k : \mathbb{X}_k \mapsto \mathbb{R}^+ \cup \{0\}$ ($k = 1, \ldots, K$) along with the sample space $\mathbb{X}$ determine the reference measure of the exponential family. A careful discussion of the choice of reference measure can be found in Krivitsky [33].

We consider here a wide range of exponential families with local dependence. A specific example of an exponential family with local dependence can be found in Section 3.3. In the following, we introduce selected exponential-family terms in order to distinguish exponential families from subfamilies of exponential families. Subfamilies of exponential families give rise to distinct theoretical challenges, and thus require a separate treatment. We therefore introduce the classic exponential-family notions of full and non-full exponential families, canonical and curved exponential families, and minimal exponential families. These exponential-family terms are taken from the monographs on exponential families by Barndorff-Nielsen

[1] and Brown [5] and have been used in other recent works as well: see, e.g., Lauritzen et al. [36], Rinaldo et al. [43], and Geyer [16]. To help ensure that parameters are identifiable, we assume that exponential families of the form (2.1) are minimal in the sense of Barndorff-Nielsen [1] and Brown [5], i.e., the closure of the convex hull of the set $\{s(\boldsymbol{x}) : \nu(\boldsymbol{x}) > 0\}$ is not contained in a proper affine subspace of $\mathbb{R}^m$ [e.g., 5, p. 2]. It is well-known that all non-minimal exponential families can be reduced to minimal exponential families [e.g., 5, Theorem 1.9, p. 13]. We consider both full and non-full exponential families of the form (2.1). An exponential family $\{\mathbb{P}_{\boldsymbol{\eta}}, \boldsymbol{\eta} \in \boldsymbol{\Xi}\}$ is full if $\boldsymbol{\Xi} = \mathfrak{N}$ and non-full if $\boldsymbol{\Xi} \subset \mathfrak{N}$, where $\mathfrak{N} = \{\boldsymbol{\eta} \in \mathbb{R}^m : \psi(\boldsymbol{\eta}) < \infty\}$ is the natural parameter space, i.e., the largest set of possible values the natural parameter vector $\boldsymbol{\eta}$ can take on. While full exponential families may be more convenient on mathematical grounds, non-full exponential families—the most important example being curved exponential families [e.g., 14, 29]—offer parsimonious parameterizations of exponential families where the dimension m_k of neighborhood-dependent natural parameter vectors $\boldsymbol{\eta}_k$ is an increasing function of the number of nodes in neighborhoods $\mathcal{A}_k$ ($k = 1, \ldots, K$), and have turned out to be useful in practice [53, 24]. A simple approach to generating non-full exponential families is to assume that $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is a known function of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^q : \psi(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}$, $\text{int}(\boldsymbol{\Theta})$ and $\text{int}(\mathfrak{N})$ denote the interiors of $\boldsymbol{\Theta}$ and $\mathfrak{N}$, respectively, and $q \leq m$. It is convenient to distinguish exponential families that can be reduced to canonical exponential families with natural parameter vectors of the form $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ from those that cannot. An exponential family can be reduced to a canonical exponential family with $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ when the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is affine. In other words, if $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{A}\,\boldsymbol{\theta} + \boldsymbol{b}$ with $\boldsymbol{A} \in \mathbb{R}^{m \times q}$ and $\boldsymbol{b} \in \mathbb{R}^m$, then the exponential family can be reduced to a canonical exponential family with $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ by absorbing $\boldsymbol{A}$ into the sufficient statistic vector and $\boldsymbol{b}$ into the reference measure. We therefore call all exponential families with affine maps $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ canonical exponential families, and call all exponential families with non-affine maps $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ curved exponential families. We note that our definition of a curved exponential family is broader than the one used in Efron [13, 14], Brown [5, pp. 81–84], and Kass and Vos [29]. The main reason is that we do not restrict the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ to be differentiable, because our concentration and consistency results in Sections 3 and 4 do not require differentiability.

Throughout, we assume that the neighborhoods are of the same order of magnitude and that the neighborhood-dependent natural parameters $\eta_{k,i}(\boldsymbol{\theta})$ are of the form $\eta_{k,i}(\boldsymbol{\theta}) = \eta_i(\boldsymbol{\theta})$ ($i = 1, \ldots, m_k$, $k = 1, \ldots, K$). We define neighborhoods of the same order of magnitude as follows.

Definition. Neighborhoods of the same order of magnitude. A sequence of neighborhoods $\mathcal{A}_1, \ldots, \mathcal{A}_K$ is of the same order of magnitude if there exists a universal constant $A > 1$ such that $\max_{1 \leq k \leq K} |\mathcal{A}_k| \leq A \min_{1 \leq k \leq K} |\mathcal{A}_k|$ $(K = 1, 2, \ldots)$.

In other words, the largest neighborhood size is a constant multiple of the smallest neighborhood size, so that the sizes of neighborhoods may not be identical, but are similar in a well-defined sense. The definition is satisfied when the sizes of neighborhoods are bounded above. When the number of neighborhoods K grows and the sizes of neighborhoods grow with K, then the definition implies that the sizes of neighborhoods grow at the same rate. We note that when the neighborhoods are not of the same order of magnitude, the natural parameters of neighborhoods may have to depend on the order of magnitude of neighborhoods [e.g., 34, 35, 7], because there are good reasons to believe that small and large within-neighborhood subgraphs are not governed by the same natural parameters [51, 11, 36]. Size-dependent parameterizations have an important place in the exponential-family random graph framework, and some promising size-dependent parameterizations have been proposed. Most of them assume that natural parameters consist of size-invariant parameters and size-dependent deviations. The size-dependent deviations may be size-dependent offsets [e.g., 34, 35, 7] or functions of size-dependent covariates [52]. Some of those size-dependent deviations can be absorbed into the sufficient statistic vector and reference measure, and are hence covered by our main concentration and consistency results in Sections 3 and 4. However, the topic of size-dependent parameterizations is an important topic in its own right, and deserves a separate treatment that is beyond the scope of our paper.

The assumption $\eta_{k,i}(\boldsymbol{\theta}) = \eta_i(\boldsymbol{\theta})$ $(i = 1, \ldots, m_k, k = 1, \ldots, K)$ implies that the exponential families considered here can be reduced to exponential families with natural parameter vectors of the form

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \quad = \quad (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta}))$$

and sufficient statistic vectors of the form

$$s(\boldsymbol{x}) \quad = \quad (s_1(\boldsymbol{x}), \ldots, s_m(\boldsymbol{x}))\,,$$

where $s_i(\boldsymbol{x}) = \sum_{k=1}^{K} s_{k,i}(\boldsymbol{x}_k)$ $(i = 1, \ldots, m)$ and $m = \max_{1 \leq k \leq K} m_k$. We assume that the dimensions m_k of the neighborhood-dependent natural parameter vectors $\boldsymbol{\eta}_k(\boldsymbol{\theta})$ are non-decreasing functions of the sizes $|\mathcal{A}_k|$ of neighborhoods $\mathcal{A}_k$ $(k = 1, \ldots, K)$, which implies that $m = \max_{1 \leq k \leq K} m_k$ is a non-decreasing function of $\|\mathcal{A}\|_\infty = \max_{1 \leq k \leq K} |\mathcal{A}_k|$. The dimensions m_k $(k = 1, \ldots, K)$ and m do not depend on other quantities. The dimension q of parameter vector $\boldsymbol{\theta}$ satisfies $q \leq m$, as mentioned above.

Notation. To prepare the ground for the concentration and consistency results in Sections 3 and 4, we introduce mean-value parameterizations of exponential families along with additional notation. Mean-value parameterizations facilitate concentration and consistency results, because concentration inequalities [4] bound probabilities of deviations from means and the mean-value parameters of an exponential family are the means of the sufficient statistics, defined by $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\, s(\boldsymbol{X}) \in \text{rint}(\mathbb{M})$ [5, p. 2 and p. 75]. Here, $\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\, s(\boldsymbol{X})$ is the expectation of $s(\boldsymbol{X})$ with respect to exponential-family distributions $\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}$ having densities of the form (2.1), $\mathbb{M}$ is the convex hull of the set $\{s(\boldsymbol{x}) : \nu(\boldsymbol{x}) > 0\}$, and $\text{rint}(\mathbb{M})$ is the relative interior of $\mathbb{M}$. We denote the data-generating parameter vector by $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$ and write $\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}$ and $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}$. An open ball in $\mathbb{R}^v$ ($v \geq 1$) with center $\boldsymbol{c} \in \mathbb{R}^v$ and radius $\rho > 0$ is denoted by $\mathcal{B}(\boldsymbol{c},\, \rho)$. We write $\|.\|_1$, $\|.\|_2$, and $\|.\|_\infty$ to refer to the ℓ_1-, ℓ_2-, and ℓ_∞-norm of vectors, respectively. Throughout, uppercase letters $A, B, C, C_1, C_2, \cdots > 0$ denote universal constants, which are independent of all other quantities of interest and may be recycled from line to line. The function $d : \mathbb{X} \times \mathbb{X} \mapsto \{0, 1, \dots\}$ denotes the Hamming metric, defined by

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) \quad = \quad \sum_{k=1}^{K} \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} \mathbb{1}(x_{1,i,j} \neq x_{2,i,j}), \quad (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X},$$

where $\mathbb{1}(x_{1,i,j} \neq x_{2,i,j})$ is an indicator function, which is 1 if $x_{1,i,j} \neq x_{2,i,j}$ and is 0 otherwise.

3. Concentration and consistency results: maximum likelihood and M-estimators. In many applications of exponential-family random graphs, the parameter vector of primary interest is $\boldsymbol{\theta}$. To estimate the parameter vector $\boldsymbol{\theta}$ of a wide range of full and non-full, curved exponential families under weak assumptions on the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$, we consider an estimating function [17]—a function $g : \boldsymbol{\Theta} \times \mathbb{X} \mapsto \mathbb{R}$ of both $\boldsymbol{\theta}$ and $\boldsymbol{X}$—of the form

$$g(\boldsymbol{\theta};\, \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) \quad = \quad \|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{3.1}$$

which is an approximation of

$$g(\boldsymbol{\theta};\, \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))) \quad = \quad \|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

where $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} = s(\boldsymbol{X})$ is an estimator of the data-generating mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}\, s(\boldsymbol{X}) \in \text{rint}(\mathbb{M})$. The fact that $g(\boldsymbol{\theta};\, \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))})$ is an approximation of $g(\boldsymbol{\theta};\, \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)))$ follows from the triangle inequality, which shows that

$$\left|g(\boldsymbol{\theta};\, \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) - g(\boldsymbol{\theta};\, \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)))\right| \quad \leq \quad \|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

along with the fact that, under suitable conditions, $\|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2$ is small with high probability, as we will show in Proposition 2 in Section 3.2.

Estimating function (3.1) has at least three advantages. First, estimating function (3.1) admits concentration and consistency results under weak assumptions on the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$. Indeed, the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ satisfies the main assumptions of Section 3 as long as the map is one-to-one and continuous, but it need not be differentiable. The weakness of these assumptions implies that the main results of Section 3 cover a vast range of full and non-full exponential families—including, but not limited to curved exponential families—and it is possible to verify these assumptions in some of the most challenging applications, as demonstrated in Section 3.3. Second, estimating function (3.1) is natural, because maximum likelihood estimation of the data-generating natural parameter vector $\boldsymbol{\eta}^\star$ of an exponential family with natural parameter vector $\boldsymbol{\eta}$ and sufficient statistic vector $s(\boldsymbol{x})$ can be based on the gradient of the loglikelihood function $\nabla_{\boldsymbol{\eta}} \log p_{\boldsymbol{\eta}}(\boldsymbol{x}) = \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}^\star)} - \boldsymbol{\mu}(\boldsymbol{\eta})$ provided $\boldsymbol{\eta} \in \text{int}(\mathfrak{N})$, where $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}^\star)} = s(\boldsymbol{x})$ [5, Lemma 5.3, p. 146]. Therefore, maximum likelihood estimation of $\boldsymbol{\eta}^\star$ can be based on estimating functions of the form $\|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}^\star)} - \boldsymbol{\mu}(\boldsymbol{\eta})\|_2$. By the parameterization-invariance of maximum likelihood estimators, maximum likelihood estimation of functions of $\boldsymbol{\eta}^\star$, such as $\boldsymbol{\theta}^\star$, can be based on $\|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2$ provided the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one. We note that estimating function (3.1) is chosen for mathematical convenience, facilitating concentration and consistency results for maximum likelihood estimators of many full and non-full, curved exponential families under weak assumptions on the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$, and is not chosen for computational convenience. Last, but not least, the simple form of estimating function (3.1) helps determine when minimizers of (3.1) exist and are unique, and how the minimizers are related to each other when there is more than one minimizer. These advantages are most useful in non-full exponential families, in particular curved exponential families.

In the following, we assume that the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one. A natural class of estimators is hence given by

$$\widehat{\boldsymbol{\theta}} \;=\; \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \;\; g(\boldsymbol{\theta};\; \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) \;=\; \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} \; g(\dot{\boldsymbol{\theta}};\; \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) \right\}.$$

If the set $\widehat{\boldsymbol{\theta}}$ is non-empty, it may contain one element (e.g., in full exponential families) or more than one element (e.g., in non-full exponential families). If the set $\widehat{\boldsymbol{\theta}}$ contains more than one element, then all elements of the set $\widehat{\boldsymbol{\theta}}$ map to mean-value parameter vectors $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ that have the same ℓ_2-distance from $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}$ by construction of estimating function (3.1). In addition, if the set $\widehat{\boldsymbol{\theta}}$ is non-empty, then all minimizers $\widehat{\boldsymbol{\theta}}$ of $g(\boldsymbol{\theta};\; \widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))})$ are approximations of the minimizer of

$g(\boldsymbol{\theta};\ \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)))$. The minimizer of $g(\boldsymbol{\theta};\ \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)))$ is unique and is given by the data-generating parameter vector $\boldsymbol{\theta}^\star$ provided $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))) \in \text{rint}(\mathbb{M})$:

$$\boldsymbol{\theta}^\star \ = \ \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta}: \ g(\boldsymbol{\theta};\ \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))) \ = \ \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} \ g(\dot{\boldsymbol{\theta}};\ \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)))\right\}.$$

The data-generating parameter vector $\boldsymbol{\theta}^\star$ is the unique minimizer of $g(\boldsymbol{\theta};\ \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)))$ $= \|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2$, because $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$ and the map $\boldsymbol{\eta}: \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one by assumption, while the map $\boldsymbol{\mu}: \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one by classic exponential-family theory [5, Theorem 3.6, p. 74]. Therefore, $\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 = 0$ holds if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$.

In the remainder of the section, we show that the estimator $\widehat{\boldsymbol{\theta}}$ is close to the data-generating parameter vector $\boldsymbol{\theta}^\star$ with high probability under weak conditions. We first sketch the main idea in Section 3.1 and then discuss concentration and consistency results for maximum likelihood and M-estimators in Sections 3.3 and 3.4, respectively. An application to popular curved exponential-family random graphs is presented in Section 3.3.

3.1. *Main idea: A non-asymptotic approach to full and non-full, curved exponential families for dependent random variables.* Establishing concentration and consistency results for estimators of full and non-full, curved exponential-family random graphs with dependent edges is non-trivial, for at least three reasons. First, the map $\boldsymbol{\eta}: \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ may not be affine and may not be differentiable. Second, in many non-full exponential families, the mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is not available in closed form and there is no simple and known relationship between the mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ evaluated at $\widehat{\boldsymbol{\theta}}$ and the sufficient statistic vector $s(\boldsymbol{X})$. Third, concentration results for functions of edges, such as $s(\boldsymbol{X})$, are more challenging when edges are dependent rather than independent. As a result, studying the behavior of the estimating function $\|s(\boldsymbol{X}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))\|_2$ and its minimizer $\widehat{\boldsymbol{\theta}}$ is not straightforward.

Our main idea is based on a simple and general exponential-family argument that may be of independent interest. It helps establish non-asymptotic probability statements about estimators of full and non-full, curved exponential families for dependent random variables under weak conditions.

We make a single weak assumption: for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exist $\gamma > 0$ and $\delta > 0$ such that

$$\boldsymbol{\theta} \notin \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \implies \boldsymbol{\eta}(\boldsymbol{\theta}) \notin \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma) \implies \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) \notin \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta).$$

A graphical representation of the main assumption can be seen in Figure 1. A formal statement of the assumption can be found in Theorem 1 in Section 3.3, where γ and δ depend on ϵ and δ depends on the sizes of neighborhoods $\mathcal{A}_1, \ldots, \mathcal{A}_K$.

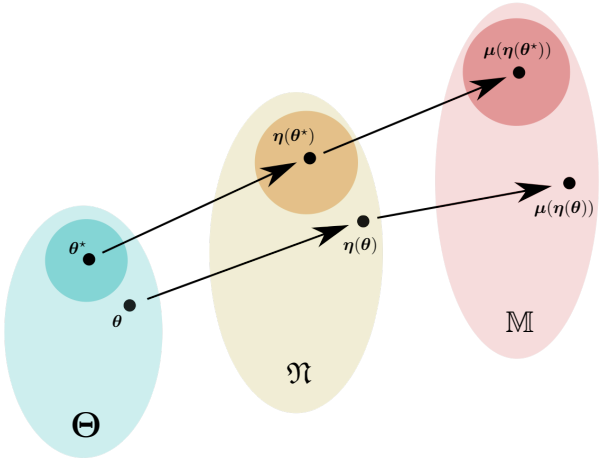

FIG 1. *The figure demonstrates the main assumption: for all* $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$*, there exist* $\gamma > 0$ *and* $\delta > 0$ *such that* $\boldsymbol{\theta} \notin \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ *implies* $\boldsymbol{\eta}(\boldsymbol{\theta}) \notin \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma)$ *and* $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) \notin \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta)$*.*

The main assumption is satisfied when the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one and continuous, but it need not be differentiable. Note that the map $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one and continuous by classic exponential-family theory [5, Theorem 3.6, p. 74].

The main assumption has a simple, but important implication. As no element of $\mathfrak{N}$ outside of the ball $\mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma)$ maps to an element of the ball $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta)$, any element $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta)$ must correspond to an element $\boldsymbol{\eta}(\boldsymbol{\theta})$ of $\mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma)$, which in turn must correspond to an element $\boldsymbol{\theta}$ of $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, so that

$$\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) \in \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta) \implies \boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma) \implies \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon).$$

As a result, the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ can be bounded by bounding the probability of event $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta)$.

A challenge, which complicates probability statements about the event $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta)$, is that in many non-full exponential families $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ is not available in closed form and there is no simple and known relationship between $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ and the sufficient statistic vector $s(\boldsymbol{X})$. To appreciate the difficulty of the problem, suppose that $s(\boldsymbol{x}) \in \text{rint}(\mathbb{M})$ is observed, so that $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} = s(\boldsymbol{x}) \in \text{rint}(\mathbb{M})$. The subset $\mathbb{M}(\boldsymbol{\Theta})$ of $\text{rint}(\mathbb{M})$ induced by $\boldsymbol{\Theta}$ is defined by

$$\mathbb{M}(\boldsymbol{\Theta}) = \{\boldsymbol{\mu}' \in \text{rint}(\mathbb{M}) : \text{ there exists } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ such that } \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \boldsymbol{\mu}'\}.$$

In full exponential families, $\mathbb{M}(\boldsymbol{\Theta}) = \text{rint}(\mathbb{M})$. Thus, there exists a minimizer $\widehat{\boldsymbol{\theta}} \in \text{int}(\boldsymbol{\Theta})$ of the estimating function $\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2$ satisfying $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) = s(\boldsymbol{x})$.

In fact, the minimizer is unique, because the maps $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ and $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ are one-to-one [5, Theorem 3.6, p. 74]. As a result, there is a simple and known relationship between $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ and $s(\boldsymbol{x})$. By contrast, in non-full exponential families, $\mathbb{M}(\boldsymbol{\Theta})$ is a proper subset of $\text{rint}(\mathbb{M})$, because non-full exponential families are subfamilies of exponential families that exclude some natural parameter vectors along with the corresponding mean-value parameter vectors. The problem, more often than not, is that it is unknown which mean-value parameter vectors are excluded, because the mean-value parameter vectors are not available in closed form. As a consequence, there is no simple and known relationship between $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ and $s(\boldsymbol{x})$, and it is not straightforward to determine where $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ is located in $\text{rint}(\mathbb{M})$, provided $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ is non-empty. So bounding the probability of event $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)), \delta)$ in non-full exponential families, in particular curved exponential families, is non-trivial.

But not all is lost. Despite the challenge of characterizing $\mathbb{M}(\boldsymbol{\Theta}) \subset \text{rint}(\mathbb{M})$ and hence $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \subset \mathbb{M}(\boldsymbol{\Theta})$, it can be shown that, under suitable conditions, the set $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \subset \mathbb{M}(\boldsymbol{\Theta})$ is non-empty and all elements of $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ are close to $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$ with high probability. Indeed, when the set $\widehat{\boldsymbol{\theta}}$ is non-empty, each element of the set $\widehat{\boldsymbol{\theta}}$ satisfies

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \quad \leq \quad \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))\|_2 + \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2.$$

While characterizing $\mathbb{M}(\boldsymbol{\Theta})$ is difficult, we do know one fact about $\mathbb{M}(\boldsymbol{\Theta})$: $\mathbb{M}(\boldsymbol{\Theta})$ contains the data-generating mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$, which implies that the ℓ_2-distance of $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ from $s(\boldsymbol{x})$ cannot exceed the ℓ_2-distance of $s(\boldsymbol{x})$ from $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \in \mathbb{M}(\boldsymbol{\Theta})$. Thus, we obtain the upper bound

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \quad \leq \quad 2\,\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2. \tag{3.2}$$

If the right-hand side of (3.2) can be shown to be small with high probability, then the problem of bounding the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ can be converted into the problem of bounding the probability of the event that $s(\boldsymbol{X})$ is close to $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}\, s(\boldsymbol{X}) \in \text{rint}(\mathbb{M})$ in a well-defined sense. All we need to bound the probabilities of those events are concentration results for the sufficient statistic vector $s(\boldsymbol{X})$, which can be established as long as (a) the dependence among edges is sufficiently weak; and (b) the sufficient statistics are sufficiently smooth functions of edges. Concentration results are facilitated by additional structure that helps control the amount of dependence induced by the model and the smoothness of sufficient statistics. We focus here on a simple form of additional structure in the form of multilevel structure, which controls the dependence among edges by constraining it to neighborhoods. But there are many other forms of additional structure that could help address (a) and (b), e.g., other forms of multilevel structure or spatial structure.

The most important implication, then, is that statistical inference for many exponential-family random graphs is possible and can be justified by statistical theory, provided a suitable form of additional structure is available. Indeed, our main idea helps establish concentration and consistency results for estimators of

- many full and non-full, curved exponential families;
- many models with dependent edges;
- finite populations of nodes;

as long as there is additional structure that helps address (a) and (b).

It is worth noting that verifying the main assumption is by no means trivial. Its verification is easiest when the sufficient statistics are monotone functions of edges, i.e., functions of a graph that do not decrease (increase) when edges are added to the graph. It is less straightforward when the sufficient statistics are not monotone functions of edges, as is the case with many popular curved exponential-family random graphs with geometrically weighted model terms [53, 24]. But we demonstrate in Section 3.3 that the main assumption can be verified even when the sufficient statistics are not monotone functions of a graph, using curved exponential-family random graphs with geometrically weighted model terms as an example.

We make these ideas rigorous in Theorem 1 in Section 3.3. An application to curved exponential-family random graphs is presented in Corollary 1 in Section 3.3. More general results for M-estimators, under correct and incorrect model specifications, are mentioned in Section 3.4. To prepare the ground for these results, we first introduce concentration results for sufficient statistics in Section 3.2.

3.2. *Concentration results for sufficient statistics.* To obtain concentration results for sufficient statistics, we need concentration results for functions of random graphs with dependent edges.

Such concentration results are non-trivial for at least two reasons. First, exponential families of the form (2.1) may induce strong dependence within neighborhoods and the sizes of neighborhoods need not be identical. Second, exponential families of the form (2.1) can induce a wide range of dependencies within neighborhoods. Therefore, we need general-purpose concentration inequalities that cover a wide range of dependencies.

The following general-purpose concentration inequality addresses the challenges discussed above. It shows that the dependence induced by exponential families of the form (2.1) may be strong within neighborhoods but is sufficiently weak overall

to obtain concentration results as long as the neighborhoods are not too large.

Proposition 1. *Consider an exponential family with countable support $\mathbb{X}$ and local dependence. Let $f : \mathbb{X} \mapsto \mathbb{R}$ be Lipschitz with respect to the Hamming metric $d : \mathbb{X} \times \mathbb{X} \mapsto \{0, 1, 2, \dots\}$ with Lipschitz coefficient $\|f\|_{Lip} > 0$ and assume that $\mathbb{E}\, f(\boldsymbol{X})$ exists. Then there exists $C > 0$ such that, for all $t > 0$,*

$$\mathbb{P}(|f(\boldsymbol{X}) - \mathbb{E}\, f(\boldsymbol{X})| \geq t) \quad \leq \quad 2 \exp\left(-\frac{t^2}{C \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2} \|\mathcal{A}\|_\infty^4 \|f\|_{Lip}^2}\right).$$

Proposition 1 covers a wide range of exponential-family random graphs with local dependence. The assumption that the function of interest is smooth, in the sense that it is Lipschitz with respect to the Hamming metric, is common in the concentration-of-measure literature: see, e.g., the concentration results for dependent random variables by Samson [46], Chatterjee [8, Theorem 4.3, p. 75], and Kontorovich and Ramanan [32]. The smoothness assumption can be weakened by using divide-and-conquer strategies: e.g., one may divide the domain of a function of interest into high- and low-probability regions and require the function to be smooth on high-probability regions, but not on low-probability regions. Such divide-and-conquer strategies were used by, e.g., Vu [55], Kim and Vu [30], and Yang et al. [61, Lemma 9]. While exploring divide-and-conquer strategies for exponential-family random graphs with local dependence would be interesting, Proposition 1 suffices for the purpose of demonstrating that statistical inference for many exponential-family random graphs with local dependence is possible.

Proposition 1 paves the way for concentration results for sufficient statistics. Proposition 2 shows that the sufficient statistic vector $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} = s(\boldsymbol{X})$ is close to the data-generating mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \in \text{rint}(\mathbb{M})$ with high probability provided the number of neighborhoods K is large relative to the size of the largest neighborhood $\|\mathcal{A}\|_\infty$ and the dimension m of $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$.

Proposition 2. *Consider a full or non-full, curved exponential family with countable support $\mathbb{X}$ and local dependence. Assume that there exists $A > 0$ such that*

$$\|s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2)\|_\infty \quad \leq \quad A\, d(\boldsymbol{x}_1, \boldsymbol{x}_2)\, \|\mathcal{A}\|_\infty \;\; \textit{for all } (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X}. \tag{3.3}$$

Then there exists $C > 0$ such that, for all deviations of the form $t = \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha$ with $\delta > 0$ and $0 \leq \alpha \leq 1$,

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \geq t\right) \quad \leq \quad 2 \exp\left(-\frac{\delta^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4(2-\alpha)}} + \log m\right).$$

The smoothness condition (3.3) of Proposition 2 is satisfied as long as changing an edge cannot change the within-neighborhood sufficient statistics by more than a constant multiple of $\|\mathcal{A}\|_\infty$. It is verified in Corollary 1 in Section 3.3.

Remark 1. The relationship between α and sparsity. Proposition 2 shows how the concentration of sufficient statistics depends on the power $\alpha \in [0, 1]$ of deviations of size $t = \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}$. The power α can be interpreted as the level of sparsity of a random graph, with lower values of α corresponding to higher levels of sparsity. The conventional definition of a sparse random graph is based on the scaling of the expected number of edges, i.e., the sufficient statistic of Bernoulli random graphs with independent edges. We use the term sparse random graph to refer to the scaling of the expectations of all sufficient statistics of exponential-family random graphs. If $\alpha = 1$, the within-neighborhood subgraphs may be called dense in the sense that the expectations of within-neighborhood sufficient statistics are non-negligible fractions of the number of edge variables $\binom{|\mathcal{A}_k|}{2}$ in neighborhood $\mathcal{A}_k$ ($k = 1, \ldots, K$). Otherwise, the within-neighborhood subgraphs may be called sparse. Note that the interpretation of α in terms of sparsity makes more sense when the neighborhoods grow than when the neighborhoods are bounded above. But, regardless of whether the neighborhoods grow, Proposition 2 shows how the concentration of sufficient statistics depends on the power α.

Remark 2. Sharpness. The concentration results discussed above are not, and cannot be sharp, because these results cover many models and many dependencies. It goes without saying that in special cases sharper results can be obtained. We are here not interested in sharp bounds in special cases, because the main appeal of the exponential-family framework is that it can capture many dependencies.

3.3. *Maximum likelihood estimators.* The main idea of Section 3.1 is made rigorous in Theorem 1, which establishes concentration results for maximum likelihood estimators of full and non-full, curved exponential-family random graphs with local dependence.

Theorem 1. *Consider a full or non-full, curved exponential-family random graph with countable support $\mathbb{X}$ and local dependence. Let*

$$\boldsymbol{\Theta} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^q : \psi(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}.$$

Assume that $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$. Let $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ be one-to-one and assume that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we have $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$. In addition, assume that there exist $\delta(\epsilon) > 0$ and $A > 0$ such that, for all $\boldsymbol{\eta}(\boldsymbol{\theta}) \in$

$\mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$,

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \;\geq\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \quad \textit{for some} \;\; 0 \leq \alpha \leq 1 \tag{3.4}$$

and, for all $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X}$,

$$\|s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2)\|_\infty \;\leq\; A\, d(\boldsymbol{x}_1, \boldsymbol{x}_2)\, \|\mathcal{A}\|_\infty. \tag{3.5}$$

Then, for all $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, *there exist* $\kappa(\epsilon) > 0$ *and* $C > 0$ *such that*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)\right) \;\geq\; 1 - 2\,\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m\right).$$

If the exponential family is full, then $\widehat{\boldsymbol{\theta}}$ *is unique in the event* $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$.

Theorem 1 shows that the estimator $\widehat{\boldsymbol{\theta}}$ is close to the data-generating parameter vector $\boldsymbol{\theta}^\star$ with high probability provided the number of neighborhoods K is large relative to $\|\mathcal{A}\|_\infty$ and m. An application of Theorem 1 to popular curved exponential-family random graphs can be found in Corollary 1. These concentration results cover inference scenarios with a finite population of nodes and a population graph generated by an exponential-family random graph model, and assume that the population graph can be observed. Inference scenarios where the population graph cannot be observed but subgraphs of the population graph can be observed are considered in Section 4. Asymptotic consistency results can be obtained by allowing the number of neighborhoods K to increase without bound. We discuss asymptotic consistency results in Remark 3 following Corollary 1.

Conditions (3.4) and (3.5) of Theorem 1 are verified in Corollary 1. As pointed out in Section 3.1, the assumptions are satisfied when the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one and continuous, but it need not be differentiable. Condition (3.4) is an identifiability assumption and covers both sparse ($0 \leq \alpha < 1$) and dense ($\alpha = 1$) within-neighborhood subgraphs. The power α can be interpreted as the level of sparsity of a random graph, as explained in Section 3.2. Theorem 1 shows that sparsity comes at a cost, because the probability of event $\widehat{\boldsymbol{\theta}} \notin \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ decays slower when the within-neighborhood subgraphs are sparse rather than dense. The fact that sparsity weakens concentration results is well-known in the concentration-of-measure literature on random graphs with independent edges [e.g., 27, 30]. Condition (3.5) is a smoothness condition, which is satisfied as long as changing an edge cannot change the within-neighborhood sufficient statistics by more than a constant multiple of $\|\mathcal{A}\|_\infty$.

It is worth noting that in full exponential families the set $\widehat{\boldsymbol{\theta}}$ contains a single element when it is non-empty, whereas in non-full exponential families it may contain more than one element. A pleasant feature of estimating function (3.1) is that, with high probability, the minimizers $\widehat{\boldsymbol{\theta}}$ of (3.1) do not give rise to global minima that are separated by large distances, under the assumptions made. The reason is that, if the set $\widehat{\boldsymbol{\theta}}$ contains more than one element, then all elements of the set $\widehat{\boldsymbol{\theta}}$ map to mean-value parameter vectors $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ whose ℓ_2-distance from $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}$ is identical and whose ℓ_2-distance from $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$ is bounded above by

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \quad \leq \quad 2\,\|\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2,$$

as explained in Section 3.1. By Proposition 2, $\widehat{\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}$ is close to $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$ with high probability provided the number of neighborhoods K is sufficiently large. Therefore, all elements of the set $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))$ are close to $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$ with high probability and hence, by the identifiability conditions of Theorem 1, all elements of the set $\widehat{\boldsymbol{\theta}}$ are close to $\boldsymbol{\theta}^\star$ with high probability.

Applications. We present two applications of Theorem 1. An application to canonical exponential-family random graphs can be found in Appendix A [see the supplement, 50]. Here, we focus on curved exponential-family random graphs with geometrically weighted model terms, which are popular in practice [e.g., 39] but are challenging on theoretical grounds.

As a specific example, consider curved exponential-family random graphs with support $\mathbb{X} = \{0,1\}^{\sum_{k=1}^K \binom{|\mathcal{A}_k|}{2}}$ and within-neighborhood edge and geometrically weighted edgewise shared partner terms [53, 24]. Such models are based on sufficient statistics of the form

$$\begin{aligned} s_{k,1}(\boldsymbol{x}_k) &= \sum_{a \in \mathcal{A}_k < b \in \mathcal{A}_k} x_{a,b} \\ s_{k,i+1}(\boldsymbol{x}_k) &= \sum_{a \in \mathcal{A}_k < b \in \mathcal{A}_k} x_{a,b}\; f_{a,b,i}(\boldsymbol{x}_k), \quad i = 1, \ldots, |\mathcal{A}_k| - 2, \end{aligned}$$

where $f_{a,b,i}(\boldsymbol{x}_k) = \mathbb{1}(\sum_{c \in \mathcal{A}_k,\, c \neq a,b} x_{a,c}\, x_{b,c} = i)$ is an indicator function, which is 1 if nodes a and b are both connected to i other nodes in neighborhood $\mathcal{A}_k$ and is 0 otherwise ($k = 1, \ldots, K$). The natural parameters are of the form

$$\begin{aligned} \eta_{k,1}(\boldsymbol{\theta}) &= \theta_1 \\ \eta_{k,i+1}(\boldsymbol{\theta}) &= \exp(\vartheta)\,\left[1 - (1 - \exp(-\vartheta))^i\right], \quad i = 1, \ldots, |\mathcal{A}_k| - 2, \end{aligned}$$

where $\vartheta > 0$ controls the rate of decay of the geometric sequence $(1 - \exp(-\vartheta))^i$, $i = 1, 2, \ldots$ ($k = 1, \ldots, K$). For convenience, we consider here the parameterization $\theta_2 = \exp(-\vartheta) \in (0,1)$, so that $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R} \times (0,1)$. Such model terms

are called geometrically weighted terms, because the natural parameters $\eta_{k,i+1}(\boldsymbol{\theta})$ are based on the geometric sequence $(1-\exp(-\vartheta))^i$, $i = 1, 2, \ldots$

While complicated, such models are able to capture transitivity in neighborhoods. As explained in Section 1, transitivity is one of the more interesting network phenomena, and induces dependence among edges. There are many models of transitivity, some of which are well-posed while others are ill-posed. An example of a model that is ill-posed in the large-graph limit is the so-called triangle model [e.g., 28, 18, 47, 2, 10]. The triangle model is a canonical exponential-family random graph model with the number of edges and triangles as sufficient statistics; note that a triangle in a random graph corresponds to three distinct nodes such that all three pairs of nodes are connected by edges. Compared with the triangle model, the curved exponential-family random graph model described above makes more reasonable assumptions:

- The curved exponential-family random graph model exploits multilevel structure to constrain the dependence among edges induced by transitivity to neighborhoods, i.e., subsets of nodes. By contrast, the triangle model does not restrict transitivity to subsets of nodes, and allows each edge to depend on many other edges in the random graph.
- The curved exponential-family random graph model implies that within neighborhoods, for each pair of nodes, the value added by additional triangles to the log odds of the conditional probability of an edge decays at a geometric rate [e.g., 23, 54]. As a result, the model encourages triangles within neighborhoods, but discourages too many of them. By contrast, the added value of additional triangles under the triangle model is constant, so that the triangle model with a positive triangle parameter places more probability mass on graphs with more triangles (among graphs with the same number of edges).

While the problematic assumptions underlying the triangle model lead to undesirable behavior in large random graphs [e.g., 28, 18, 47, 2, 10], curved exponential-family random graphs with geometrically weighted edgewise shared partner terms have turned out to be well-behaved [e.g., 24, 47] and have been widely used [see, e.g., 53, 25, 23, 24, 39, 54]. A full-fledged discussion of these complex models is beyond the scope of our paper. We therefore refer the interested reader to the above-cited literature and focus here on concentration and consistency results.

Curved exponential-family random graphs with within-neighborhood edge and geometrically weighted edgewise shared partner terms are popular in practice but are challenging on theoretical grounds, for several reasons. First, the dimension of the natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathbb{R}^{\|\mathcal{A}\|_\infty - 1}$ is an increasing function of the number of nodes in the largest neighborhood(s), $\|\mathcal{A}\|_\infty$. Second, the natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathbb{R}^{\|\mathcal{A}\|_\infty - 1}$ is a non-affine function of a lower-dimensional parameter

vector $\boldsymbol{\theta} \in \mathbb{R} \times (0, 1)$. Third, the mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is not available in closed form. Finally, the sufficient statistics $s_2(\boldsymbol{x}), \ldots, s_{\|\mathcal{A}\|_\infty - 1}(\boldsymbol{x})$ are not monotone functions of graphs, which complicates the verification of the main assumption of Theorem 1, as mentioned in Section 3.1. Despite these challenges, it is possible to verify all conditions of Theorem 1 and obtain the following concentration result. It shows that the estimator $\widehat{\boldsymbol{\theta}}$ is close to the data-generating parameter vector $\boldsymbol{\theta}^\star$ with high probability provided K is large relative to $\|\mathcal{A}\|_\infty^6 \log\|\mathcal{A}\|_\infty$.

Corollary 1. *Consider a curved exponential-family random graph with within-neighborhood edge and geometrically weighted edgewise shared partner terms. Let* $\boldsymbol{\Theta} = \mathbb{R} \times (0, 1)$ *and assume that* $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$. *Then all conditions of Theorem 1 are satisfied and hence, for all* $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, *there exist* $\kappa(\epsilon) > 0$ *and* $C > 0$ *such that*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)\right) \geq 1 - 2\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{\|\mathcal{A}\|_\infty^6} + \log\|\mathcal{A}\|_\infty\right),$$

provided $|\mathcal{A}_k| \geq 4$ *(*$k = 1, \ldots, K$*) and* $K \geq 2$.

Corollary 1 is the first concentration result for estimators of exponential-family random graphs with dependence among edges induced by transitivity, one of the more interesting network phenomena. In addition, it is the first concentration result for curved exponential-family random graphs with geometrically weighted model terms, which are popular in practice [e.g., 53, 24]. The concentration result assumes that each neighborhood $\mathcal{A}_k$ consists of $|\mathcal{A}_k| \geq 4$ nodes, because $\boldsymbol{\theta}^\star$ is not identifiable when $|\mathcal{A}_k| \leq 3$ ($k = 1, \ldots, K$). As mentioned above, the set $\widehat{\boldsymbol{\theta}}$ may contain more than one element, but all elements of the set $\widehat{\boldsymbol{\theta}}$ are close to $\boldsymbol{\theta}^\star$ with high probability provided K is large relative to $\|\mathcal{A}\|_\infty^6 \log\|\mathcal{A}\|_\infty$. Concentration results for other curved exponential-family random graphs with geometrically weighted model terms [e.g., 53, 24] can be established along the same lines.

Remark 3. *Asymptotic consistency results.* As pointed out in Section 1, we state all theoretical results for finite populations of nodes, because in practice all populations are finite. Asymptotic consistency results can be obtained by allowing the number of neighborhoods K to grow without bound. If there exists a universal constant $C > 0$ such that $|\mathcal{A}_k| < C$ ($k = 1, 2, \ldots$), then the main idea described in Section 3.1 along with the concentration results in Section 3.2 imply that $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}^\star$ with rate of convergence $K^{1/2}$. As the units of statistical analysis are neighborhoods, the rate $K^{1/2}$ resembles the rate in classical statistical problems where the rate is the square root of the sample size, albeit with two notable differences: first, the units are subsets of nodes (neighborhoods) rather than nodes or edges; and, second, the sizes of units are not identical, but the size of the largest unit is a constant multiple of the size of the smallest. Last, but not least, it is

possible to obtain asymptotic consistency results when K grows and the neighborhoods grow with K, which implies that $\|\mathcal{A}\|_\infty$ grows with K. Then, as long as K grows faster than $\|\mathcal{A}\|_\infty^6 \log \|\mathcal{A}\|_\infty$ in the sense that $K \,/\, (\|\mathcal{A}\|_\infty^6 \log \|\mathcal{A}\|_\infty) \to \infty$, the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ tends to 1.

3.4. *M-estimators, correct and incorrect model specifications.* The concentration and consistency results for maximum likelihood estimators in Section 3.3 are special cases of more general results for M-estimators. To demonstrate, we introduce a natural class of M-estimators in Appendix B.1, which includes both likelihood- and moment-based estimators, and present concentration results in Appendix B.2 along with an application to misspecified models with omitted covariate terms. These results cover both correct and incorrect model specifications, as the example with omitted covariate terms demonstrates. Due to space restrictions, we provide details in Appendix B [see the supplement, 50].

4. Extendability and subgraph-to-graph estimators. A question that has been asked about exponential-family random graphs is whether it is possible to extend, in a well-defined sense, an exponential-family random graph with a given set of nodes to an exponential-family random graph with more nodes [51, 11, 36]. We show that multilevel structure helps extend an exponential-family random graph with a given set of neighborhoods to an exponential-family random graph with more neighborhoods (Section 4.1) and hence facilitates subgraph-to-graph estimation (Section 4.2). The importance of these results lies in the fact that subgraph-to-graph estimation for exponential-family random graphs is believed to be difficult [e.g., 51], but our results demonstrate that additional structure facilitates it.

4.1. *Extendability.* While many exponential-family random graphs with a given set of nodes cannot be extended to exponential-family random graphs with more nodes [51, 11, 36], an exponential-family random graph with a given set of neighborhoods can be extended to an exponential-family random graph with more neighborhoods.

To demonstrate, consider a population graph $(\boldsymbol{X}_\mathcal{L}, \boldsymbol{Y}_\mathcal{L})$ with a set of neighborhoods $\mathcal{L} = \{\mathcal{A}_1, \ldots, \mathcal{A}_L\}$, where $\boldsymbol{X}_\mathcal{L} \in \mathbb{X}_\mathcal{L}$ and $\boldsymbol{Y}_\mathcal{L} \in \mathbb{Y}_\mathcal{L}$ denote the sequences of within- and between-neighborhood edge variables based on the set of neighborhoods $\mathcal{L}$, respectively. As before, assume that $\boldsymbol{X}_\mathcal{L}$ is governed by an exponential family with countable support $\mathbb{X}_\mathcal{L}$ and local dependence, with neighborhood-dependent natural parameters $\eta_{\mathcal{A},i}(\boldsymbol{\theta}) = \eta_i(\boldsymbol{\theta})$ and sufficient statistics $s_{\mathcal{A},i}(\boldsymbol{x}_\mathcal{A})$ ($i = 1, \ldots, m_\mathcal{A}$, $\mathcal{A} \in \mathcal{L}$). Therefore, the exponential family can be reduced to an exponential family with natural parameter vector

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \quad = \quad (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta})) \tag{4.1}$$

and sufficient statistic vector

$$s(\boldsymbol{x}_{\mathcal{L}}) \;=\; (s_1(\boldsymbol{x}_{\mathcal{L}}), \ldots, s_m(\boldsymbol{x}_{\mathcal{L}}))\,,$$

where $s_i(\boldsymbol{x}_{\mathcal{L}}) = \sum_{\mathcal{A}\in\mathcal{L}} s_{\mathcal{A},i}(\boldsymbol{x}_{\mathcal{A}})$ $(i = 1, \ldots, m)$ and $m = \max_{\mathcal{A}\in\mathcal{L}} m_{\mathcal{A}}$.

Consider a subgraph $(\boldsymbol{X}_{\mathcal{K}}, \boldsymbol{Y}_{\mathcal{K}})$ induced by a subset of neighborhoods $\mathcal{K} \subseteq \mathcal{L}$. Then the subgraph $(\boldsymbol{X}_{\mathcal{K}}, \boldsymbol{Y}_{\mathcal{K}})$ with subset of neighborhoods $\mathcal{K}$ is extendable to the population graph $(\boldsymbol{X}_{\mathcal{L}}, \boldsymbol{Y}_{\mathcal{L}})$ with set of neighborhoods $\mathcal{L} \supset \mathcal{K}$ as follows.

Proposition 3. *Consider a full or non-full, curved exponential-family random graph with set of neighborhoods $\mathcal{L}$, countable support $\mathbb{X}_{\mathcal{L}}$, and local dependence. Assume that, for all $\boldsymbol{y}_{\mathcal{L}} \in \mathbb{Y}_{\mathcal{L}}$,*

$$\mathbb{P}(\boldsymbol{Y}_{\mathcal{L}} = \boldsymbol{y}_{\mathcal{L}}) \;=\; \prod_{\mathcal{C}\in\mathcal{L},\, \mathcal{D}\in\mathcal{L},\, \mathcal{C}\neq\mathcal{D}} \mathbb{P}(\boldsymbol{Y}_{\mathcal{C},\mathcal{D}} = \boldsymbol{y}_{\mathcal{C},\mathcal{D}}),$$

where $\boldsymbol{Y}_{\mathcal{C},\mathcal{D}} = (Y_{i,j})_{i\in\mathcal{C},\, j\in\mathcal{D}}$. Then, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^q :\ \psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}$, all $\mathcal{K} \subseteq \mathcal{L}$, and all $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ and $\boldsymbol{y}_{\mathcal{K}} \in \mathbb{Y}_{\mathcal{K}}$,

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \\ =\;& \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}}), \end{aligned}$$

where

$$\psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \;=\; \sum_{\mathcal{A}\in\mathcal{L}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \;=\; \sum_{\mathcal{A}\in\mathcal{L}} \log \sum_{\boldsymbol{x}_{\mathcal{A}}\in\mathbb{X}_{\mathcal{A}}} \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})\rangle\right)\, \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}).$$

The marginal density of a subgraph $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ of $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$ induced by $\mathcal{K} \subseteq \mathcal{L}$ is an exponential-family density with support $\mathbb{X}_{\mathcal{K}}$ and local dependence:

$$\sum_{\boldsymbol{x}_{\mathcal{L}\setminus\mathcal{K}}\in\mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}} p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{L}}) = p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{K}}) = \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), s(\boldsymbol{x}_{\mathcal{K}})\rangle - \psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \nu_{\mathcal{K}}(\boldsymbol{x}_{\mathcal{K}}),$$

where $\psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \sum_{\mathcal{A}\in\mathcal{K}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$, $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta}))$, $s(\boldsymbol{x}_{\mathcal{K}}) = (\sum_{\mathcal{A}\in\mathcal{K}} s_{\mathcal{A},1}(\boldsymbol{x}_{\mathcal{A}}), \ldots, \sum_{\mathcal{A}\in\mathcal{K}} s_{\mathcal{A},m}(\boldsymbol{x}_{\mathcal{A}}))$, and $\nu_{\mathcal{K}}(\boldsymbol{x}_{\mathcal{K}}) = \prod_{\mathcal{A}\in\mathcal{K}} \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$.

Thus, in the above-mentioned sense, the exponential-family random graph induced by a subset of neighborhoods $\mathcal{K}$ can be extended to the exponential-family random graph with set of neighborhoods $\mathcal{L} \supset \mathcal{K}$. A more restrictive result was proved by Schweinberger and Handcock [48, Theorem 1].

4.2. *Subgraph-to-graph estimators.* The extendability of exponential-family random graphs with multilevel structure discussed in Section 4.1 facilitates subgraph-to-graph estimation.

To demonstrate, let $\mathcal{L}$ be the set of neighborhoods of the population graph and assume that $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$ was generated by an exponential family with countable support $\mathbb{X}_{\mathcal{L}}$ and local dependence. Suppose that it is infeasible to observe $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$, but it is feasible to sample a subset of neighborhoods $\mathcal{K} \subseteq \mathcal{L}$ and collect data on the subgraphs induced by $\mathcal{K} \subseteq \mathcal{L}$. We assume henceforth that the sampling design is ignorable in the sense of Rubin [45] and Handcock and Gile [19], i.e., the probability of observing subgraphs does not depend on the unobserved subgraphs. A simple example is a sampling design that samples neighborhoods at random and collects data on the subgraphs induced by the sampled neighborhoods.

By Proposition 3 and the ignorability of the sampling design [45, 19], the observed-data likelihood function based on the observed subgraph $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ of $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$ is proportional to

$$\sum_{\boldsymbol{x}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}} p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{L}}) \quad = \quad p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{K}}), \qquad \boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}},$$

where $\boldsymbol{\eta}(\boldsymbol{\theta})$ is of the form (4.1). In other words, maximum likelihood estimation can be based on $p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{K}})$. Motivated by the same considerations we outlined in Section 3, we therefore consider an estimating function of the form

$$g_{\mathcal{K}}(\boldsymbol{\theta};\, \widehat{\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) \quad = \quad \|\widehat{\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} - \boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2,$$

where $\widehat{\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))} = s(\boldsymbol{x}_{\mathcal{K}})$, $\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\, s(\boldsymbol{X}_{\mathcal{K}})$, and $s(\boldsymbol{x}_{\mathcal{K}})$ is defined in Proposition 3. The data-generating parameter vector $\boldsymbol{\theta}^\star$ of the population graph can hence be estimated by the estimator $\widehat{\boldsymbol{\theta}}_{\mathcal{K}}$ based on the observed subgraph $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ of $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$:

$$\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \quad = \quad \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \;\; g_{\mathcal{K}}(\boldsymbol{\theta};\, \widehat{\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) \quad = \quad \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} \; g_{\mathcal{K}}(\dot{\boldsymbol{\theta}};\, \widehat{\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))}) \right\}.$$

The following concentration result shows that, with high probability, the estimator $\widehat{\boldsymbol{\theta}}_{\mathcal{K}}$ based on the observed subgraph induced by $\mathcal{K} \subseteq \mathcal{L}$ is close to the data-generating parameter vector $\boldsymbol{\theta}^\star$ of the population graph as long as the number of sampled neighborhoods $|\mathcal{K}|$ is large relative to $\|\mathcal{L}\|_\infty = \max_{\mathcal{A}\in\mathcal{L}} |\mathcal{A}|$ and $m = \max_{\mathcal{A}\in\mathcal{L}} m_{\mathcal{A}}$.

Theorem 2. *Consider a full or non-full, curved exponential-family random graph with set of neighborhoods $\mathcal{L}$, countable support $\mathbb{X}_{\mathcal{L}}$, and local dependence. Let*

$$\boldsymbol{\Theta} \quad \subseteq \quad \{\boldsymbol{\theta} \in \mathbb{R}^q : \psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}.$$

Assume that $\boldsymbol{\theta}^\star \in \mathrm{int}(\boldsymbol{\Theta})$ and that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \mathrm{int}(\boldsymbol{\Theta})$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we have $\boldsymbol{\eta}(\boldsymbol{\theta}) \in$

$\mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$. *In addition, assume that there exist* $\delta(\epsilon) > 0$ *and* $A > 0$ *such that, for all* $\mathcal{K} \subseteq \mathcal{L}$ *and all* $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$,

$$\|\boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \;\geq\; \delta(\epsilon) \sum_{\mathcal{A} \in \mathcal{K}} \binom{|\mathcal{A}|}{2}^{\alpha} \text{ for some } 0 \leq \alpha \leq 1 \tag{4.2}$$

and, for all $\mathcal{K} \subseteq \mathcal{L}$ *and all* $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X}_{\mathcal{K}} \times \mathbb{X}_{\mathcal{K}}$,

$$\|s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2)\|_\infty \;\leq\; A\, d(\boldsymbol{x}_1, \boldsymbol{x}_2)\, \|\mathcal{K}\|_\infty, \tag{4.3}$$

where $\|\mathcal{K}\|_\infty = \max_{\mathcal{A} \in \mathcal{K}} |\mathcal{A}|$. *Then, for all* $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, *there exist* $\kappa(\epsilon) > 0$ *and* $C > 0$ *such that*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)\right) \;\geq\; 1 - 2\, \exp\left(-\frac{\kappa(\epsilon)^2\, C\, |\mathcal{K}|}{m\, \|\mathcal{L}\|_\infty^{4(2-\alpha)}} + \log m\right).$$

If the exponential family is full, then $\widehat{\boldsymbol{\theta}}_{\mathcal{K}}$ *is unique in the event* $\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$.

Theorem 2 shows that there are costs associated with observing a subset of neighborhoods $\mathcal{K} \subseteq \mathcal{L}$ rather than the whole set of neighborhoods $\mathcal{L}$ of the population graph: The probability of event $\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \notin \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ decays with the number of sampled neighborhoods $|\mathcal{K}|$ and is hence lowest when the whole set of neighborhoods $\mathcal{L}$ of the population graph is sampled.

As a specific example, consider the main example of Section 3.3: curved exponential-family random graphs with within-neighborhood edge and geometrically weighted edgewise shared partner terms.

Corollary 2. *Consider a curved exponential-family random graph with set of neighborhoods* $\mathcal{L}$, *countable support* $\mathbb{X}_{\mathcal{L}}$, *and local dependence induced by within-neighborhood edge and geometrically weighted edgewise shared partner terms. Let* $\boldsymbol{\Theta} = \mathbb{R} \times (0, 1)$ *and assume that* $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$. *Then all conditions of Theorem 2 are satisfied and hence, for all* $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, *there exist* $\kappa(\epsilon) > 0$ *and* $C > 0$ *such that*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)\right) \;\geq\; 1 - 2\, \exp\left(-\frac{\kappa(\epsilon)^2\, C\, |\mathcal{K}|}{\|\mathcal{L}\|_\infty^6} + \log\|\mathcal{L}\|_\infty\right),$$

provided $|\mathcal{A}| \geq 4$ $(\mathcal{A} \in \mathcal{L})$ *and* $|\mathcal{K}| \geq 2$.

Remark 4. "Bad" subsets of neighborhoods $\mathcal{K} \subseteq \mathcal{L}$. Since the neighborhoods need not have the same size, it is natural to ask whether it is possible to sample a "bad" subset of neighborhoods $\mathcal{K}$ with too small or too large neighborhoods, which could make it challenging to estimate some of the parameters. However, the

assumptions of Theorem 2 rule out "bad" subsets of neighborhoods $\mathcal{K}$, for two reasons. First, while some neighborhoods may be larger than others, Theorem 2 assumes that the neighborhoods are of the same order of magnitude, as defined in Section 2. In other words, the neighborhoods have similar sizes. Second, the conditions of Theorem 2 assume that the model satisfies identifiability and smoothness conditions for all possible subsets of neighborhoods $\mathcal{K} \subseteq \mathcal{L}$. Corollary 2 shows that, in the special case of curved exponential-family random graphs with within-neighborhood edge and geometrically weighted edgewise shared partner terms, the identifiability conditions require $|\mathcal{A}| \geq 4$ for all neighborhoods $\mathcal{A} \in \mathcal{L}$ of the population graph. Thus, no neighborhood can be too small, and no neighborhood can be too large, because all neighborhoods are of the same order of magnitude. As a consequence, under the stated assumptions, it is impossible to sample a "bad" subset of neighborhoods $\mathcal{K}$ with too small or too large neighborhoods.

5. Comparison with existing consistency results. To compare our concentration and consistency results to existing consistency results, we focus on exponential-family random graphs with dependent edges. It is worth noting that there are consistency results for exponential-family random graphs with independence assumptions—see, e.g., Diaconis et al. [12], Rinaldo et al. [44], Krivitsky and Kolaczyk [35], and Yan et al. [60, 59]—but such independence assumptions may not be satisfied in applications, as discussed in Section 1.

Concerning exponential-family random graphs with dependent edges, Shalizi and Rinaldo [51] showed that maximum likelihood estimators of natural parameters of fixed dimension are consistent provided exponential-family random graphs satisfy strong extendability or projectability assumptions. However, those projectability assumptions rule out dependencies induced by transitivity and many other interesting network phenomena. Xiang and Neville [58] reported consistency results under weak dependence assumptions, but did not give any example of an exponential-family random graph with dependent edges that satisfies those assumptions. Mukherjee [40] showed that consistent estimation of the so-called two-star model is possible, but those results have not been extended to other exponential-family random graphs. In addition, Shalizi and Rinaldo [51], Xiang and Neville [58], and Mukherjee [40] focus on consistency results for estimators of natural parameter vectors whose dimensions do not depend on the number of nodes. By contrast, we advance the statistical theory of exponential-family random graphs by providing the first concentration and consistency results that cover

- a wide range of exponential-family random graphs with dependence among edges induced by transitivity and other interesting network phenomena;
- curved exponential-family random graphs with dependent edges and parameter vectors whose dimension depends on the number of nodes (Section 3);

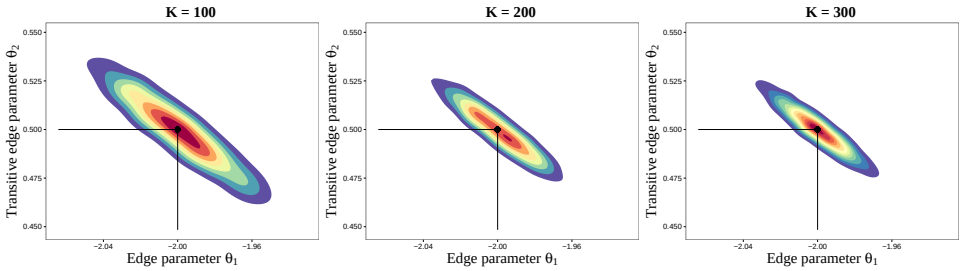

FIG 2. *1,000 estimates of the exponential-family random graph with within-neighborhood edge and transitive edge terms, where each graph consists of* $K = 100, 200,$ *and* 300 *neighborhoods of size* 50 *with natural parameter vectors* $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1, \theta_2)$. *The horizontal and vertical lines indicate the coordinates of the data-generating parameter vector* $\boldsymbol{\theta}^\star = (\theta_1^\star, \theta_2^\star)$.

- maximum likelihood and M-estimators (Section 3 and Appendix B);
- correct and incorrect model specifications (Section 3 and Appendix B);
- subgraph-to-graph estimators (Section 4).

These results underscore the importance of additional structure: It is the additional structure in the form of multilevel structure that facilitates these results.

6. Simulation results. To shed light on the finite-graph properties of maximum likelihood estimators, we generated data from the canonical and curved exponential-family random graphs mentioned in Section 3.3. We used R package hergm [49] to generate 1,000 graphs from each model and estimated the data-generating parameter vector by Monte Carlo maximum likelihood estimators [24].

We first consider canonical exponential-family random graphs with support $\mathbb{X} = \{0,1\}^{\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}}$ and local dependence induced by within-neighborhood edge and transitive edge terms [26]. Within-neighborhood edge and transitive edge terms correspond to neighborhood-dependent natural parameters $\eta_{k,1}(\boldsymbol{\theta}) = \theta_1$ and $\eta_{k,2}(\boldsymbol{\theta}) = \theta_2$ and sufficient statistics $s_{k,1}(\boldsymbol{x}_k)$ and $s_{k,2}(\boldsymbol{x}_k)$ given by

$$\begin{aligned} s_{k,1}(\boldsymbol{x}_k) &= \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} x_{i,j} \\ s_{k,2}(\boldsymbol{x}_k) &= \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} x_{i,j} \max_{h \in \mathcal{A}_k, h \neq i,j} x_{i,h}\, x_{j,h}, \end{aligned}$$

where $k = 1, \ldots, K$. It is worth noting that the number of transitive edges is not the same as the number of triangles. A discussion of the model along with concentration results for maximum likelihood estimators can be found in Appendix A [see the supplement, 50]. Figure 2 shows 1,000 estimates of the exponential-family random graph with within-neighborhood edge and transitive edge terms,

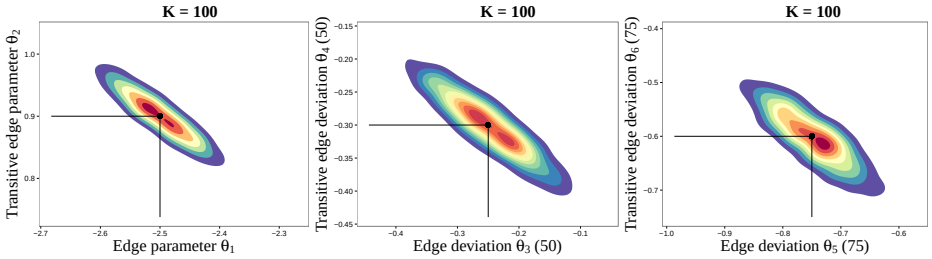

FIG 3. *1,000 estimates of the exponential-family random graph with within-neighborhood edge and transitive edge terms, where each graph consists of* 33 *neighborhoods of size* 25 *with natural parameter vectors* $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1, \theta_2)$*,* 34 *neighborhoods of size* 50 *with natural parameter vectors* $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1 + \theta_3, \theta_2 + \theta_4)$*, and* 33 *neighborhoods of size* 75 *with natural parameter vectors* $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1 + \theta_5, \theta_2 + \theta_6)$*. The horizontal and vertical lines indicate the coordinates of the data-generating parameter vector* $\boldsymbol{\theta}^\star = (\theta_1^\star, \theta_2^\star, \theta_3^\star, \theta_4^\star, \theta_5^\star, \theta_6^\star)$*.*

where each graph consists of $K = 100$, 200, and 300 neighborhoods of size 50 with natural parameter vectors $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1, \theta_2)$ $(k = 1, \ldots, K)$. The figure suggests that the probability mass of estimators becomes more and more concentrated in a neighborhood of the data-generating parameters as the number of neighborhoods K increases from 100 to 300, demonstrating that the concentration results in Section 3 are manifest when K is in the low hundreds and $\|\mathcal{A}\|_\infty = 50$.

Figure 3 sheds light on the performance of a simple form of a size-dependent parameterization that allows small and large neighborhoods to have different parameters. We consider exponential-family random graphs with within-neighborhood edge and transitive edge terms, where each graph consists of 33 neighborhoods of size 25 with natural parameter vectors $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1, \theta_2)$ $(k = 1, \ldots, 33)$, 34 neighborhoods of size 50 with natural parameter vectors $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1 + \theta_3, \theta_2 + \theta_4)$ $(k = 34, \ldots, 67)$, and 33 neighborhoods of size 75 with natural parameter vectors $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1 + \theta_5, \theta_2 + \theta_6)$ $(k = 68, \ldots, 100)$. Figure 3 demonstrates that the estimates of the baseline edge and transitive edge parameters θ_1 and θ_2 tend to be closer to the data-generating parameters than the deviation parameters $\theta_3, \theta_4, \theta_5$, and θ_6, because θ_1 and θ_2 are estimated from all neighborhoods whereas $\theta_3, \theta_4, \theta_5$, and θ_6 are estimated from a subset of neighborhoods.

Figure 4 shows 1,000 estimates of the curved exponential-family random graph with within-neighborhood edge and geometrically weighted edgewise shared partner terms described in Section 3.3. Each graph consists of $K = 100$ neighborhoods of size 50 with natural parameter vectors $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1, \eta_{k,1}(\theta_2), \ldots, \eta_{k,48}(\theta_2))$, where θ_1 is the natural parameter of the edge term and $\eta_{k,1}(\theta_2), \ldots, \eta_{k,48}(\theta_2)$ are the natural parameters of the geometrically weighted edgewise shared partner term $(k = 1, \ldots, 100)$. The figure shows that the probability mass of the estimators is concentrated in a small neighborhood of the data-generating parameters.

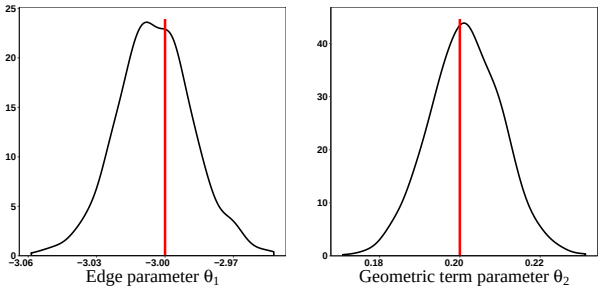

FIG 4. *1,000 estimates of the curved exponential-family random graph with within-neighborhood edge and geometrically weighted edgewise shared partner terms, where each graph consists of* $K = 100$ *neighborhoods of size* 50 *with natural parameter vectors* $\boldsymbol{\eta}_k(\boldsymbol{\theta}) = (\theta_1, \eta_{k,1}(\theta_2), \ldots, \eta_{k,48}(\theta_2))$. *The vertical lines indicate the coordinates of the data-generating parameter vector* $\boldsymbol{\theta}^\star = (\theta_1^\star, \theta_2^\star)$.

7. Discussion. We have taken constructive steps to demonstrate that statistical inference for exponential-family random graphs with dependence among edges induced by transitivity and other interesting network phenomena is possible, provided additional structure in the form of multilevel structure is available. The theoretical results reported here underscore the importance of additional structure. In practice, many other forms of additional structure exist and could be used for the purpose of facilitating statistical inference for exponential-family random graphs (e.g., other forms of multilevel structure or spatial structure).

Last, but not least, while we have focused here on theoretical results showing that multilevel structure facilitates statistical inference, it is worth noting that multilevel structure has computational benefits as well: The contributions of neighborhoods to estimating functions—e.g., the expectations of within-neighborhood sufficient statistics—may be computed or approximated by exploiting parallel computing on computing clusters.

Acknowledgements. We acknowledge support from the National Science Foundation (NSF awards DMS-1513644 and DMS-1812119).

Supplementary materials. All results are proved in the supplement [50]. In addition, the supplement contains concentration results for M-estimators. These results cover correct and incorrect model specifications.

References.

[1] Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: John Wiley & Sons.

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

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

[4] Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press.

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

[6] Butts, C. T. (2011), "Bernoulli graph bounds for general random graph models," *Sociological Methodology*, 41, 299–345.

[7] Butts, C. T., and Almquist, Z. W. (2015), "A Flexible Parameterization for Baseline Mean Degree in Multiple-Network ERGMs," *Journal of Mathematical Sociology*, 39, 163–167.

[8] Chatterjee, S. (2005), "Concentration Inequalities with Exchangeable Pairs," Ph.D. thesis, Department of Statistics, Stanford University.

[9] — (2007), "Estimation in spin glasses: A first step," *The Annals of Statistics*, 35, 1931–1946.

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

[11] Crane, H., and Dempsey, W. (2015), "A framework for statistical network modeling," Available at https://arxiv.org/abs/1509.08185.v4.

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

[13] Efron, B. (1975), "Defining the curvature of a statistical problem (with applications to second order efficiency)," *The Annals of Statistics*, 3, 1189–1242.

[14] — (1978), "The geometry of exponential families," *The Annals of Statistics*, 6, 362–376.

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

[16] Geyer, C. J. (2009), "Likelihood inference in exponential families and directions of recession," *Electronic Journal of Statistics*, 3, 259–289.

[17] Godambe, V. P. (1991), *Estimating Functions*, Oxford: Oxford University Press.

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

[19] Handcock, M. S., and Gile, K. (2010), "Modeling social networks from sampled data," *The Annals of Applied Statistics*, 4, 5–25.

[20] Harris, J. K. (2013), *An Introduction to Exponential Random Graph Modeling*, Thousand Oaks, California: Sage.

[21] Holland, P. W., and Leinhardt, S. (1976), "Local structure in social networks," *Sociological Methodology*, 1–45.

[22] Hollway, J., Lomi, A., Pallotti, F., and Stadtfeld, C. (2017), "Multilevel social spaces: The network dynamics of organizational fields," *Network Science*, 5, 187–212.

[23] Hunter, D. R. (2007), "Curved exponential family models for social networks," *Social Networks*, 29, 216–230.

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

[25] Hunter, D. R., and Handcock, M. S. (2006), "Inference in curved exponential family models

for networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

[26] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), "Computational statistical methods for social network models," *Journal of Computational and Graphical Statistics*, 21, 856–882.

[27] Janson, S., and Rucinski, A. (2002), "The infamous upper tail," *Random Structures and Algorithms*, 20, 317–342.

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

[29] Kass, R., and Vos, P. (1997), *Geometrical foundations of asymptotic inference*, New York: Wiley.

[30] Kim, J. H., and Vu, V. H. (2004), "Divide and conquer martingales and the number of triangles in a random graph," *Random Structures & Algorithms*, 24, 166–174.

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

[32] Kontorovich, L., and Ramanan, K. (2008), "Concentration inequalities for dependent random variables via the martingale method," *The Annals of Probability*, 36, 2126–2158.

[33] Krivitsky, P. N. (2012), "Exponential-family models for valued networks," *Electronic Journal of Statistics*, 6, 1100–1128.

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

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

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

[37] Lazega, E., and Snijders, T. A. B. (eds.) (2016), *Multilevel Network Analysis for the Social Sciences*, Switzerland: Springer-Verlag.

[38] Lomi, A., Robins, G., and Tranmer, M. (2016), "Introduction to multilevel social networks," *Social Networks*, 266–268.

[39] Lusher, D., Koskinen, J., and Robins, G. (2013), *Exponential Random Graph Models for Social Networks*, Cambridge, UK: Cambridge University Press.

[40] Mukherjee, S. (2013), "Consistent estimation in the two star exponential random graph model," Tech. rep., Department of Statistics, Columbia University, arXiv:1310.4526v1.

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

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

[43] Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), "On the geometry of discrete exponential families with application to exponential random graph models," *Electronic Journal of Statistics*, 3, 446–484.

[44] Rinaldo, A., Petrovic, S., and Fienberg, S. E. (2013), "Maximum likelihood estimation in network models," *The Annals of Statistics*, 41, 1085–1110.

[45] Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

[46] Samson, P. M. (2000), "Concentration of measure inequalities for Markov chains and Φ-mixing processes," *The Annals of Probability*, 28, 416–461.

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

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

[49] Schweinberger, M., and Luna, P. (2018), "HERGM: Hierarchical exponential-family random graph models," *Journal of Statistical Software*, 85, 1–39.

[50] Schweinberger, M., and Stewart, J. (2018), "Supplement: Finite-graph concentration and consistency results for canonical and curved exponential-family models of random graphs," Tech. rep., Department of Statistics, Rice University.

[51] Shalizi, C. R., and Rinaldo, A. (2013), "Consistency under sampling of exponential random graph models," *The Annals of Statistics*, 41, 508–535.

[52] Slaughter, A. J., and Koehly, L. M. (2016), "Multilevel models for social networks: hierarchical Bayesian approaches to exponential random graph modeling," *Social Networks*, 44, 334–345.

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

[54] Stewart, J., Schweinberger, M., Bojanowski, M., and Morris, M. (2018), "Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms," *Social Networks*, to appear.

[55] Vu, V. H. (2002), "Concentration of non-Lipschitz functions and applications," *Random Structures & Algorithms*, 20, 262–316.

[56] Wang, P., Robins, G., Pattison, P., and Lazega, E. (2013), "Exponential random graph models for multilevel networks," *Social Networks*, 35, 96–115.

[57] Wasserman, S., and Pattison, P. (1996), "Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^*," *Psychometrika*, 61, 401–425.

[58] Xiang, R., and Neville, J. (2011), "Relational learning with one network: an asymptotic analysis," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1–10.

[59] Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2018), "Statistical inference in a directed network model with covariates," *Journal of the American Statistical Association*, 1–33, to appear.

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

[61] Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015), "Graphical models via univariate exponential family distributions," *Journal of Machine Learning Research*, 16, 3813–3847.

[62] Zappa, P., and Lomi, A. (2015), "The analysis of multilevel networks in organizations: models and empirical tests," *Organizational Research Methods*, 18, 542–569.

SUPPLEMENT: CONCENTRATION AND CONSISTENCY RESULTS FOR CANONICAL AND CURVED EXPONENTIAL-FAMILY MODELS OF RANDOM GRAPHS

BY MICHAEL SCHWEINBERGER AND JONATHAN STEWART

Rice University

We first present an additional application of concentration results for maximum likelihood estimators in Appendix A and discuss concentration results for M-estimators in Appendix B. We then prove the main concentration results for random graphs with dependent edges in Appendix C and the main concentration results for maximum likelihood estimators, M-estimators, and subgraph-to-graph estimators in Appendices D, E, and F, respectively. Auxiliary lemmas are proved in Appendix G.

APPENDIX A: CONCENTRATION: MAXIMUM LIKELIHOOD ESTIMATORS

In addition to the application to curved exponential-family random graphs in Section 3.3, we present here an application to canonical exponential-family random graphs that induces dependence among edges through transitivity.

We consider canonical exponential-family random graphs with support $\mathbb{X} = \{0,1\}^{\sum_{k=1}^{K}\binom{|\mathcal{A}_k|}{2}}$ and local dependence induced by within-neighborhood edge and transitive edge terms [26]. Within-neighborhood edge and transitive edge terms correspond to neighborhood-dependent natural parameters $\eta_{k,1}(\boldsymbol{\theta}) = \theta_1$ and $\eta_{k,2}(\boldsymbol{\theta}) = \theta_2$ and sufficient statistics $s_{k,1}(\boldsymbol{x}_k)$ and $s_{k,2}(\boldsymbol{x}_k)$ given by

$$\begin{aligned} s_{k,1}(\boldsymbol{x}_k) &= \sum_{i\in\mathcal{A}_k < j\in\mathcal{A}_k} x_{i,j} \\ s_{k,2}(\boldsymbol{x}_k) &= \sum_{i\in\mathcal{A}_k < j\in\mathcal{A}_k} x_{i,j} \max_{h\in\mathcal{A}_k,\, h\neq i,j} x_{i,h}\, x_{j,h}, \end{aligned}$$

where $k = 1,\ldots,K$.

Corollary 3. *Consider an exponential-family random graph with within-neighborhood edge and transitive edge terms. Let* $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$*, where* $\boldsymbol{\Theta} = \mathbb{R}\times\mathbb{R}^+$*. Then conditions* (3.4) *and* (3.5) *of Theorem 1 are satisfied and hence, for all* $\epsilon > 0$*, there exist* $\kappa(\epsilon) > 0$ *and* $C > 0$ *such that* $\widehat{\boldsymbol{\theta}}$ *exists, is unique, and*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star,\epsilon)\right) \geq 1 - 4\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{\|\mathcal{A}\|_\infty^4}\right),$$

provided $|\mathcal{A}_k| \geq 3$ *(*$k = 1, \ldots, K$*) and* $K \geq 2$*.*

Remark 5. Transitive edge terms versus triangle terms. Transitive edge terms are not the same as triangle terms. As pointed out in Section 3.3, there are many models of transitivity, some of which are well-posed while others are ill-posed [e.g., the triangle model with edge and triangle terms, 28, 18, 47, 2, 10]. As the curved exponential-family random graphs in Section 3.3, canonical exponential-family random graphs with within-neighborhood edge and transitive edge terms constrain the dependence among edges induced by transitivity to neighborhoods. In addition, the model ensures that the added value of additional triangles decreases provided $\theta_2 > 0$. In fact, for each pair of nodes, the value added by the first triangle to the log odds of the conditional probability of an edge is $\theta_2 > 0$, whereas the added value of additional triangles is 0. Thus, the model has similar properties as the curved exponential-family random graphs discussed in Section 3.3.

APPENDIX B: CONCENTRATION RESULTS: M-ESTIMATORS

The concentration and consistency results for maximum likelihood estimators in Section 3.3 are special cases of more general results for M-estimators. To demonstrate, we introduce a natural class of M-estimators in Appendix B.1, which includes both likelihood- and moment-based estimators, and present concentration results in Appendix B.2 along with an application to misspecified models with omitted covariate terms. These results cover both correct and incorrect model specifications, as the example with omitted covariate terms demonstrates.

B.1. M-estimators. A natural class of M-estimators, which includes both likelihood- and moment-based estimators, can be constructed as follows.

Let $b : \mathbb{X} \mapsto \mathbb{R}^m$ be a vector of statistics. These statistics might be

- the sufficient statistics of the data-generating exponential family;
- the sufficient statistics of an exponential family other than the data-generating exponential family, which implies that the model is misspecified;
- statistics motivated by computational considerations.

An example is presented in Corollary 4, where we consider a misspecified exponential family with omitted covariate terms.

A natural extension of the class of likelihood-based estimating functions in Section 3 is given by estimating functions of the form

$$g(\boldsymbol{\theta};\, b(\boldsymbol{x})) \quad = \quad \|b(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{B.1}$$

which are approximations of

$$g(\boldsymbol{\theta};\, \mathbb{E}\, b(\boldsymbol{X})) \quad = \quad \|\mathbb{E}\, b(\boldsymbol{X}) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

provided $\mathbb{E}\, b(\boldsymbol{X})$ exists. Here, $\mathbb{E}\, b(\boldsymbol{X})$ is the expectation of $b(\boldsymbol{X})$ under the data-generating exponential-family distribution. The vector $\boldsymbol{\beta} : \boldsymbol{\Theta} \mapsto \mathbb{R}^m$ is a vector-valued function of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ defined by $\boldsymbol{\beta}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$, where $\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ is the expectation of $b(\boldsymbol{X})$ under a distribution parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, provided $\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ exists. The distribution may not belong to the data-generating exponential family or any other exponential family, i.e., the model may be misspecified.

Estimating functions of the form (B.1) cover both likelihood- and moment-based estimating functions:

- Likelihood-based estimating functions: The choice $b(\boldsymbol{x}) = s(\boldsymbol{x})$ gives $g(\boldsymbol{\theta};\, s(\boldsymbol{x})) = \|s(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X})\|_2$, which is based on the gradient of the loglikelihood function with respect to the natural parameter vector of the data-generating exponential family and is therefore based on moments of the sufficient statistics of the data-generating exponential family.
- Moment-based estimating functions: The choice $b(\boldsymbol{x}) \neq s(\boldsymbol{x})$ gives $g(\boldsymbol{\theta};\, b(\boldsymbol{x})) = \|b(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})\|_2$, which is based on moments of statistics other than the sufficient statistics of the data-generating exponential family.

M-estimators based on estimating functions of the form (B.1) are defined by

$$\widehat{\boldsymbol{\theta}} \;=\; \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} :\;\; g(\boldsymbol{\theta};\, b(\boldsymbol{x})) \;=\; \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}}\; g(\dot{\boldsymbol{\theta}};\, b(\boldsymbol{x}))\right\},$$

which are estimators of

$$\boldsymbol{\theta}_0 \;=\; \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} :\;\; g(\boldsymbol{\theta};\, \mathbb{E}\, b(\boldsymbol{X})) \;=\; \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}}\; g(\dot{\boldsymbol{\theta}};\, \mathbb{E}\, b(\boldsymbol{X}))\right\}.$$

If, given an observation $\boldsymbol{x}$ of a random graph $\boldsymbol{X}$, the estimating function $g(\boldsymbol{\theta};\, b(\boldsymbol{x}))$ is close to $g(\boldsymbol{\theta};\, \mathbb{E}\, b(\boldsymbol{X}))$, then we would expect the minimizers $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ of $g(\boldsymbol{\theta}; b(\boldsymbol{x}))$ and $g(\boldsymbol{\theta}; \mathbb{E}\, b(\boldsymbol{X}))$ to be close under suitable conditions. To show that $\widehat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}_0$ with high probability, we make the following assumptions. In the following, $\mathbb{B}$ denotes the convex hull of the set $\{b(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\}$.

[C.1] The expectation $\mathbb{E}\, b(\boldsymbol{X}) \in \text{rint}(\mathbb{B})$ exists and there exists $A > 0$ such that, for all $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X}$,

$$\|b(\boldsymbol{x}_1) - b(\boldsymbol{x}_2)\|_\infty \;\leq\; A\, d(\boldsymbol{x}_1, \boldsymbol{x}_2)\, \|\mathcal{A}\|_\infty.$$

[C.2] The parameter space $\boldsymbol{\Theta}$ is an open subset of $\mathbb{R}^q$, the expectation $\boldsymbol{\beta}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and there exists a unique element $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $\boldsymbol{\beta}(\boldsymbol{\theta}_0) = \mathbb{E}\, b(\boldsymbol{X}) \in \text{rint}(\mathbb{B})$.

[C.3] For all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \boldsymbol{\Theta}$, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$,

$$\|\boldsymbol{\beta}(\boldsymbol{\theta}_0) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2 \;\geq\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \quad \text{for some} \quad 0 \leq \alpha \leq 1.$$

Conditions [C.1]—[C.3] are verified in Corollary 4. Conditions [C.1] and [C.3] resemble the conditions of Theorem 1 and are verified in Corollaries 1 and 3 in the special case $b(\boldsymbol{x}) = s(\boldsymbol{x})$. Condition [C.2] is a moment-matching condition: It assumes that the family of distributions parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which may not be the data-generating exponential family, is able to match the first moment $\mathbb{E}\, b(\boldsymbol{X})$ of $b(\boldsymbol{X})$ under the data-generating exponential-family distribution. The unique parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ that matches the moment $\mathbb{E}\, b(\boldsymbol{X})$ is the unique minimizer of $g(\boldsymbol{\theta};\, \mathbb{E}\, b(\boldsymbol{X}))$, i.e., $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\, g(\boldsymbol{\theta};\, \mathbb{E}\, b(\boldsymbol{X}))$.

B.2. Concentration results. We establish concentration results for the class of M-estimators introduced in Section B.1.

To do so, we need to show that the probability mass of statistic vector $b(\boldsymbol{X})$ concentrates around its expectation $\mathbb{E}\, b(\boldsymbol{X})$ under the data-generating exponential-family distribution.

Proposition 4. *Consider an exponential family with countable support $\mathbb{X}$ and local dependence. Let $b : \mathbb{X} \mapsto \mathbb{R}^m$ and assume that condition [C.1] is satisfied. Then there exists $C > 0$ such that, for all deviations of the form $t = \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}$ with $\delta > 0$ and $0 \leq \alpha \leq 1$,*

$$\mathbb{P}\left(\|b(\boldsymbol{X}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \,\geq\, t\right) \quad\leq\quad 2\, \exp\left(-\frac{\delta^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m\right).$$

Proposition 4 paves the way for concentration results for the class of M-estimators introduced in Section B.1. The following concentration result shows that M-estimators $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\, g(\boldsymbol{\theta};\, b(\boldsymbol{X}))$ based on estimating functions of the form (B.1) are close to $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\, g(\boldsymbol{\theta};\, \mathbb{E}\, b(\boldsymbol{X}))$ with high probability as long as the neighborhoods and m are small relative to the number of neighborhoods K.

Theorem 3. *Consider an exponential-family random graph with countable support $\mathbb{X}$ and local dependence. Assume that conditions [C.1]—[C.3] are satisfied. Then, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exist $\kappa(\epsilon) > 0$ and $C > 0$ such that*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \,\in\, \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)\right) \,\geq\, 1 - 2\, \exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m\right).$$

Conditions [C.1]—[C.3] of Theorem 3 are verified by Corollary 4. It is worth noting that $\widehat{\boldsymbol{\theta}}$ may not be unique, but an argument along the lines of Section 3.3 shows that, with high probability, the minimizers of estimating function (B.1) do

not give rise to global minima that are separated by large distances, under the stated assumptions.

Example: misspecified exponential family with omitted covariate term. To demonstrate, we consider an extension of the exponential-family random graph with within-neighborhood edge and transitive edge terms, as described in Appendix A. The extended model includes an additional same-attribute edge term with natural parameters $\eta_{k,3}(\boldsymbol{\theta}) = \theta_3$ and sufficient statistics $s_{k,3}(\boldsymbol{x}_k)$,

$$s_{k,3}(\boldsymbol{x}_k) \quad = \quad \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} x_{i,j}\, \mathbb{1}(c_i = c_j),$$

where $\mathbb{1}(c_i = c_j)$ is an indicator function, which is 1 if $c_i = c_j$ and is 0 otherwise, and $c_i \in \{C_1, \ldots, C_H\}$ ($C_h \in \mathbb{R},\ h = 1, \ldots, H,\ H \geq 2$) is a categorical attribute of node $i \in \mathcal{A}_k$ ($k = 1, \ldots, K$). Same-attribute edge terms are popular in applications and capture excesses in the expected number of edges among nodes with the same attribute [e.g., 24]: e.g., students may prefer to befriend other students of the same region of origin.

Suppose that researchers are unaware that the attributes c_i of nodes i are important predictors of edges and estimate the edge and transitive edge parameters θ_1 and θ_2 based on the misspecified exponential family with natural parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and sufficient statistic vector $b(\boldsymbol{x}) = (s_1(\boldsymbol{x}), s_2(\boldsymbol{x}))$, where $s_1(\boldsymbol{x}) = \sum_{k=1}^{K} s_{k,1}(\boldsymbol{x}_k)$ and $s_2(\boldsymbol{x}) = \sum_{k=1}^{K} s_{k,2}(\boldsymbol{x}_k)$ are defined in Appendix A. In other words, suppose that statistical inference is based on the estimating function

$$g(\boldsymbol{\theta};\ b(\boldsymbol{x})) \quad = \quad \|b(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})\|_2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_0,$$

where $\boldsymbol{\Theta}_0 = \mathbb{R} \times \mathbb{R}^+$ is the parameter space of the misspecified exponential family, which is a two-dimensional subspace of the three-dimensional parameter space $\boldsymbol{\Theta}^\star = \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ of the data-generating exponential family. The following concentration result shows that the M-estimator $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} g(\boldsymbol{\theta};\ b(\boldsymbol{X}))$ is close to $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} g(\boldsymbol{\theta};\ \mathbb{E}\, b(\boldsymbol{X}))$ with high probability as long as the neighborhoods are small relative to the number of neighborhoods K.

Corollary 4. *Consider an exponential-family random graph with within-neighborhood edge, transitive edge, and same-attribute edge terms. Suppose that an observation of the random graph is generated by $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta}^\star$. Then all conditions of Theorem 3 are satisfied and hence, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \boldsymbol{\Theta}_0$, there exist $\kappa(\epsilon) > 0$ and $C > 0$ such that $\widehat{\boldsymbol{\theta}}$ exists, is unique, and*

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \ \in \ \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)\right) \quad \geq \quad 1 - 4\, \exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{\|\mathcal{A}\|_\infty^4}\right),$$

provided $|\mathcal{A}_k| \geq 3$ *(*$k = 1, \ldots, K$*) and* $K \geq 2$.

It is worth noting that $\boldsymbol{\theta}_0 \in \mathbb{R} \times \mathbb{R}^+$ and $\boldsymbol{\theta}^\star \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ do not have the same dimension and cannot be identical, but the distribution parameterized by $\boldsymbol{\theta}_0$ is as close as possible to the data-generating distribution parameterized by $\boldsymbol{\theta}^\star$ in terms of Kullback-Leibler divergence: by construction, $\boldsymbol{\theta}_0$ minimizes $\|\mathbb{E}\, b(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})\|_2 = \|\mathbb{E}\, \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{X})\|_2$ and hence maximizes $\mathbb{E} \log p_{\boldsymbol{\theta}}(\boldsymbol{X})$, where $p_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp(\langle \boldsymbol{\theta}, b(\boldsymbol{x}) \rangle)$ denotes the density of $\boldsymbol{x} \in \mathbb{X}$ under the misspecified exponential-family distribution with natural parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and sufficient statistic vector $b(\boldsymbol{x}) = (s_1(\boldsymbol{x}), s_2(\boldsymbol{x}))$. Therefore, $\boldsymbol{\theta}_0$ satisfies

$$\boldsymbol{\theta}_0 = \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbb{E}\, \log p_{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathbb{E}\, \log p_{\boldsymbol{\theta}^\star}(\boldsymbol{X}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} KL(\mathbb{P}_{\boldsymbol{\theta}^\star}; \mathbb{P}_{\boldsymbol{\theta}}), \tag{B.2}$$

where $KL(\mathbb{P}_{\boldsymbol{\theta}^\star}; \mathbb{P}_{\boldsymbol{\theta}})$ is the Kullback-Leibler divergence from $\mathbb{P}_{\boldsymbol{\theta}^\star}$ to $\mathbb{P}_{\boldsymbol{\theta}}$ and the expectations $\mathbb{E}\, \log p_{\boldsymbol{\theta}}(\boldsymbol{X})$ and $\mathbb{E}\, \log p_{\boldsymbol{\theta}^\star}(\boldsymbol{X})$ are with respect to $\mathbb{P}_{\boldsymbol{\theta}^\star}$. Owing to the dependence of within-neighborhood edge variables and sufficient statistics, it is not straightforward to bound $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_0\|_2$, but—by the properties of maximum likelihood estimation—we are assured that the distribution parameterized by $\boldsymbol{\theta}_0$ is as close as possible to the distribution parameterized by the data-generating parameter vector $\boldsymbol{\theta}^\star$ in terms of Kullback-Leibler divergence, as shown by (B.2).

APPENDIX C: PROOFS: CONCENTRATION RESULTS FOR RANDOM GRAPHS WITH DEPENDENT EDGES

We prove the main concentration results of Sections 3.2 and 3.4 and Appendix B, Propositions 1, 2, and 4.

Proof of Proposition 1. By assumption, $\mathbb{E}\, f(\boldsymbol{X})$ exists. We are interested in deviations of the form $|f(\boldsymbol{X}) - \mathbb{E}\, f(\boldsymbol{X})| \geq t$, where $t > 0$. Choose any $t > 0$ and let $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{X} : |f(\boldsymbol{x}) - \mathbb{E}\, f(\boldsymbol{X})| \geq t\}$. Since within-neighborhood edges do not depend on between-neighborhood edges,

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{X}, \boldsymbol{Y} \in \mathbb{Y}) \quad = \quad \mathbb{P}(\boldsymbol{X} \in \mathcal{X}).$$

In the following, we denote by $\mathbb{P}$ a probability measure on $(\mathbb{X}, \mathbb{S})$ with densities of the form (2.1), where $\mathbb{S}$ is the power set of the countable set $\mathbb{X}$. Keep in mind that $\boldsymbol{X} = (\boldsymbol{X}_k)_{k=1}^K$ denotes the sequence of within-neighborhood edge variables, where $\boldsymbol{X}_k = (X_{i,j})_{i \in \mathcal{A}_k < j \in \mathcal{A}_k}$. In an abuse of notation, we denote the elements of the sequence of edge variables $\boldsymbol{X}$ by $X_1, \ldots, X_w$ with sample spaces $\mathbb{X}_1, \ldots, \mathbb{X}_w$, respectively, where $w = \sum_{k=1}^K \binom{|\mathcal{A}_k|}{2}$ is the number of within-neighborhood edge variables. Let $\boldsymbol{X}_{i:j} = (X_i, \ldots, X_j)$ be a subsequence of edge variables with sample space $\mathbb{X}_{i:j}$, where $i \leq j$. By applying Theorem 1.1 of Kontorovich and Ramanan [32] to $\|f\|_{\text{Lip}}$-Lipschitz functions $f : \mathbb{X} \mapsto \mathbb{R}$ defined on the countable set

$\mathbb{X}$,

$$\mathbb{P}(|f(\boldsymbol{X}) - \mathbb{E}\, f(\boldsymbol{X})| \;\geq\; t) \;\;\leq\;\; 2\,\exp\left(-\frac{t^2}{2\,w\,\|\Phi\|_\infty^2\,\|f\|_{\text{Lip}}^2}\right),$$

where Φ is the $w \times w$-upper triangular matrix with entries

$$\phi_{i,j} \;\;=\;\; \begin{cases} \varphi_{i,j} & \text{if } i < j \\ 1 & \text{if } i = j \\ 0 & \text{if } i > j \end{cases}$$

and

$$\|\Phi\|_\infty \;\;=\;\; \max_{1 \leq i \leq w} \left| 1 + \sum_{j=i+1}^{w} \varphi_{i,j} \right|.$$

The coefficients $\varphi_{i,j}$ are known as mixing coefficients and are defined by

$$\varphi_{i,j} \equiv \sup_{\substack{\boldsymbol{x}_{1:i-1} \in \mathbb{X}_{1:i-1} \\ (x_i, x_i^\star) \in \mathbb{X}_i \times \mathbb{X}_i}} \varphi_{i,j}(\boldsymbol{x}_{1:i-1}, x_i, x_i^\star) \;\;=\;\; \sup_{\substack{\boldsymbol{x}_{1:i-1} \in \mathbb{X}_{1:i-1} \\ (x_i, x_i^\star) \in \mathbb{X}_i \times \mathbb{X}_i}} \|\pi_{x_i} - \pi_{x_i^\star}\|_{\text{TV}},$$

where $\|\pi_{x_i} - \pi_{x_i^\star}\|_{\text{TV}}$ is the total variation distance between the distributions π_{x_i} and $\pi_{x_i^\star}$ given by

$$\pi_{x_i} \equiv \pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i) = \mathbb{P}(\boldsymbol{X}_{j:w} = \boldsymbol{x}_{j:w} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1}, X_i = x_i)$$

and

$$\pi_{x_i^\star} \equiv \pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i^\star) = \mathbb{P}(\boldsymbol{X}_{j:w} = \boldsymbol{x}_{j:w} \mid \boldsymbol{X}_{1:i-1} = \boldsymbol{x}_{1:i-1}, X_i = x_i^\star).$$

Since the support of π_{x_i} and $\pi_{x_i^\star}$ is countable,

$$\|\pi_{x_i} - \pi_{x_i^\star}\|_{\text{TV}} \;\;=\;\; \frac{1}{2} \sum_{\boldsymbol{x}_{j:w} \in \mathbb{X}_{j:w}} |\pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i) - \pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i^\star)|.$$

An upper bound on $\|\Phi\|_\infty$ can be obtained by bounding the mixing coefficients $\varphi_{i,j}$ as follows. Consider any pair of edge variables X_i and X_j. If X_i and X_j involve nodes in more than one neighborhood, the mixing coefficient $\varphi_{i,j}$ vanishes by the local dependence induced by exponential families of the form (2.1). If the pair of nodes corresponding to X_i and the pair of nodes corresponding to X_j belong to the same neighborhood, the mixing coefficient $\varphi_{i,j}$ can be bounded as follows:

$$\varphi_{i,j}(\boldsymbol{x}_{1:i-1}, x_i, x_i^\star) \;=\; \frac{1}{2} \sum_{\boldsymbol{x}_{j:w} \in \mathbb{X}_{j:w}} |\pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i) - \pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i^\star)|$$

$$\leq\; \frac{1}{2} \sum_{\boldsymbol{x}_{j:w} \in \mathbb{X}_{j:w}} \pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i) + \frac{1}{2} \sum_{\boldsymbol{x}_{j:w} \in \mathbb{X}_{j:w}} \pi(\boldsymbol{x}_{j:w} \mid \boldsymbol{x}_{1:i-1}, x_i^\star) \;=\; 1,$$

because π_{x_i} and $\pi_{x_i^\star}$ are conditional probability mass functions with countable support $\mathbb{X}_{j:w}$. We note that the upper bound is not sharp, but it has the advantage that it covers a wide range of dependencies within neighborhoods. Thus,

$$\|\Phi\|_\infty \;=\; \max_{1\le i\le w}\left|1+\sum_{j=i+1}^{w}\varphi_{i,j}\right| \;\le\; \binom{\|\mathcal{A}\|_\infty}{2},$$

because each edge variable X_i can depend on at most $\binom{\|\mathcal{A}\|_\infty}{2}-1$ other edge variables corresponding to pairs of nodes belonging to the same pair of neighborhoods. Therefore, there exists $C>0$ such that, for all $t>0$,

$$\mathbb{P}(|f(\boldsymbol{X})-\mathbb{E}\,f(\boldsymbol{X})|\;\ge\;t)\;\le\;2\,\exp\left(-\frac{t^2}{C\,\sum_{k=1}^{K}\binom{|\mathcal{A}_k|}{2}\,\|\mathcal{A}\|_\infty^4\,\|f\|_{\text{Lip}}^2}\right),$$

where $\|\mathcal{A}\|_\infty \ge 1$ because all neighborhoods $\mathcal{A}_k$ are non-empty and $\|f\|_{\text{Lip}}>0$ by assumption.

Proof of Proposition 2. By assumption, the neighborhood-dependent natural parameters $\eta_{k,i}(\boldsymbol{\theta})$ are of the form $\eta_{k,i}(\boldsymbol{\theta})=\eta_i(\boldsymbol{\theta})$ $(i=1,\ldots,m_k, k=1,\ldots,K)$. Therefore, the exponential family can be reduced to an exponential family with natural parameter vector

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \;=\; (\eta_1(\boldsymbol{\theta}),\ldots,\eta_m(\boldsymbol{\theta}))$$

and sufficient statistic vector

$$s(\boldsymbol{x}) \;=\; (s_1(\boldsymbol{x}),\ldots,s_m(\boldsymbol{x}))\,,$$

where $m=\max_{1\le k\le K} m_k$. Here, the sufficient statistics $s_i(\boldsymbol{x})$ are sums of within-neighborhood sufficient statistics $s_{k,i}(\boldsymbol{x}_k)$ $(k=1,\ldots,K)$:

$$s_i(\boldsymbol{x}) \;=\; \sum_{k=1}^{K} s_{k,i}(\boldsymbol{x}_k),\quad i=1,\ldots,m.$$

Observe that $\boldsymbol{\mu}(\widehat{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)})=s(\boldsymbol{X})$ and that $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))=\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}\,s(\boldsymbol{X})=\mathbb{E}\,s(\boldsymbol{X})\in \text{rint}(\mathbb{M})$ exists, because $\boldsymbol{\eta}(\boldsymbol{\theta}^\star)\in\text{int}(\mathfrak{N})$ [5, Theorem 2.2, pp. 34–35]. We bound the probability of deviations of $s(\boldsymbol{X})$ from $\mathbb{E}\,s(\boldsymbol{X})$ in terms of the ℓ_∞- and ℓ_2-norm below.

ℓ_∞-norm. For all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\mu}(\widehat{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_\infty \;\geq\; \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \\
=\;& \mathbb{P}\left(\|s(\boldsymbol{X}) - \mathbb{E}\, s(\boldsymbol{X})\|_\infty \;\geq\; \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \\
\leq\;& \mathbb{P}\left(\bigcup_{i=1}^{m} \left(|s_i(\boldsymbol{X}) - \mathbb{E}\, s_i(\boldsymbol{X})| \;\geq\; \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \right).
\end{aligned}$$

By condition (3.3) of Proposition 2, there exists $A > 0$ such that the Lipschitz coefficient of $s_i : \mathbb{X} \mapsto \mathbb{R}$ with respect to the Hamming metric $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_0^+$ is bounded above by $\|s_i\|_{\text{Lip}} \leq A\, \|\mathcal{A}\|_\infty$ $(i = 1, \ldots, m)$. Thus, by a union bound over the m components of $s(\boldsymbol{X})$ and by applying Proposition 1 to deviations of size $t = \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha$, we obtain, for all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i=1}^{m} \left(|s_i(\boldsymbol{X}) - \mathbb{E}\, s_i(\boldsymbol{X})| \;\geq\; \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \right) \\
\leq\;& 2 \exp\left(- \frac{\delta^2 \left(\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right)^2}{B \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2} \|\mathcal{A}\|_\infty^4 \|\mathcal{A}\|_\infty^2} + \log m \right),
\end{aligned}$$

where $B > 0$. Since all neighborhoods are of the same order of magnitude, there exist $C_1 > 0$ and $C_2 > 0$ such that $C_1 \|\mathcal{A}\|_\infty \leq |\mathcal{A}_k| \leq C_2 \|\mathcal{A}\|_\infty$. Thus, there exists $C_3 > 0$ such that the term $(\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha)^2$ in the numerator of the exponent can be bounded below by

$$\left(\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right)^2 \;\geq\; C_3\, K^2\, \|\mathcal{A}\|_\infty^{4\alpha},$$

and there exists $C_4 > 0$ such that the term $\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}$ in the denominator of the exponent can be bounded above by

$$\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2} \;\leq\; C_4\, K\, \|\mathcal{A}\|_\infty^2.$$

As a result, there exists $C > 0$ such that, for all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\mu}(\widehat{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_\infty \;\geq\; \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \\
\leq\;& 2\exp\left(-\frac{\delta^2\, C\, K}{\|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right).
\end{aligned}$$

ℓ_2-norm. The same argument used above shows that, for all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\mu}(\widehat{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \geq \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \\
\leq\;& \mathbb{P}\left(\|\boldsymbol{\mu}(\widehat{\boldsymbol{\eta}(\boldsymbol{\theta}^\star)}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_\infty \geq \frac{\delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha}{\sqrt{m}} \right) \\
\leq\;& 2\exp\left(-\frac{\delta^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right).
\end{aligned}$$

Proof of Proposition 4. Condition [C.1] implies that $\mathbb{E}\, b(\boldsymbol{X})$ exists. Thus, for all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\|b(\boldsymbol{X}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;\geq\; \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right) \\
\leq\;& \mathbb{P}\left(\|b(\boldsymbol{X}) - \mathbb{E}\, b(\boldsymbol{X})\|_\infty \geq \frac{\delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha}{\sqrt{m}} \right) \\
\leq\;& \mathbb{P}\left(\bigcup_{i=1}^{m} \left(|b_i(\boldsymbol{X}) - \mathbb{E}\, b_i(\boldsymbol{X})| \;\geq\; \frac{\delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha}{\sqrt{m}} \right) \right).
\end{aligned}$$

By condition [C.1], there exists $A > 0$ such that the Lipschitz coefficient of $b_i : \mathbb{X} \mapsto \mathbb{R}$ with respect to the Hamming metric $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_0^+$ is bounded above by $\|b_i\|_{\text{Lip}} \leq A\ \|\mathcal{A}\|_\infty$ $(i = 1, \ldots, m)$. Thus, by a union bound over the m components of $b(\boldsymbol{X})$ and by applying Proposition 1 to deviations of size $t = \delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha / \sqrt{m}$, we obtain, for all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i=1}^{m} \left(|b_i(\boldsymbol{X}) - \mathbb{E}\, b_i(\boldsymbol{X})| \;\geq\; \frac{\delta \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha}{\sqrt{m}} \right) \right) \\
\leq\;& 2\exp\left(-\frac{\delta^2 \left(\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^\alpha \right)^2}{B\, m \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2} \|\mathcal{A}\|_\infty^4\, \|\mathcal{A}\|_\infty^2} + \log m \right),
\end{aligned}$$

where $B > 0$. We showed in the proof of Proposition 2 that there exist $C_1 > 0$ and $C_2 > 0$ such that $(\sum_{k=1}^K \binom{|\mathcal{A}_k|}{2}^\alpha)^2 \geq C_1 \; K^2 \; \|\mathcal{A}\|_\infty^{4\,\alpha}$ and $\sum_{k=1}^K \binom{|\mathcal{A}_k|}{2} \leq C_2 \, K \, \|\mathcal{A}\|_\infty^2$. As a consequence, there exists $C > 0$ such that, for all $\delta > 0$,

$$\begin{aligned} & \mathbb{P}\left(\|b(\boldsymbol{X}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \; \geq \; \delta \; \sum_{k=1}^K \binom{|\mathcal{A}_k|}{2}^\alpha \right) \\ \leq \; & 2 \, \exp\left(- \frac{\delta^2 \; C \; K}{m \, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right). \end{aligned}$$

APPENDIX D: PROOFS: CONCENTRATION RESULTS FOR MAXIMUM LIKELIHOOD ESTIMATORS

We prove the main concentration results of Section 3, Theorem 1 along with Corollaries 1 and 3. Auxiliary lemmas are proved in Appendix G.

Proof of Theorem 1. By assumption, $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$. Observe that $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$ implies $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \in \text{rint}(\mathbb{M})$, because the maps $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ and $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ are one-to-one: the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one by assumption while the map $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one by classic exponential-family theory [5, Theorem 3.6, p. 74].

We are interested in estimators of the form

$$\widehat{\boldsymbol{\theta}} \;\; = \;\; \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \;\; \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \;\; = \;\; \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\dot{\boldsymbol{\theta}}))\|_2 \right\},$$

where

$$\boldsymbol{\Theta} \;\; \subseteq \;\; \{\boldsymbol{\theta} \in \mathbb{R}^q : \; \psi(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}.$$

The following proof covers both full and non-full exponential families, including curved exponential families. We first discuss the existence of $\widehat{\boldsymbol{\theta}}$ in full and non-full, curved exponential families and then bound the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$.

Existence of $\widehat{\boldsymbol{\theta}}$: full and non-full, curved exponential families. For any given $\omega > 0$, let

$$\mathbb{M}(\omega) \;\; = \;\; \left\{ \boldsymbol{\mu}' \in \mathbb{M} : \; \|\boldsymbol{\mu}' - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \; < \; \omega \; \sum_{k=1}^K \binom{|\mathcal{A}_k|}{2}^\alpha \right\}$$

be the subset of $\boldsymbol{\mu}' \in \mathbb{M}$ close to the data-generating mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \in \text{rint}(\mathbb{M})$ in the sense that $\|\boldsymbol{\mu}' - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \; < \; \omega \, \sum_{k=1}^K \binom{|\mathcal{A}_k|}{2}^\alpha$.

Choose $\omega > 0$ small enough so that $\mathbb{M}(2\,\omega) \subseteq \text{rint}(\mathbb{M})$ and let

$$\mathcal{G}(\omega) \;=\; \left\{ \boldsymbol{x} \in \mathbb{X} :\; \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;<\; \omega \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \right\}$$

be the subset of $\boldsymbol{x} \in \mathbb{X}$ such that $s(\boldsymbol{x}) \in \mathbb{M}(\omega) \subseteq \text{rint}(\mathbb{M})$. In the following, we show that in the event $\mathcal{G}(\omega)$ the set $\widehat{\boldsymbol{\theta}}$ is non-empty, i.e., $\widehat{\boldsymbol{\theta}}$ exists. To see that, let $\mathbb{M}(\boldsymbol{\Theta})$ be the set of mean-value parameter vectors induced by $\boldsymbol{\Theta}$, i.e.,

$$\mathbb{M}(\boldsymbol{\Theta}) \;=\; \{\boldsymbol{\mu}' \in \text{rint}(\mathbb{M}) :\; \text{there exists } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ such that } \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \boldsymbol{\mu}'\}\,.$$

If the exponential family is full, then $\mathbb{M}(\boldsymbol{\Theta}) = \text{rint}(\mathbb{M})$, otherwise $\mathbb{M}(\boldsymbol{\Theta}) \subset \text{rint}(\mathbb{M})$, because non-full exponential families exclude some natural parameter vectors along with the corresponding mean-value parameter vectors. To show that the set $\widehat{\boldsymbol{\theta}}$ is non-empty in the event $\mathcal{G}(\omega)$, note that, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$ and hence all $s(\boldsymbol{x}) \in \mathbb{M}(\omega) \subseteq \text{rint}(\mathbb{M})$, there exists at least one element $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathbb{M}(\boldsymbol{\Theta}) \subseteq \text{rint}(\mathbb{M})$ such that

$$\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))\|_2 \;\leq\; \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2,$$

because the data-generating mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \in \mathbb{M}(\boldsymbol{\Theta}) \subseteq \text{rint}(\mathbb{M})$ is known to be an element of $\mathbb{M}(\boldsymbol{\Theta}) \subseteq \text{rint}(\mathbb{M})$ and any minimizer of $\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2$ cannot be farther from $s(\boldsymbol{x}) \in \mathbb{M}(\omega) \subseteq \text{rint}(\mathbb{M})$ than the data-generating mean-value parameter vector $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \in \mathbb{M}(\boldsymbol{\Theta}) \subseteq \text{rint}(\mathbb{M})$. In addition, for each element $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathbb{M}(\boldsymbol{\Theta}) \subseteq \text{rint}(\mathbb{M})$, we have

$$\begin{aligned} \|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;&\leq\; \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))\|_2 + \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \\ &\leq\; 2\,\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2, \end{aligned}$$

which implies that each element $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathbb{M}(\boldsymbol{\Theta}) \subseteq \text{rint}(\mathbb{M})$ is contained in $\mathbb{M}(2\,\omega) \subseteq \text{rint}(\mathbb{M})$; note that $\omega > 0$ was chosen small enough so that $\mathbb{M}(2\,\omega) \subseteq \text{rint}(\mathbb{M})$. Therefore, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$, the set $\widehat{\boldsymbol{\theta}}$ contains at least one element. If there exists a unique element $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) = s(\boldsymbol{x}) \in \mathbb{M}(\omega) \subseteq \text{rint}(\mathbb{M})$, then the set $\widehat{\boldsymbol{\theta}}$ contains one and only one element. In particular, in full exponential families with $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$, there exists a unique element $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) = s(\boldsymbol{x}) \in \mathbb{M}(\omega) \subseteq \text{rint}(\mathbb{M})$ [5, Theorem 3.6, p. 74]. Therefore, in full exponential families $\widehat{\boldsymbol{\theta}}$ exists and is unique in the event $\mathcal{G}(\omega)$, whereas in non-full exponential families $\widehat{\boldsymbol{\theta}}$ exists but may not be unique in the event $\mathcal{G}(\omega)$.

Bounding the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$: full and non-full, curved exponential families. By the identifiability conditions of Theorem 1, for all $\epsilon > 0$

small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we have $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$. In addition, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$,

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\; \geq \;\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\; \text{for some} \;\; 0 \leq \alpha \leq 1.$$

Therefore,

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\; < \;\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}$$

implies $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$, which in turn implies $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$.

To bound the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we exploit the fact that, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$ and hence all $s(\boldsymbol{x}) \in \mathbb{M}(\omega) \subseteq \text{rint}(\mathbb{M})$, the set $\widehat{\boldsymbol{\theta}}$ is non-empty and, for each element of the set $\widehat{\boldsymbol{\theta}}$,

$$\begin{aligned} \|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\; &\leq \;\; \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}}))\|_2 + \|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \\ &\leq \;\; 2\,\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2, \end{aligned}$$

which implies that $\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) \in \mathbb{M}(2\,\omega) \subseteq \text{rint}(\mathbb{M})$. Thus, the probability of event

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\; < \;\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\; \cap \;\; \mathcal{G}(\omega)$$

can be bounded from below by bounding the probability of event

$$2\,\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\; < \;\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\; \cap \;\; \mathcal{G}(\omega),$$

which, by definition of event $\mathcal{G}(\omega)$, is equivalent to bounding the probability of event

$$\|s(\boldsymbol{x}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\; < \;\; \min\left(\frac{\delta(\epsilon)}{2},\, \omega\right) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}.$$

These results show that the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ can be bounded from below as follows:

$$\begin{aligned} &\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \;\subseteq\; \text{int}(\boldsymbol{\Theta})\right) \;\geq\; \mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \;\subseteq\; \text{int}(\boldsymbol{\Theta}) \;\cap\; \mathcal{G}(\omega)\right) \\ \geq \;\; &\mathbb{P}\left(\|\boldsymbol{\mu}(\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;<\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\cap\; \mathcal{G}(\omega)\right) \\ = \;\; &\mathbb{P}\left(\|s(\boldsymbol{X}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;<\; \min\left(\frac{\delta(\epsilon)}{2},\, \omega\right) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}\right). \end{aligned}$$

To bound the probability of the event on the right-hand side of the inequality above, note that the smoothness condition (3.5) of Theorem 1 ensures that the smoothness condition (3.3) of Proposition 2 is satisfied. Therefore, Proposition 2 can be invoked to show that there exists $C > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\|s(\boldsymbol{X}) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))\|_2 \;\geq\; \min\left(\frac{\delta(\epsilon)}{2},\, \omega\right) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \right) \\
\leq\;& 2\,\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right),
\end{aligned}$$

where $\kappa(\epsilon) = \min(\delta(\epsilon)\,/\,2,\; \omega) > 0$. Collecting results shows that

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \;\subseteq\; \text{int}(\boldsymbol{\Theta})\right) \;\geq\; 1 - 2\,\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right).$$

Proof of Corollary 1. Corollary 1 follows from Theorem 1, because all conditions of Theorem 1 are satisfied. First, by Lemma 3 in Appendix G.1, the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one and, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we have $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star),\, \gamma(\epsilon))$. Second, condition (3.4) of Theorem 1 is satisfied with $\alpha = 3/4$ by Lemma 4 in Appendix G.1 provided $|\mathcal{A}_k| \geq 4$ $(k = 1, \ldots, K)$. Last, but not least, condition (3.5) of Theorem 1 is satisfied, because changing an edge cannot change the number of within-neighborhood edges by more than 1 and the number of within-neighborhood connected pairs of nodes with i shared partners by more than $2\,(\|\mathcal{A}\|_\infty - 2) + 1$ $(i = 1, \ldots, |\mathcal{A}_k| - 2,\, k = 1, \ldots, K)$. Thus, by Theorem 1, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exist $\kappa(\epsilon) > 0$ and $C > 0$ such that

$$\begin{aligned}
\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \;\subseteq\; \text{int}(\boldsymbol{\Theta})\right) \;&\geq\; 1 - 2\,\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right) \\
&\geq\; 1 - 2\,\exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{\|\mathcal{A}\|_\infty^{6}} + \log \|\mathcal{A}\|_\infty \right),
\end{aligned}$$

where we used the fact that $m = \max_{1 \leq k \leq K} m_k = \|\mathcal{A}\|_\infty - 1$ and $\alpha = 3/4$.

Proof of Corollary 3. Corollary 3 follows from Theorem 1, because conditions (3.4) and (3.5) of Theorem 1 are satisfied. Condition (3.4) is satisfied with $\alpha = 1$ by Lemma 5 in Appendix G.1 provided $|\mathcal{A}_k| \geq 3$ $(k = 1, \ldots, K)$. Condition

(3.5) is satisfied, because changing an edge cannot change the number of within-neighborhood edges by more than 1 and the number of within-neighborhood transitive edges by more than $2\,(\|\mathcal{A}\|_\infty - 2) + 1$. Thus, Theorem 1 can be invoked to conclude that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exist $\kappa(\epsilon) > 0$, $C_1 > 0$, and $C_2 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})\right) &\geq\ 1 - 2\,\exp\left(-\frac{\kappa(\epsilon)^2\, C_1\, K}{q\,\|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log q\right) \\ &=\ 1 - 4\,\exp\left(-\frac{\kappa(\epsilon)^2\, C_2\, K}{\|\mathcal{A}\|_\infty^4}\right),\end{aligned}$$

where we used the fact that $q = 2$ and $\alpha = 1$.

APPENDIX E: PROOFS: CONCENTRATION RESULTS FOR M-ESTIMATORS

We prove the main concentration result of Appendix B, Theorem 3 along with Corollary 4.

Proof of Theorem 3. Theorem 3 can be proved along the same lines as Theorem 1.

By condition [C.1], the expectation $\mathbb{E}\, b(\boldsymbol{X})$ of $b(\boldsymbol{X})$ under the data-generating exponential-family distribution exists. Let $\mathbb{B}$ be the convex hull of the set $\{b(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\}$ and, given any $\omega > 0$, let

$$\mathbb{B}(\omega) \quad = \quad \left\{\boldsymbol{\beta}' \in \mathbb{B} :\ \|\boldsymbol{\beta}' - \mathbb{E}\, b(\boldsymbol{X})\|_2 \ < \ \omega \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}\right\}$$

be the subset of $\boldsymbol{\beta}' \in \mathbb{B}$ close to the expectation $\mathbb{E}\, b(\boldsymbol{X})$ of $b(\boldsymbol{X})$ under the data-generating exponential-family distribution in the sense that $\|\boldsymbol{\beta}' - \mathbb{E}\, b(\boldsymbol{X})\|_2 < \omega \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}$. Choose $\omega > 0$ small enough so that $\mathbb{B}(2\,\omega) \subseteq \text{rint}(\mathbb{B})$ and let

$$\mathcal{G}(\omega) \quad = \quad \left\{\boldsymbol{x} \in \mathbb{X} :\ \|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 < \omega \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}\right\}$$

be the subset of $\boldsymbol{x} \in \mathbb{X}$ such that $b(\boldsymbol{x}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$.

We are interested in estimators of the form

$$\widehat{\boldsymbol{\theta}} \quad = \quad \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} :\ \ \|b(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2 \quad = \quad \inf_{\dot{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} \|b(\boldsymbol{x}) - \boldsymbol{\beta}(\dot{\boldsymbol{\theta}})\|_2\right\},$$

where $\boldsymbol{\Theta}$ is an open subset of $\mathbb{R}^q$ and the expectation $\boldsymbol{\beta}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ by condition [C.2].

In the following, we do not assume that the family of distributions parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is an exponential family and we do not assume that the set of expectations $\{\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ covers the whole set $\text{rint}(\mathbb{B})$. In other words, the family may not be able to match all possible $b(\boldsymbol{x}) \in \text{rint}(\mathbb{B})$ in the sense that there exists $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X}) = b(\boldsymbol{x}) \in \text{rint}(\mathbb{B})$. The critical assumption exploited by the following proof is that the family can match the expectation $\mathbb{E}\, b(\boldsymbol{X})$ of $b(\boldsymbol{X})$ under the data-generating exponential-family distribution in the sense that there exists one and only one element $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $\mathbb{E}_{\boldsymbol{\theta}_0}\, b(\boldsymbol{X}) = \mathbb{E}\, b(\boldsymbol{X}) \in \text{rint}(\mathbb{B})$, along with the fact that $b(\boldsymbol{X})$ falls with high probability into a small ball centered at $\mathbb{E}_{\boldsymbol{\theta}_0}\, b(\boldsymbol{X}) = \mathbb{E}\, b(\boldsymbol{X}) \in \text{rint}(\mathbb{B})$ under suitable conditions. We first discuss the existence of $\widehat{\boldsymbol{\theta}}$ and then bound the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$.

Existence of $\widehat{\boldsymbol{\theta}}$. We show that in the event $\mathcal{G}(\omega)$ the set $\widehat{\boldsymbol{\theta}}$ is non-empty, i.e., $\widehat{\boldsymbol{\theta}}$ exists. To show that the set $\widehat{\boldsymbol{\theta}}$ is non-empty in the event $\mathcal{G}(\omega)$, note that, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$ and hence all $b(\boldsymbol{x}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$, there exists at least one element $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that

$$\|b(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2 \;\leq\; \|b(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{\theta}_0)\|_2 \;=\; \|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2,$$

because, by condition [C.2], there exists $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $\boldsymbol{\beta}(\boldsymbol{\theta}_0) = \mathbb{E}\, b(\boldsymbol{X}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$ and any minimizer of $\|b(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2$ cannot be farther from $b(\boldsymbol{x}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$ than $\boldsymbol{\beta}(\boldsymbol{\theta}_0) = \mathbb{E}\, b(\boldsymbol{X}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$. Therefore, the set $\widehat{\boldsymbol{\theta}}$ is non-empty in the event $\mathcal{G}(\omega)$. In addition, for each element of the set $\widehat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}\|\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;&\leq\; \|b(\boldsymbol{x}) - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\|_2 + \|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \\ &\leq\; 2\,\|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2,\end{aligned}$$

which implies that $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) \in \mathbb{B}(2\,\omega) \subseteq \text{rint}(\mathbb{B})$; note that $\omega > 0$ was chosen small enough so that $\mathbb{B}(2\,\omega) \subseteq \text{rint}(\mathbb{B})$. Therefore, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$ and hence all $b(\boldsymbol{x}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$, the set $\widehat{\boldsymbol{\theta}}$ contains at least one element. If the set $\widehat{\boldsymbol{\theta}}$ contains more than one element, then all elements of the set $\widehat{\boldsymbol{\theta}}$ map to expectations $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\widehat{\boldsymbol{\theta}}}\, b(\boldsymbol{X}) \in \mathbb{B}(2\,\omega) \subseteq \text{rint}(\mathbb{B})$ that have the same ℓ_2-distance from $b(\boldsymbol{x}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$ by construction of estimating function $\|b(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{\theta})\|_2$.

Bounding the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$. To bound the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$, note that, by condition [C.2], the expectation $\boldsymbol{\beta}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and, by condition [C.3], for all $\epsilon > 0$ small enough so that

$\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \boldsymbol{\Theta}$, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$,

$$\|\boldsymbol{\beta}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\boldsymbol{\theta}_0)\|_2 \;=\; \|\boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;\geq\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}$$

for some $0 \leq \alpha \leq 1$. Therefore,

$$\|\boldsymbol{\beta}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\boldsymbol{\theta}_0)\|_2 \;=\; \|\boldsymbol{\beta}(\boldsymbol{\theta}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;<\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}$$

implies $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$.

To bound the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$, we exploit the fact that, for all $\boldsymbol{x} \in \mathcal{G}(\omega)$ and hence all $b(\boldsymbol{x}) \in \mathbb{B}(\omega) \subseteq \text{rint}(\mathbb{B})$, the set $\widehat{\boldsymbol{\theta}}$ is non-empty and, for each element of the set $\widehat{\boldsymbol{\theta}}$,

$$\begin{aligned} \|\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;&\leq\; \|b(\boldsymbol{x}) - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\|_2 + \|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \\ &\leq\; 2\, \|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2, \end{aligned}$$

which implies that $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) \in \mathbb{B}(2\,\omega) \subseteq \text{rint}(\mathbb{B})$, as pointed out above. Thus, the probability of event

$$\|\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;<\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\cap\; \mathcal{G}(\omega)$$

can be bounded from below by bounding the probability of event

$$2\, \|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;<\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\cap\; \mathcal{G}(\omega),$$

which, by definition of event $\mathcal{G}(\omega)$, is equivalent to bounding the probability of event

$$\|b(\boldsymbol{x}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;<\; \min\left(\frac{\delta(\epsilon)}{2}, \omega\right) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}.$$

These results show that the probability of event $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$ can be bounded from below as follows:

$$\begin{aligned} &\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \;\subseteq\; \boldsymbol{\Theta}\right) \;\geq\; \mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \;\subseteq\; \boldsymbol{\Theta} \;\cap\; \mathcal{G}(\omega)\right) \\ \geq\; &\mathbb{P}\left(\|\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;<\; \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \;\cap\; \mathcal{G}(\omega)\right) \\ \geq\; &\mathbb{P}\left(\|b(\boldsymbol{X}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;<\; \min\left(\frac{\delta(\epsilon)}{2}, \omega\right) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha}\right). \end{aligned}$$

To bound the probability of the event on the right-hand side of the inequality above, Proposition 4 can be invoked to show that there exists $C > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\|b(\boldsymbol{X}) - \mathbb{E}\, b(\boldsymbol{X})\|_2 \;\geq\; \min\left(\frac{\delta(\epsilon)}{2},\, \omega\right) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \right) \\
\leq\;& 2\, \exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right),
\end{aligned}$$

where $\kappa(\epsilon) = \min(\delta(\epsilon)\,/\,2,\; \omega) > 0$. Collecting results shows that

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \;\subseteq\; \boldsymbol{\Theta}\right) \;\geq\; 1 - 2\, \exp\left(-\frac{\kappa(\epsilon)^2\, C\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right).$$

Proof of Corollary 4. Corollary 4 follows from Theorem 3, because all conditions of Theorem 3 are satisfied. Condition [C.1] is satisfied, because $\mathbb{E}\, b(\boldsymbol{X}) \in \text{rint}(\mathbb{B})$ follows from $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta}^\star)$ and because changing an edge cannot change the number of within-neighborhood edges by more than 1 and the number of within-neighborhood transitive edges by more than $2\,(\|\mathcal{A}\|_\infty - 2) + 1$. Conditions [C.2] and [C.3] follow from classic exponential-family theory [5], because $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is the natural parameter vector and $\boldsymbol{\beta}(\boldsymbol{\theta}) \in \text{rint}(\mathbb{B})$ is the mean-value parameter vector of the misspecified exponential family. Condition [C.2] is satisfied, because $\boldsymbol{\beta}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ by classic exponential-family theory [5, Theorem 2.2, pp. 34–35] and there exists a unique element $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ such that $\boldsymbol{\beta}(\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}\, b(\boldsymbol{X}) = \mathbb{E}\, b(\boldsymbol{X})$, as the map $\boldsymbol{\beta} : \text{int}(\boldsymbol{\Theta}_0) \mapsto \text{rint}(\mathbb{B})$ is one-to-one by classic exponential-family theory [e.g., Theorem 3.6, 5, p. 74]. Condition [C.3] is satisfied with $\alpha = 1$ by Lemma 6 in Appendix G.2 provided $|\mathcal{A}_k| \geq 3$ $(k = 1, \ldots, K)$. Since all conditions of Theorem 3 are satisfied, we can invoke Theorem 3 to conclude that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \boldsymbol{\Theta}_0$, there exist $\kappa(\epsilon) > 0$, $C_1 > 0$, and $C_2 > 0$ such that the M-estimator $\widehat{\boldsymbol{\theta}}$ exists, is unique, and is contained in $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$ with probability

$$\begin{aligned}
\mathbb{P}\left(\widehat{\boldsymbol{\theta}} \;\in\; \mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \;\subseteq\; \boldsymbol{\Theta}_0\right) \;&\geq\; 1 - 2\exp\left(-\frac{\kappa(\epsilon)^2\, C_1\, K}{m\, \|\mathcal{A}\|_\infty^{4\,(2-\alpha)}} + \log m \right) \\
&=\; 1 - 4\, \exp\left(-\frac{\kappa(\epsilon)^2\, C_2\, K}{\|\mathcal{A}\|_\infty^4} \right),
\end{aligned}$$

where we used the fact that $m = 2$ and $\alpha = 1$. Observe that uniqueness follows from the fact that there is a unique $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ such that $\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X}) = b(\boldsymbol{x}) \in \text{rint}(\mathbb{B})$ for all possible $b(\boldsymbol{x}) \in \text{rint}(\mathbb{B})$, because $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is the natural parameter vector and

$\boldsymbol{\beta}(\boldsymbol{\theta}) \in \text{rint}(\mathbb{B})$ is the mean-value parameter vector of the misspecified exponential family and therefore the map $\boldsymbol{\beta} : \text{int}(\boldsymbol{\Theta}_0) \mapsto \text{rint}(\mathbb{B})$ is one-to-one [e.g., Theorem 3.6, 5, p. 74].

APPENDIX F: PROOFS: CONCENTRATION RESULTS FOR SUBGRAPH-TO-GRAPH ESTIMATORS

We prove the main concentration result of Section 4, Theorem 2 along with Proposition 3 and Corollary 2.

Proof of Proposition 3. By the factorization properties implied by local dependence, we have, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^q : \ \psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}$, all $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$, and all $\mathcal{K} \subseteq \mathcal{L}$,

$$
\begin{aligned}
& \sum_{\boldsymbol{x}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}} p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{L}}) \\
= & \sum_{\boldsymbol{x}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}} \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{L}})\rangle - \psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \nu_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}}) \\
= & \sum_{\boldsymbol{x}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}} \prod_{\mathcal{A}\in\mathcal{L}} \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{A}})\rangle - \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}) \\
= & \prod_{\mathcal{A}\in\mathcal{K}} \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{A}})\rangle - \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}) \\
= & \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{K}})\rangle - \psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \nu_{\mathcal{K}}(\boldsymbol{x}_{\mathcal{K}}) \\
= & \ p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{K}}),
\end{aligned}
\tag{F.1}
$$

where we exploited the fact that, by local dependence, $\nu_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{K}})$ satisfies

$$
\nu_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}}) \ = \ \prod_{\mathcal{A}\in\mathcal{L}} \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}),
$$

while $\psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ satisfies

$$
\psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \ = \ \sum_{\mathcal{A}\in\mathcal{L}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))
$$

and

$$
\psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \ = \ \sum_{\boldsymbol{x}_{\mathcal{A}} \in \mathbb{X}_{\mathcal{A}}} \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{A}})\rangle\right) \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}), \quad \mathcal{A} \in \mathcal{L}.
$$

We note that the natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta})$ takes the form

$$
\boldsymbol{\eta}(\boldsymbol{\theta}) \ = \ (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta})),
$$

while the sufficient statistic vector takes the form

$$s(\boldsymbol{x}_{\mathcal{K}}) \;=\; (s_1(\boldsymbol{x}_{\mathcal{K}}), \ldots, s_m(\boldsymbol{x}_{\mathcal{K}})),$$

where the sufficient statistics $s_i(\boldsymbol{x}_{\mathcal{K}})$ are based on the sufficient statistics $s_{\mathcal{A},i}(\boldsymbol{x}_{\mathcal{A}})$ of neighborhoods $\mathcal{A} \in \mathcal{K}$:

$$s_i(\boldsymbol{x}_{\mathcal{K}}) \;=\; \sum_{\mathcal{A}\in\mathcal{K}} s_{\mathcal{A},i}(\boldsymbol{x}_{\mathcal{A}}), \quad i = 1, \ldots, m.$$

In addition, note that

$$\nu_{\mathcal{K}}(\boldsymbol{x}_{\mathcal{K}}) \;=\; \prod_{\mathcal{A}\in\mathcal{K}} \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$$

and that

$$\psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \;=\; \sum_{\mathcal{A}\in\mathcal{K}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$$

is finite provided $\psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \sum_{\mathcal{A}\in\mathcal{L}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is finite.

To prove that, for all $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ and all $\boldsymbol{y}_{\mathcal{K}} \in \mathbb{Y}_{\mathcal{K}}$,

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \\ =\;& \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}}), \end{aligned}$$

observe that the independence of within- and between-neighborhood subgraphs implies that

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \\ =\;& \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}})\, \mathbb{P}(\boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}). \end{aligned} \tag{F.2}$$

The first term on the right-hand side of (F.2) can be dealt with by using (F.1), which implies that

$$\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}) \;=\; \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}}).$$

The second term on the right-hand side of (F.2) can be dealt with by using the assumption that the between-neighborhood subgraphs are independent:

$$\mathbb{P}(\boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \;=\; \mathbb{P}(\boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}})\, \mathbb{P}(\boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \;=\; \mathbb{P}(\boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}}).$$

Collecting results gives

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \\ =\;& \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}})\, \mathbb{P}(\boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}}) \;=\; \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}}). \end{aligned}$$

Proof of Theorem 2. By Proposition 3, we have, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^q : \psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) < \infty\}$, all $\mathcal{K} \subseteq \mathcal{L}$, and all $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ and all $\boldsymbol{y}_{\mathcal{K}} \in \mathbb{Y}_{\mathcal{K}}$,

$$\begin{aligned} & \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}},\ \boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{Y}_{\mathcal{L}\setminus\mathcal{K}}) \\ = \ & \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{x}_{\mathcal{K}},\ \boldsymbol{Y}_{\mathcal{K}} = \boldsymbol{y}_{\mathcal{K}}) \end{aligned}$$

and, for all $\boldsymbol{x}_{\mathcal{L}} \in \mathbb{X}_{\mathcal{L}}$,

$$\sum_{\boldsymbol{x}_{\mathcal{L}\setminus\mathcal{K}} \in \mathbb{X}_{\mathcal{L}\setminus\mathcal{K}}} p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{L}}) \ = \ p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{K}}),$$

where

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta})}(\boldsymbol{x}_{\mathcal{K}}) \ = \ \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{K}})\rangle - \psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \nu_{\mathcal{K}}(\boldsymbol{x}_{\mathcal{K}}).$$

Here, $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $s(\boldsymbol{x}_{\mathcal{K}})$ have dimension $m = \max_{\mathcal{A}\in\mathcal{L}} m_{\mathcal{A}}$. We note that $\psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ satisfies

$$\psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \ = \ \sum_{\mathcal{A}\in\mathcal{K}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta})),$$

where

$$\psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \ = \ \sum_{\boldsymbol{x}_{\mathcal{A}}\in\mathbb{X}_{\mathcal{A}}} \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}),\, s(\boldsymbol{x}_{\mathcal{A}})\rangle\right)\, \nu_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}), \quad \mathcal{A} \in \mathcal{K}.$$

In addition, note that $\psi_{\mathcal{K}}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \sum_{\mathcal{A}\in\mathcal{K}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is finite provided $\psi_{\mathcal{L}}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \sum_{\mathcal{A}\in\mathcal{L}} \psi_{\mathcal{A}}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is finite. In other words, local dependence implies that $\boldsymbol{X}_{\mathcal{K}}$ is independent of $\boldsymbol{X}_{\mathcal{L}\setminus\mathcal{K}}$ and the marginal density of $\boldsymbol{x}_{\mathcal{K}} \in \mathbb{X}_{\mathcal{K}}$ induced by $\mathcal{K} \subseteq \mathcal{L}$ is an exponential-family density with support $\mathbb{X}_{\mathcal{K}}$ and local dependence, with natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta})$ and sufficient statistic vector $s(\boldsymbol{x}_{\mathcal{K}})$. As a result, Theorem 1 can be applied to the exponential family with support $\mathbb{X}_{\mathcal{K}}$ and local dependence, with natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta})$ and sufficient statistic vector $s(\boldsymbol{x}_{\mathcal{K}})$. Observe that the conditions of Theorem 2 ensure that all conditions of Theorem 1 are satisfied for all $\mathcal{K} \subseteq \mathcal{L}$. Therefore, for full exponential families, Theorem 1 shows that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exist $\kappa_1(\epsilon) > 0$ and $C_1 > 0$ such that $\widehat{\boldsymbol{\theta}}_{\mathcal{K}}$ exists, is unique, and is contained in $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ with probability

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})\right) \geq 1 - 2\,\exp\left(-\frac{\kappa_1(\epsilon)^2\, C_1\, |\mathcal{K}|}{q\, \|\mathcal{K}\|_\infty^{4\,(2-\alpha)}} + \log q\right),$$

where $\|\mathcal{K}\|_\infty = \max_{\mathcal{A}\in\mathcal{K}} |\mathcal{A}|$. For non-full, curved exponential families, Theorem 1 shows that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exist

$\kappa_2(\epsilon) > 0$ and $C_2 > 0$ such that $\widehat{\boldsymbol{\theta}}_{\mathcal{K}}$ exists and is contained in $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ with probability

$$\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})\right) \geq 1 - 2\exp\left(-\frac{\kappa_2(\epsilon)^2 C_2 |\mathcal{K}|}{m \|\mathcal{K}\|_\infty^{4(2-\alpha)}} + \log m\right).$$

Thus, there exist $\kappa(\epsilon) = \min(\kappa_1(\epsilon), \kappa_2(\epsilon)) > 0$ and $C = \min(C_1, C_2) > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{K}} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})\right) &\geq 1 - 2\exp\left(-\frac{\kappa(\epsilon)^2 C |\mathcal{K}|}{m \|\mathcal{K}\|_\infty^{4(2-\alpha)}} + \log m\right) \\ &\geq 1 - 2\exp\left(-\frac{\kappa(\epsilon)^2 C |\mathcal{K}|}{m \|\mathcal{L}\|_\infty^{4(2-\alpha)}} + \log m\right),\end{aligned}$$

where we used the fact that $\|\mathcal{K}\|_\infty = \max_{\mathcal{A}\in\mathcal{K}} |\mathcal{A}| \leq \max_{\mathcal{A}\in\mathcal{L}} |\mathcal{A}| = \|\mathcal{L}\|_\infty$ and $m = \max_{\mathcal{A}\in\mathcal{L}} m_{\mathcal{A}} = q$ in full exponential families with natural parameter vectors of the form $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

Proof of Corollary 2. Corollary 2 follows from Theorem 2, because the conditions of Theorem 2 are satisfied. The proof of Corollary 2 resembles the proof of Corollary 1 and is therefore omitted.

APPENDIX G: PROOFS: AUXILIARY LEMMAS

We prove auxiliary lemmas that are useful for verifying the conditions of Theorem 1 (Appendix G.1) and Theorem 3 (Appendix G.2). We start with two lemmas that are useful for bounding expectations of sufficient statistics. Throughout, we write

$$\Lambda(\boldsymbol{\eta}) = \min_{i\in\mathcal{A}_k < j\in\mathcal{A}_k,\, k=1,\ldots,K} \min_{\boldsymbol{x}_{-i,j}\in\mathbb{X}_{-i,j}} \mathbb{P}_{\boldsymbol{\eta}}(X_{i,j} = 1 \mid \boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j})$$

and

$$\Omega(\boldsymbol{\eta}) = \max_{i\in\mathcal{A}_k < j\in\mathcal{A}_k,\, k=1,\ldots,K} \max_{\boldsymbol{x}_{-i,j}\in\mathbb{X}_{-i,j}} \mathbb{P}_{\boldsymbol{\eta}}(X_{i,j} = 1 \mid \boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j}),$$

where $\boldsymbol{\eta}$ is the natural parameter vector of the exponential family under consideration and $\boldsymbol{x}_{-i,j} \in \mathbb{X}_{-i,j}$ corresponds to $\boldsymbol{x} \in \mathbb{X}$ excluding $x_{i,j} \in \mathbb{X}_{i,j}$ ($i \in \mathcal{A}_k < j \in \mathcal{A}_k$, $k = 1, \ldots, K$). Let $f : \mathbb{X} \mapsto \mathbb{R}$ be a function of the random graph. We denote by $\mathbb{E}_{\Lambda(\boldsymbol{\eta})} f(\boldsymbol{X})$ and $\mathbb{E}_{\Omega(\boldsymbol{\eta})} f(\boldsymbol{X})$ the expectations of $f : \mathbb{X} \mapsto \mathbb{R}$ under the Bernoulli random graph model with probabilities $\Lambda(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$, respectively, provided $\mathbb{E}_{\Lambda(\boldsymbol{\eta})} f(\boldsymbol{X})$ and $\mathbb{E}_{\Omega(\boldsymbol{\eta})} f(\boldsymbol{X})$ exist. Here, the Bernoulli random graph model

refers to the exponential-family random graph model which assumes that the edge variables $X_{i,j}$ are independent Bernoulli random variables with probabilities $\Lambda(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$, respectively.

Lemma 1. *Consider a full or non-full, canonical or curved exponential family with support* $\mathbb{X} = \{0,1\}^{\sum_{k=1}^{K}\binom{|\mathcal{A}_k|}{2}}$ *and local dependence and natural parameter vector* $\boldsymbol{\eta} \in \text{int}(\mathfrak{N})$. *Then there exists* $C(\boldsymbol{\eta}) \in [\Lambda(\boldsymbol{\eta}),\, \Omega(\boldsymbol{\eta})]$ *such that*

$$\mathbb{E}_{\boldsymbol{\eta}} \sum_{k=1}^{K} \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} X_{i,j} \;\; = \;\; C(\boldsymbol{\eta}) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2},$$

where $C(\boldsymbol{\eta}) \in [\Lambda(\boldsymbol{\eta}),\, \Omega(\boldsymbol{\eta})]$ *denotes the probability of an edge under the Bernoulli*$(C(\boldsymbol{\eta}))$ *random graph model. In addition, if* $f : \mathbb{X} \mapsto \mathbb{R}$ *is a function of the random graph that is non-decreasing in the number of edges, then*

$$\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \;\; \leq \;\; \mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X}) \;\; \leq \;\; \mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X}),$$

where $\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X})$ *and* $\mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X})$ *are the expectations of* $f : \mathbb{X} \mapsto \mathbb{R}$ *under the Bernoulli random graph model with probabilities* $\Lambda(\boldsymbol{\theta})$ *and* $\Omega(\boldsymbol{\theta})$*, respectively. If* $f : \mathbb{X} \mapsto \mathbb{R}$ *is a function of the random graph that is non-increasing in the number of edges, then the inequalities are reversed.*

Proof of Lemma 1. To ease the presentation, we consider $K = 1$ neighborhood and drop the subscript k from all neighborhood-dependent quantities. The extension to $K \geq 2$ neighborhoods is straightforward. Observe that $\mathbb{E}_{\boldsymbol{\eta}}\, X_{i,j} = \mathbb{P}_{\boldsymbol{\eta}}(X_{i,j} = 1)$ can be written as

$$\mathbb{P}_{\boldsymbol{\eta}}(X_{i,j} = 1) \;\; = \sum_{\boldsymbol{x}_{-i,j} \in \mathbb{X}_{-i,j}} \mathbb{P}_{\boldsymbol{\eta}}(X_{i,j} = 1 \,|\, \boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j})\, \mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j}),$$

which implies that $\mathbb{E}_{\boldsymbol{\eta}}\, X_{i,j}$ is bounded below by

$$\sum_{\boldsymbol{x}_{-i,j} \in \mathbb{X}_{-i,j}} \Lambda(\boldsymbol{\eta})\, \mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j}) \;\; = \;\; \Lambda(\boldsymbol{\eta})$$

and bounded above by

$$\sum_{\boldsymbol{x}_{-i,j} \in \mathbb{X}_{-i,j}} \Omega(\boldsymbol{\eta})\, \mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j}) \;\; = \;\; \Omega(\boldsymbol{\eta}).$$

Since the expectation $\mathbb{E}_{\boldsymbol{\eta}} X_{i,j}$ is contained in the convex set $[\Lambda(\boldsymbol{\eta}),\, \Omega(\boldsymbol{\eta})]$, there exists $C(\boldsymbol{\eta}) \in [\Lambda(\boldsymbol{\eta}), \Omega(\boldsymbol{\eta})]$ such that

$$\mathbb{E}_{\boldsymbol{\eta}} \sum_{i \in \mathcal{A} < j \in \mathcal{A}} X_{i,j} \;=\; \sum_{i \in \mathcal{A} < j \in \mathcal{A}} \mathbb{E}_{\boldsymbol{\eta}}\, X_{i,j} \;=\; C(\boldsymbol{\eta}) \binom{|\mathcal{A}|}{2}.$$

In addition, Butts [6, Corollary 1, p. 306] proved that, if $f : \mathbb{X} \mapsto \mathbb{R}$ is a function of the random graph that is non-decreasing in the number of edges, then

$$\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \;\leq\; \mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X}) \;\leq\; \mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X}),$$

where the inequalities are reversed when $f : \mathbb{X} \mapsto \mathbb{R}$ is a function of the random graph that is non-increasing in the number of edges. The result follows from the recursive factorization of the probability mass function $\mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{X} = \boldsymbol{x})$. Denote the elements of the sequence of within-neighborhood edge variables $\boldsymbol{X}$ by $X_1, \ldots, X_w$ and the corresponding sample spaces by $\mathbb{X}_1, \ldots, \mathbb{X}_w$, where $w = \binom{|\mathcal{A}|}{2}$. Then, for all $\boldsymbol{x} \in \mathbb{X}$, we have

$$\mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{X} = \boldsymbol{x}) \;=\; \prod_{i=1}^{w} \mathbb{P}_{\boldsymbol{\eta}}(X_i = x_i \mid X_j = x_j,\; j = 1, \ldots, i-1),$$

where the conditional probabilities $\mathbb{P}_{\boldsymbol{\eta}}(X_i = 1 \mid X_j = x_j,\; j = 1, \ldots, i-1)$ are bounded by

$$\begin{aligned}
\Lambda(\boldsymbol{\eta}) \;&=\; \min_{1 \leq i \leq w} \min_{\boldsymbol{x}_{-i} \in \mathbb{X}_{-i}} \mathbb{P}_{\boldsymbol{\eta}}(X_i = 1 \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) \\
&\leq\; \mathbb{P}_{\boldsymbol{\eta}}(X_i = 1 \mid X_j = x_j,\, j = 1, \ldots, i-1) \\
&\leq\; \max_{1 \leq i \leq w} \max_{\boldsymbol{x}_{-i} \in \mathbb{X}_{-i}} \mathbb{P}_{\boldsymbol{\eta}}(X_i = 1 \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) \;=\; \Omega(\boldsymbol{\eta}),
\end{aligned}$$

where $\boldsymbol{x}_{-i} \in \mathbb{X}_{-i}$ corresponds to $\boldsymbol{x} \in \mathbb{X}$ excluding $x_i \in \mathbb{X}_i$ $(i = 1, \ldots, w)$. Therefore, $\mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{X} = \boldsymbol{x})$ can be bounded below and above by the probability mass function of the Bernoulli random graph model with probabilities $\Lambda(\boldsymbol{\eta})$ and $\Omega(\boldsymbol{\eta})$, respectively. As a result, the cumulative distribution function and expectation of any function $f : \mathbb{X} \mapsto \mathbb{R}$ of the random graph that is non-decreasing in the number of edges can be bounded by the corresponding cumulative distribution functions and expectations of the Bernoulli random graph model with probabilities $\Lambda(\boldsymbol{\eta})$ and $\Omega(\boldsymbol{\eta})$ [6, Corollary 1, p. 306]:

$$\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \;\leq\; \mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X}) \;\leq\; \mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X}),$$

where the inequalities are reversed when $f : \mathbb{X} \mapsto \mathbb{R}$ is a function of the random graph that is non-increasing in the number of edges.

Lemma 2. *Consider a full or non-full, canonical or curved exponential family with support* $\mathbb{X} = \{0,1\}^{\sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}}$ *and local dependence and natural parameter vector* $\boldsymbol{\eta} \in \text{int}(\mathfrak{N})$. *Let* $f : \mathbb{X} \mapsto \mathbb{R}$ *be the number of within-neighborhood transitive edges defined by*

$$f(\boldsymbol{x}) \;=\; \sum_{k=1}^{K} \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} x_{i,j} \max_{h \in \mathcal{A}_k,\, h \neq i,j} x_{i,h}\, x_{j,h}, \quad \boldsymbol{x} \in \mathbb{X},$$

which may or may not be a sufficient statistic of the exponential family under consideration. If $|\mathcal{A}_k| \geq 3$ $(k = 1, \ldots, K)$, *then there exists* $C(\boldsymbol{\eta}) > 0$ *such that*

$$\mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X}) \;=\; C(\boldsymbol{\eta}) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2},$$

where $C(\boldsymbol{\eta})$ *satisfies, for some* $\lambda \in (0,1)$,

$$0 \;<\; \lambda\, \Lambda(\boldsymbol{\eta})^3 + (1-\lambda)\, \Omega(\boldsymbol{\eta})^3 \;\leq\; C(\boldsymbol{\eta}) \;\leq\; \lambda\, \Lambda(\boldsymbol{\eta}) + (1-\lambda)\, \Omega(\boldsymbol{\eta}) \;<\; 1.$$

Proof of Lemma 2. To ease the presentation, we consider $K = 1$ neighborhood and drop the subscript k from all neighborhood-dependent quantities. The extension to $K \geq 2$ neighborhoods is straightforward. Since the number of transitive edges $f : \mathbb{X} \mapsto \mathbb{R}$ is a function of the random graph that is non-decreasing in the number of edges, Lemma 1 implies that the expectation $\mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X})$ can be bounded by the expectations $\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X})$ and $\mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X})$ under the Bernoulli random graph model with edge probabilities $\Lambda(\boldsymbol{\eta})$ and $\Omega(\boldsymbol{\eta})$, respectively:

$$\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \;\leq\; \mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X}) \;\leq\; \mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X}),$$

where $\Lambda(\boldsymbol{\eta})$ and $\Omega(\boldsymbol{\eta})$ are defined in the introduction of Appendix G and all expectations exist, because the sample space $\mathbb{X}$ is finite. The expectation $\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X})$ can be written as

$$\begin{aligned}
\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \;&=\; \mathbb{E}_{\Lambda(\boldsymbol{\eta})} \left[\sum_{i \in \mathcal{A} < j \in \mathcal{A}} X_{i,j} \max_{h \in \mathcal{A},\, h \neq i,j} X_{i,h}\, X_{j,h} \right] \\
&=\; \mathbb{E}_{\Lambda(\boldsymbol{\eta})} \left[\sum_{i \in \mathcal{A} < j \in \mathcal{A}} X_{i,j}\, \mathbb{1} \left(\sum_{h \in \mathcal{A},\, h \neq i,j} X_{i,h}\, X_{j,h} \geq 1 \right) \right] \\
&=\; \sum_{i \in \mathcal{A} < j \in \mathcal{A}} \mathbb{E}_{\Lambda(\boldsymbol{\eta})} \left[X_{i,j}\, \mathbb{1} \left(\sum_{h \in \mathcal{A},\, h \neq i,j} X_{i,h}\, X_{j,h} \geq 1 \right) \right],
\end{aligned}$$

where $\mathbb{1}(\sum_{h\in\mathcal{A},\, h\neq i,j} x_{i,h}\, x_{j,h} \geq 1)$ is an indicator function, which is 1 if nodes i and j have one or more shared partners in neighborhood $\mathcal{A}$ and is 0 otherwise. Here, we exploited the fact that the number of transitive edges in neighborhood $\mathcal{A}$ is equal to the number of pairs of nodes with one or more edgewise shared partners. To evaluate the expectation $\mathbb{E}_{\Lambda(\boldsymbol{\eta})}[X_{i,j}\ \mathbb{1}(\sum_{h\in\mathcal{A},\, h\neq i,j} X_{i,h}\, X_{j,h} \geq 1)]$, we take advantage of the independence of edge variables under the Bernoulli$(\Lambda(\boldsymbol{\eta}))$ random graph model, which implies that

$$\begin{aligned}
&\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\left[X_{i,j}\ \mathbb{1}\left(\sum_{h\in\mathcal{A},\, h\neq i,j} X_{i,h}\, X_{j,h} \geq 1\right)\right] \\
&= \mathbb{E}_{\Lambda(\boldsymbol{\eta})}(X_{i,j})\ \mathbb{E}_{\Lambda(\boldsymbol{\eta})}\left[\mathbb{1}\left(\sum_{h\in\mathcal{A},\, h\neq i,j} X_{i,h}\, X_{j,h} \geq 1\right)\right] \\
&= \Lambda(\boldsymbol{\eta})\ \mathbb{P}_{\Lambda(\boldsymbol{\eta})}\left(\sum_{h\in\mathcal{A},\, h\neq i,j} X_{i,h}\, X_{j,h}\ \geq\ 1\right).
\end{aligned}$$

In addition, the independence of edge variables under the Bernoulli$(\Lambda(\boldsymbol{\eta}))$ random graph model implies that, for all $i \in \mathcal{A}$ and all $j \in \mathcal{A}$ such that $i \neq j$, the distribution of $\sum_{h\in\mathcal{A},\, h\neq i,j} X_{i,h}\, X_{j,h}$ is Binomial$(|\mathcal{A}| - 2,\, \Lambda(\boldsymbol{\eta})^2)$. Therefore,

$$\mathbb{P}_{\Lambda(\boldsymbol{\eta})}\left(\sum_{h\in\mathcal{A},\, h\neq i,j} X_{i,h}\, X_{j,h}\ \geq\ 1\right) \quad = \quad 1 - (1 - \Lambda(\boldsymbol{\eta})^2)^{|\mathcal{A}|-2},$$

which implies that

$$\mathbb{E}_{\Lambda(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \quad = \quad \Lambda(\boldsymbol{\eta})\,[1 - (1 - \Lambda(\boldsymbol{\eta})^2)^{|\mathcal{A}|-2}]\,\binom{|\mathcal{A}|}{2}.$$

Along the same lines, it can be shown that the expectation $\mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X})$ is given by

$$\mathbb{E}_{\Omega(\boldsymbol{\eta})}\, f(\boldsymbol{X}) \quad = \quad \Omega(\boldsymbol{\eta})\,[1 - (1 - \Omega(\boldsymbol{\eta})^2)^{|\mathcal{A}|-2}]\,\binom{|\mathcal{A}|}{2}.$$

Let $C_1(\boldsymbol{\eta}) = \Lambda(\boldsymbol{\eta})\,[1-(1-\Lambda(\boldsymbol{\eta})^2)^{|\mathcal{A}|-2}]$ and $C_2(\boldsymbol{\eta}) = \Omega(\boldsymbol{\eta})\,[1-(1-\Omega(\boldsymbol{\eta})^2)^{|\mathcal{A}|-2}]$. Observe that $\Lambda(\boldsymbol{\eta}) \in (0,1)$ implies that $C_1(\boldsymbol{\eta})$ is bounded below by $\Lambda(\boldsymbol{\eta})^3$ and bounded above by $\Lambda(\boldsymbol{\eta})$ for all $|\mathcal{A}| \geq 3$. Likewise, $C_2(\boldsymbol{\eta})$ is bounded below by $\Omega(\boldsymbol{\eta})^3$ and bounded above by $\Omega(\boldsymbol{\eta})$ for all $|\mathcal{A}| \geq 3$. Since the expectation $\mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X})$ is contained in the convex set $[C_1(\boldsymbol{\eta})\,\binom{|\mathcal{A}|}{2},\ C_2(\boldsymbol{\eta})\,\binom{|\mathcal{A}|}{2}]$, there exists $\lambda \in (0,1)$ such that

$$\mathbb{E}_{\boldsymbol{\eta}}\, f(\boldsymbol{X}) \quad = \quad \lambda\, C_1(\boldsymbol{\eta})\,\binom{|\mathcal{A}|}{2} + (1-\lambda)\, C_2(\boldsymbol{\eta})\,\binom{|\mathcal{A}|}{2} \quad = \quad C(\boldsymbol{\eta})\,\binom{|\mathcal{A}|}{2},$$

where $C(\boldsymbol{\eta})$ satisfies

$$0 \;<\; \lambda\,\Lambda(\boldsymbol{\eta})^3 + (1-\lambda)\,\Omega(\boldsymbol{\eta})^3 \;\;\leq\;\; C(\boldsymbol{\eta}) \;\;\leq\;\; \lambda\,\Lambda(\boldsymbol{\eta}) + (1-\lambda)\,\Omega(\boldsymbol{\eta}) \;<\; 1.$$

G.1. Auxiliary lemmas: Theorem 1. We prove auxiliary lemmas that are useful for verifying the conditions of Theorem 1.

Lemma 3. *Consider a curved exponential-family random graph with within-neighborhood edge and geometrically weighted edgewise shared partner terms as defined in Section 3.3. Let* $\boldsymbol{\Theta} = \mathbb{R} \times (0,1)$. *Then the map* $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ *is one-to-one and, for all* $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, *there exists* $\gamma(\epsilon) > 0$ *such that, for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, *we have* $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$ *provided* $|\mathcal{A}_k| \geq 4$ *(*$k = 1, \ldots, K$*).*

Proof of Lemma 3. It is straightforward to show that the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one and that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$,

$$\|\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^\star)\|_2 \;\;=\;\; \sqrt{\sum_{i=1}^{m} (\eta_i(\boldsymbol{\theta}) - \eta_i(\boldsymbol{\theta}^\star))^2} \;\;\geq\;\; \gamma(\epsilon).$$

To see that, note that $m = \|\mathcal{A}\|_\infty - 1 \geq 3$ since $|\mathcal{A}_k| \geq 4$ $(k = 1, \ldots, K)$. Therefore,

$$\|\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^\star)\|_2 \;\;=\;\; \sqrt{\sum_{i=1}^{m} (\eta_i(\boldsymbol{\theta}) - \eta_i(\boldsymbol{\theta}^\star))^2} \;\;\geq\;\; \sqrt{\sum_{i=1}^{3} (\eta_i(\boldsymbol{\theta}) - \eta_i(\boldsymbol{\theta}^\star))^2},$$

where

$$\begin{aligned} \eta_1(\boldsymbol{\theta}) &= \theta_1 \\ \eta_2(\boldsymbol{\theta}) &= 1 \\ \eta_3(\boldsymbol{\theta}) &= 2 - \theta_2. \end{aligned}$$

Thus, for all $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta}) = \mathbb{R} \times (0,1)$ and all $\boldsymbol{\delta} \in \mathbb{R}^2$ such that $\boldsymbol{\theta}^\star + \boldsymbol{\delta} \in \text{int}(\boldsymbol{\Theta}) = \mathbb{R} \times (0,1)$,

$$\begin{aligned} (\eta_1(\boldsymbol{\theta}^\star + \boldsymbol{\delta}) - \eta_1(\boldsymbol{\theta}^\star))^2 &= (\theta_1^\star + \delta_1 - \theta_1^\star)^2 \;=\; \delta_1^2 \\ (\eta_2(\boldsymbol{\theta}^\star + \boldsymbol{\delta}) - \eta_2(\boldsymbol{\theta}^\star))^2 &= 0 \\ (\eta_3(\boldsymbol{\theta}^\star + \boldsymbol{\delta}) - \eta_3(\boldsymbol{\theta}^\star))^2 &= (2 - \theta_2^\star - \delta_2 - 2 + \theta_2^\star)^2 \;=\; \delta_2^2, \end{aligned}$$

which implies that

$$\|\boldsymbol{\eta}(\boldsymbol{\theta}^\star + \boldsymbol{\delta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^\star)\|_2 \geq \sqrt{\sum_{i=1}^{3}(\eta_i(\boldsymbol{\theta}^\star + \boldsymbol{\delta}) - \eta_i(\boldsymbol{\theta}^\star))^2} = \sqrt{\delta_1^2 + \delta_2^2} = \|\boldsymbol{\delta}\|_2.$$

Therefore, the ℓ_2-distance of $\boldsymbol{\eta}(\boldsymbol{\theta}^\star + \boldsymbol{\delta})$ from $\boldsymbol{\eta}(\boldsymbol{\theta}^\star) \in \text{int}(\mathfrak{N})$ in the natural parameter space $\mathfrak{N}$ is a strictly increasing function of the ℓ_2-distance $\epsilon = \|\boldsymbol{\delta}\|_2$ of $\boldsymbol{\theta}^\star + \boldsymbol{\delta}$ from $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$ in the parameter space $\boldsymbol{\Theta}$. As a result, for all $\epsilon > 0$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$,

$$\|\boldsymbol{\eta}(\boldsymbol{\theta}^\star) - \boldsymbol{\eta}(\boldsymbol{\theta})\|_2 = \sqrt{\sum_{i=1}^{m}(\eta_i(\boldsymbol{\theta}^\star) - \eta_i(\boldsymbol{\theta}))^2} \geq \gamma(\epsilon).$$

Lemma 4. *Consider a curved exponential-family random graph with within-neighborhood edge and geometrically weighted edgewise shared partner terms as defined in Section 3.3. Let* $\boldsymbol{\Theta} = \mathbb{R} \times (0, 1)$ *and assume that* $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$. *Then, for all* $\epsilon > 0$ *small enough so that* $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, *there exists* $\gamma(\epsilon) > 0$ *such that, for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, *we have* $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$. *In addition, there exists* $\delta(\epsilon) > 0$ *such that, for all* $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$,

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \geq \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{3/4},$$

provided $|\mathcal{A}_k| \geq 4$ $(k = 1, \ldots, K)$. *Therefore, identifiability condition (3.4) of Theorem 1 is satisfied with* $\alpha = 3/4$ *provided* $|\mathcal{A}_k| \geq 4$ $(k = 1, \ldots, K)$.

Proof of Lemma 4. To ease the presentation, we consider $K = 1$ neighborhood and drop the subscript k from all neighborhood-dependent quantities. The extension to $K \geq 2$ neighborhoods is straightforward. By Lemma 3, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\gamma(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we have $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$. Therefore, it is enough to show that there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$,

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \geq \delta(\epsilon) \binom{|\mathcal{A}|}{2}^{\alpha} \quad \text{for some} \quad 0 \leq \alpha \leq 1.$$

To do so, observe that

$$\begin{aligned} &\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \\ &= \sqrt{(\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta})))^2 + \sum_{i=2}^{|\mathcal{A}|-1}(\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})))^2} > 0, \end{aligned}$$

where strict positivity follows from the fact that the maps $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ and $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ are one-to-one and hence $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star$ implies $\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 > 0$; note that the map $\boldsymbol{\eta} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{int}(\mathfrak{N})$ is one-to-one by Lemma 3 while the map $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one by classic exponential-family theory [5, Theorem 3.6, p. 74].

We distinguish two cases, the case $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \neq \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))$ and the case $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) = \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))$. We note that $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star$ and $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) = \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))$ imply $\sum_{i=2}^{|\mathcal{A}|-1}(\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})))^2 > 0$, because $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star$ implies $\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 > 0$.

Case $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \neq \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))$. We have

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \quge \quad |\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))| \quad > \quad 0.$$

Observe that $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\, s_1(\boldsymbol{X})$ is the expected number of edges in neighborhood $\mathcal{A}$. Therefore, Lemma 1 can be invoked to show that there exist $C_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) > 0$ and $C_1(\boldsymbol{\eta}(\boldsymbol{\theta})) > 0$ such that

$$|\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))| \quad = \quad |C_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - C_1(\boldsymbol{\eta}(\boldsymbol{\theta}))| \binom{|\mathcal{A}|}{2},$$

where $\Lambda(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \leq C_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) \leq \Omega(\boldsymbol{\eta}(\boldsymbol{\theta}^\star))$ and $\Lambda(\boldsymbol{\eta}(\boldsymbol{\theta})) \leq C_1(\boldsymbol{\eta}(\boldsymbol{\theta})) \leq \Omega(\boldsymbol{\eta}(\boldsymbol{\theta}))$ for all $\boldsymbol{\theta} \in \text{int}(\boldsymbol{\Theta}) = \mathbb{R} \times (0, 1)$. We note that $\Lambda(\boldsymbol{\eta}(\boldsymbol{\theta}))$ and $\Omega(\boldsymbol{\eta}(\boldsymbol{\theta}))$ are defined in the introduction of Appendix G and that $\Lambda(\boldsymbol{\eta}(\boldsymbol{\theta})) \geq [1 + \exp(-\theta_1)]^{-1} > 0$ while $\Omega(\boldsymbol{\eta}(\boldsymbol{\theta}))$ satisfies $0 < \Lambda(\boldsymbol{\eta}(\boldsymbol{\theta})) \leq \Omega(\boldsymbol{\eta}(\boldsymbol{\theta})) < 1$ for all $\boldsymbol{\theta} \in \text{int}(\boldsymbol{\Theta}) = \mathbb{R} \times (0, 1)$.

Case $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) = \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))$. As pointed out above, $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star$ and $\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) = \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}))$ imply $\sum_{i=2}^{|\mathcal{A}|-1}(\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})))^2 > 0$, where $\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\, s_i(\boldsymbol{X})$ is the expected number of pairs of nodes with $i-1$ edgewise shared partners in neighborhood $\mathcal{A}$ ($i = 2, \ldots, |\mathcal{A}| - 1$). Bounding the term $\sum_{i=2}^{|\mathcal{A}|-1}(\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})))^2$ is more challenging than bounding the term $(\mu_1(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_1(\boldsymbol{\eta}(\boldsymbol{\theta})))^2$, because the numbers of pairs of nodes with $i-1$ edgewise shared partners are neither non-decreasing nor non-increasing functions of the number of edges ($i = 2, \ldots, |\mathcal{A}| - 1$). Therefore, Lemma 1 cannot be applied to the expectations $\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\, s_i(\boldsymbol{X})$ ($i = 2, \ldots, |\mathcal{A}| - 1$). But it turns out to be possible to bound $\sum_{i=2}^{|\mathcal{A}|-1}(\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})))^2$ from below in terms of absolute deviations of the expected numbers of transitive edges under $\boldsymbol{\eta}(\boldsymbol{\theta}^\star)$ and $\boldsymbol{\eta}(\boldsymbol{\theta})$. The advantage of doing so is that the number of transitive edges is a non-decreasing function of the number of edges and hence Lemma 1 can be applied via Lemma 2. To see that the expected numbers of pairs of nodes with one or more edgewise shared partners is related to the expected number of transitive edges, note

that the expected number of transitive edges in neighborhood $\mathcal{A}$ can be written as

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})} \sum_{a\in\mathcal{A}<b\in\mathcal{A}} X_{a,b} \max_{c\in\mathcal{A},\, c\neq a,b} X_{a,c}\, X_{b,c} \\
= \;& \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})} \sum_{i=2}^{|\mathcal{A}|-1} \sum_{a\in\mathcal{A}<b\in\mathcal{A}} X_{a,b}\, \mathbb{1}\left(\sum_{c\in\mathcal{A},\, c\neq a,b} X_{a,c}\, X_{b,c} = i-1\right) \\
= \;& \sum_{i=2}^{|\mathcal{A}|-1} \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})} \sum_{a\in\mathcal{A}<b\in\mathcal{A}} X_{a,b}\, \mathbb{1}\left(\sum_{c\in\mathcal{A},\, c\neq a,b} X_{a,c}\, X_{b,c} = i-1\right) \\
= \;& \sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})),
\end{aligned}$$

which shows that $\sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is equal to the expected number of transitive edges under $\boldsymbol{\eta}(\boldsymbol{\theta})$. To bound $\sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta})))^2$ from below in terms of absolute deviations of the expected numbers of transitive edges under $\boldsymbol{\eta}(\boldsymbol{\theta}^\star)$ and $\boldsymbol{\eta}(\boldsymbol{\theta})$, it is convenient to operate in the natural parameter space $\mathfrak{N} = \mathbb{R}^{|\mathcal{A}|-1}$ of the full exponential family with natural parameter vector $\boldsymbol{\eta}$ and sufficient statistic vector $s(\boldsymbol{x})$. Write $\boldsymbol{\eta} \equiv \boldsymbol{\eta}(\boldsymbol{\theta})$ and $\boldsymbol{\eta}^\star \equiv \boldsymbol{\eta}(\boldsymbol{\theta}^\star)$. Then, to bound the term $\sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\boldsymbol{\eta}))^2$ from below in terms of absolute deviations of the expected numbers of transitive edges under $\boldsymbol{\eta}^\star$ and $\boldsymbol{\eta}$ while avoiding the trivial lower bound of 0, we note that it is possible to choose $\dot{\boldsymbol{\eta}} \in \text{int}(\mathfrak{N})$ such that

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\boldsymbol{\eta})\|_2 \;\; \geq \;\; \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_2$$

and such that the expected number of transitive edges under $\dot{\boldsymbol{\eta}}$ is not identical to the expected number of transitive edges under $\boldsymbol{\eta}^\star$. To see that, note that

$$\|\boldsymbol{\mu}(\boldsymbol{\eta})\|_1 \;\; = \;\; \sum_{i=1}^{|\mathcal{A}|-1} |\mu_i(\boldsymbol{\eta})| \;\; = \;\; \mu_1(\boldsymbol{\eta}) + \sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}),$$

where $\mu_1(\boldsymbol{\eta})$ is the expected number of edges while $\sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta})$ is the expected number of transitive edges under $\boldsymbol{\eta}$. Therefore, if $\mu_1(\boldsymbol{\eta}^\star) = \mu_1(\boldsymbol{\eta})$ and $\|\boldsymbol{\mu}(\boldsymbol{\eta})\|_1 = \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$, then the expected numbers of transitive edges under $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^\star$ are identical. Let

$$\mathbb{T} \;\; = \;\; \{\boldsymbol{\mu}' \in \text{rint}(\mathbb{M}) : \;\; \mu_1' \;=\; \mu_1(\boldsymbol{\eta}^\star), \;\; \|\boldsymbol{\mu}'\|_1 = \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1\}$$

be the set of mean-value parameter vectors for which the expected numbers of edges and the expected numbers of transitive edges are identical. To see that it

is possible to choose $\dot{\boldsymbol{\eta}} \in \text{int}(\mathfrak{N})$ such that $\mu_1(\dot{\boldsymbol{\eta}}) = \mu_1(\boldsymbol{\eta}^\star)$ and $\|\boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_1 \neq \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$, note that the set $\mathbb{T}$ is a subset of the boundary of the ℓ_1-ball with center $\mathbf{0} \in \mathbb{R}^{|\mathcal{A}|-1}$ and radius $\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$. Suppose that the ℓ_2-distance of $\boldsymbol{\mu}(\boldsymbol{\eta})$ from $\boldsymbol{\mu}(\boldsymbol{\eta}^\star)$ is equal to $\rho_1 > 0$; note that the ℓ_2-distance is strictly positive because $\boldsymbol{\eta} \neq \boldsymbol{\eta}^\star$ implies $\boldsymbol{\mu}(\boldsymbol{\eta}) \neq \boldsymbol{\mu}(\boldsymbol{\eta}^\star)$. Observe that the ℓ_2-ball $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), \rho_1)$ with center $\boldsymbol{\mu}(\boldsymbol{\eta}^\star) \in \text{rint}(\mathbb{M})$ and radius $\rho_1 > 0$ need not be contained in $\text{rint}(\mathbb{M})$, but, owing to the fact that the set $\mathbb{M}$ is convex by construction, it is possible to construct a smaller ℓ_2-ball $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), \rho_2)$ with the same center $\boldsymbol{\mu}(\boldsymbol{\eta}^\star) \in \text{rint}(\mathbb{M})$ but smaller radius $0 < \rho_2 < \rho_1$ such that the resulting ℓ_2-ball $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), \rho_2)$ is contained in $\text{rint}(\mathbb{M})$. The ℓ_1-ball with center $\mathbf{0} \in \mathbb{R}^{|\mathcal{A}|-1}$ and radius $\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$ and the ℓ_2-ball $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), \rho_2) \subseteq \text{rint}(\mathbb{M})$ with center $\boldsymbol{\mu}(\boldsymbol{\eta}^\star) \in \text{rint}(\mathbb{M})$ and radius ρ_2 intersect, but it is not too hard to see that it is possible to choose $\dot{\boldsymbol{\mu}} \in \text{rint}(\mathbb{M})$ inside the ℓ_2-ball $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), \rho_2) \subseteq \text{rint}(\mathbb{M})$ such that $\dot{\mu}_1 = \mu_1(\boldsymbol{\eta}^\star)$ and $\|\dot{\boldsymbol{\mu}}\|_1 \neq \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$. In addition, due to the fact that the map $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is a homeomorphism [5, Theorem 3.6, p. 74], there exists $\dot{\boldsymbol{\eta}} \in \text{int}(\mathfrak{N})$ such that $\boldsymbol{\mu}(\dot{\boldsymbol{\eta}}) = \dot{\boldsymbol{\mu}} \in \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), \rho_2) \subseteq \text{rint}(\mathbb{M})$, which implies that $\mu_1(\dot{\boldsymbol{\eta}}) = \mu_1(\boldsymbol{\eta}^\star)$ and $\|\boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_1 \neq \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$. In conclusion, there exists $\dot{\boldsymbol{\eta}} \in \text{int}(\mathfrak{N})$ such that $\mu_1(\dot{\boldsymbol{\eta}}) = \mu_1(\boldsymbol{\eta}^\star)$ and $\|\boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_1 \neq \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star)\|_1$ and

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\boldsymbol{\eta})\|_2 \;\geq\; \|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_2 \;=\; \sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}}))^2 \;>\; 0.$$

As a consequence, we can bound $\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_2$ from below in terms of absolute deviations of the expected numbers of transitive edges under $\boldsymbol{\eta}^\star$ and $\dot{\boldsymbol{\eta}}$ while avoiding the trivial lower bound of 0.

To do so, we first use the Cauchy-Schwarz inequality to obtain

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_2 \;=\; \sqrt{(\mu_1(\boldsymbol{\eta}^\star) - \mu_1(\dot{\boldsymbol{\eta}}))^2 + \sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}}))^2}$$

$$\geq \sqrt{\sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}}))^2} \geq \frac{1}{\sqrt{|\mathcal{A}| - 2}} \sum_{i=2}^{|\mathcal{A}|-1} |\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}}))|$$

and then use the triangle inequality to obtain

$$\sum_{i=2}^{|\mathcal{A}|-1} |\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}})| \;\geq\; \left| \sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}})) \right|$$

$$= \left| \sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}^\star) - \sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\dot{\boldsymbol{\eta}}) \right|.$$

The sums $\sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}^\star)$ and $\sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\dot{\boldsymbol{\eta}})$ are equal to the expected number of transitive edges under $\boldsymbol{\eta}^\star$ and $\dot{\boldsymbol{\eta}}$, respectively, as shown above. Lemma 2 can hence be applied to show that, for all $\boldsymbol{\eta} \in \text{int}(\mathfrak{N})$, there exists $C_2(\boldsymbol{\eta}) > 0$ such that

$$\sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}) \;=\; \mathbb{E}_{\boldsymbol{\eta}} \sum_{a \in \mathcal{A} < b \in \mathcal{A}} X_{a,b} \max_{c \in \mathcal{A},\, c \neq a,b} X_{a,c}\, X_{b,c} \;=\; C_2(\boldsymbol{\eta}) \binom{|\mathcal{A}|}{2},$$

where $C_2(\boldsymbol{\eta})$ satisfies

$$\lambda\, \Lambda(\boldsymbol{\eta})^3 + (1-\lambda)\, \Omega(\boldsymbol{\eta})^3 \;\leq\; C_2(\boldsymbol{\theta}) \;\leq\; \lambda\, \Lambda(\boldsymbol{\eta}) + (1-\lambda)\, \Omega(\boldsymbol{\eta}),$$

provided $|\mathcal{A}| \geq 3$. Collecting results shows that

$$\begin{aligned}
\|\boldsymbol{\mu}(\boldsymbol{\eta}^\star) - \boldsymbol{\mu}(\dot{\boldsymbol{\eta}})\|_2 \;&=\; \sqrt{(\mu_1(\boldsymbol{\eta}^\star) - \mu_1(\dot{\boldsymbol{\eta}}))^2 + \sum_{i=2}^{|\mathcal{A}|-1} (\mu_i(\boldsymbol{\eta}^\star) - \mu_i(\dot{\boldsymbol{\eta}}))^2} \\
&\geq\; \frac{1}{\sqrt{|\mathcal{A}|-2}} \left| \sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \sum_{i=2}^{|\mathcal{A}|-1} \mu_i(\boldsymbol{\eta}(\dot{\boldsymbol{\theta}})) \right| \\
&=\; \frac{1}{\sqrt{|\mathcal{A}|-2}}\; |C_2(\boldsymbol{\eta}^\star) - C_2(\dot{\boldsymbol{\eta}})|\; \binom{|\mathcal{A}|}{2} \\
&\geq\; \frac{1}{|\mathcal{A}|^{1/4}\,(|\mathcal{A}|-1)^{1/4}}\; |C_2(\boldsymbol{\eta}^\star) - C_2(\dot{\boldsymbol{\eta}})|\; \binom{|\mathcal{A}|}{2} \\
&=\; \frac{1}{2^{1/4}}\; |C_2(\boldsymbol{\eta}^\star) - C_2(\dot{\boldsymbol{\eta}})|\; \binom{|\mathcal{A}|}{2}^{3/4},
\end{aligned}$$

where $|C_2(\boldsymbol{\eta}^\star) - C_2(\dot{\boldsymbol{\eta}})| > 0$ by the choice of $\dot{\boldsymbol{\eta}}$, as explained above.

We can therefore conclude that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, there exists a function $C(\boldsymbol{\theta}^\star, \boldsymbol{\theta}) > 0$ such that

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \;\;\geq\;\; C(\boldsymbol{\theta}^\star, \boldsymbol{\theta}) \binom{|\mathcal{A}|}{2}^{3/4}.$$

Last, but not least, we show that there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star), \gamma(\epsilon))$, we have $\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \geq \delta(\epsilon) \binom{|\mathcal{A}|}{2}^{3/4} > 0$. To do so, it is convenient to operate in the natural parameter space $\mathfrak{N} = \mathbb{R}^{|\mathcal{A}|-1}$ of the full exponential family with natural parameter vector $\boldsymbol{\eta}$ and sufficient statistic vector $s(\boldsymbol{x})$. Write $\boldsymbol{\eta} \equiv \boldsymbol{\eta}(\boldsymbol{\theta})$ and $\boldsymbol{\eta}^\star \equiv \boldsymbol{\eta}(\boldsymbol{\theta}^\star)$. By classic exponential family theory, the map $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is a homeomorphism, i.e., $\boldsymbol{\mu} : \text{int}(\mathfrak{N}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one and continuous, and so is its inverse $\boldsymbol{\mu}^{-1} : \text{rint}(\mathbb{M}) \mapsto \text{int}(\mathfrak{N})$

[5, Theorem 3.6, p. 74]. By the continuity of $\boldsymbol{\mu}^{-1} : \text{rint}(\mathbb{M}) \mapsto \text{int}(\mathfrak{N})$, we know that, for each $\epsilon > 0$ and each $\gamma(\epsilon) > 0$, there exists $a(\epsilon) > 0$ such that $\boldsymbol{\mu}^{-1}(\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), a(\epsilon))) \subseteq \mathcal{B}(\boldsymbol{\eta}^\star, \gamma(\epsilon))$. In other words, all elements of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), a(\epsilon))$ map to elements of $\mathcal{B}(\boldsymbol{\eta}^\star, \gamma(\epsilon))$ and thus no element of $\mathfrak{N} \setminus \mathcal{B}(\boldsymbol{\eta}^\star, \gamma(\epsilon))$ can map to an element of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\eta}^\star), a(\epsilon))$. In addition, we have shown above that $\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \geq C(\boldsymbol{\theta}^\star, \boldsymbol{\theta}) \left(\binom{|\mathcal{A}|}{2}\right)^{3/4}$. Combining these results shows that there exists $\delta(\epsilon) > 0$ such that $a(\epsilon)$ can be written as $a(\epsilon) = \delta(\epsilon) \left(\binom{|\mathcal{A}|}{2}\right)^{3/4}$. In summary, no element of $\boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$ can map to an element of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\theta}^\star), \delta(\epsilon) \left(\binom{|\mathcal{A}|}{2}\right)^{3/4})$.

Taken together, these results show that there exists $\epsilon > 0$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$,

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \quad \geq \quad \delta(\epsilon) \left(\frac{|\mathcal{A}|}{2}\right)^{3/4},$$

provided $|\mathcal{A}| \geq 4$. Therefore, identifiability condition (3.4) of Theorem 1 is satisfied with $\alpha = 3/4$ provided $|\mathcal{A}| \geq 4$.

Lemma 5. *Consider an exponential-family random graph with within-neighborhood edge and transitive edge terms as defined in Appendix A. Let $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$ and assume that $\boldsymbol{\theta}^\star \in \text{int}(\boldsymbol{\Theta})$. Then, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$,*

$$\|\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}^\star)) - \boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta}))\|_2 \quad \geq \quad \delta(\epsilon) \sum_{k=1}^{K} \left(\frac{|\mathcal{A}_k|}{2}\right),$$

provided $|\mathcal{A}_k| \geq 3$ ($k = 1, \ldots, K$). Therefore, identifiability condition (3.4) of Theorem 1 is satisfied with $\alpha = 1$ provided $|\mathcal{A}_k| \geq 3$ ($k = 1, \ldots, K$).

Proof of Lemma 5. To ease the presentation, we consider $K = 1$ neighborhood and drop the subscript k from all neighborhood-dependent quantities. The extension to $K \geq 2$ neighborhoods is straightforward. In the following, we write $\boldsymbol{\mu}(\boldsymbol{\eta}(\boldsymbol{\theta})) = \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\, s(\boldsymbol{X})$ and $\psi(\boldsymbol{\eta}(\boldsymbol{\theta})) = \psi(\boldsymbol{\theta})$, because the natural parameter vector of the exponential family is given by $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

To verify identifiability condition (3.4) of Theorem 1, we need to verify that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$,

$$\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 \quad \geq \quad \delta(\epsilon) \left(\frac{|\mathcal{A}|}{2}\right)^{\alpha} \quad \text{for some} \quad 0 \leq \alpha \leq 1.$$

To do so, observe that

$$\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 \;=\; \sqrt{(\mu_1(\boldsymbol{\theta}^\star) - \mu_1(\boldsymbol{\theta}))^2 + (\mu_2(\boldsymbol{\theta}^\star) - \mu_2(\boldsymbol{\theta}))^2}.$$

The deviation $|\mu_1(\boldsymbol{\theta}^\star) - \mu_1(\boldsymbol{\theta})|$ can be dealt with by using Lemma 1, which shows that there exists $C_1(\boldsymbol{\theta}) \in [\Lambda(\boldsymbol{\theta}),\, \Omega(\boldsymbol{\theta})]$ such that

$$\mu_1(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\boldsymbol{\theta}} \sum_{i \in \mathcal{A} < j \in \mathcal{A}} X_{i,j} \;=\; C_1(\boldsymbol{\theta}) \binom{|\mathcal{A}|}{2},$$

where $\Lambda(\boldsymbol{\theta})$ and $\Omega(\boldsymbol{\theta})$ are defined in the introduction of Appendix G. Here, $\Lambda(\boldsymbol{\theta})$ is given by $\Lambda(\boldsymbol{\theta}) = [1+\exp(-\theta_1)]^{-1} > 0$, while $\Omega(\boldsymbol{\theta})$ satisfies $0 < \Lambda(\boldsymbol{\theta}) \leq \Omega(\boldsymbol{\theta}) < 1$ for all $\boldsymbol{\theta} \in \mathbb{R} \times \mathbb{R}^+$. Therefore, the deviation $|\mu_1(\boldsymbol{\theta}^\star) - \mu_1(\boldsymbol{\theta})|$ satisfies

$$|\mu_1(\boldsymbol{\theta}^\star) - \mu_1(\boldsymbol{\theta})| \;=\; |C_1(\boldsymbol{\theta}^\star) - C_1(\boldsymbol{\theta})| \binom{|\mathcal{A}|}{2}.$$

To deal with the deviation $|\mu_2(\boldsymbol{\theta}^\star) - \mu_2(\boldsymbol{\theta})|$, observe that, by Lemma 2, there exists $C_2(\boldsymbol{\theta}) > 0$ such that

$$\mu_2(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\boldsymbol{\theta}} \sum_{i \in \mathcal{A} < j \in \mathcal{A}} X_{i,j} \max_{h \in \mathcal{A},\, h \neq i,j} X_{i,h}\, X_{j,h} \;=\; C_2(\boldsymbol{\theta}) \binom{|\mathcal{A}|}{2},$$

where $C_2(\boldsymbol{\theta})$ satisfies

$$\lambda\, \Lambda(\boldsymbol{\theta})^3 + (1-\lambda)\, \Omega(\boldsymbol{\theta})^3 \;\leq\; C_2(\boldsymbol{\theta}) \;\leq\; \lambda\, \Lambda(\boldsymbol{\theta}) + (1-\lambda)\, \Omega(\boldsymbol{\theta}), \quad 0 < \lambda < 1.$$

As a result,

$$|\mu_2(\boldsymbol{\theta}^\star) - \mu_2(\boldsymbol{\theta})| \;=\; |C_2(\boldsymbol{\theta}^\star) - C_2(\boldsymbol{\theta})| \binom{|\mathcal{A}|}{2}.$$

We can hence conclude that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$, there exists a function $C(\boldsymbol{\theta}, \boldsymbol{\theta}^\star) > 0$ such that

$$\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 \;=\; C(\boldsymbol{\theta}^\star, \boldsymbol{\theta}) \binom{|\mathcal{A}|}{2} \;>\; 0,$$

where strict positivity follows from the fact that the map $\boldsymbol{\mu} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one by classic exponential-family theory [5, Theorem 3.6, p. 74], which implies that $\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 > 0$.

Last, but not least, we show that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)$, we have $\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 \geq \delta(\epsilon) \binom{|\mathcal{A}|}{2} > 0$. To do so, note that $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ is the natural parameter

vector of the exponential family. By classic exponential family theory, the map $\boldsymbol{\mu} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{rint}(\mathbb{M})$ is a homeomorphism, i.e., $\boldsymbol{\mu} : \text{int}(\boldsymbol{\Theta}) \mapsto \text{rint}(\mathbb{M})$ is one-to-one and continuous, and so is its inverse $\boldsymbol{\mu}^{-1} : \text{rint}(\mathbb{M}) \mapsto \text{int}(\boldsymbol{\Theta})$ [5, Theorem 3.6, p. 74]. By the continuity of $\boldsymbol{\mu}^{-1} : \text{rint}(\mathbb{M}) \mapsto \text{int}(\boldsymbol{\Theta})$, we know that, for each $\epsilon > 0$, there exists $a(\epsilon) > 0$ such that $\boldsymbol{\mu}^{-1}(\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\theta}^\star),\, a(\epsilon))) \subseteq \mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$. Thus, all elements of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\theta}^\star),\, a(\epsilon))$ map to elements of $\mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$ and thus no element of $\boldsymbol{\Theta} \backslash \mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$ can map to an element of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\theta}^\star),\, a(\epsilon))$. In addition, we have shown above that $\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 = C(\boldsymbol{\theta}^\star, \boldsymbol{\theta}) \binom{|\mathcal{A}|}{2}$. Combining these results shows that there exists $\delta(\epsilon) > 0$ such that $a(\epsilon)$ can be written as $a(\epsilon) = \delta(\epsilon) \binom{|\mathcal{A}|}{2}$. In summary, no element of $\boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$ can map to an element of $\mathcal{B}(\boldsymbol{\mu}(\boldsymbol{\theta}^\star),\, \delta(\epsilon) \binom{|\mathcal{A}|}{2})$.

As a result, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}^\star, \epsilon) \subseteq \text{int}(\boldsymbol{\Theta})$, there exists $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$,

$$\|\boldsymbol{\mu}(\boldsymbol{\theta}^\star) - \boldsymbol{\mu}(\boldsymbol{\theta})\|_2 \quad \geq \quad \delta(\epsilon) \binom{|\mathcal{A}|}{2},$$

provided $|\mathcal{A}| \geq 3$. Therefore, identifiability condition (3.4) of Theorem 1 is satisfied with $\alpha = 1$ provided $|\mathcal{A}| \geq 3$.

G.2. Auxiliary lemmas: Theorem 3. We prove an auxiliary lemma that is useful for verifying the conditions of Theorem 3.

Lemma 6. *Consider an exponential-family random graph with within-neighborhood edge, transitive edge, and same-attribute edge terms as defined in Section B.2. Then identifiability condition [C.3] of Theorem 3 is satisfied with* $\alpha = 1$ *provided* $|\mathcal{A}_k| \geq 3$ *(*$k = 1, \ldots, K$*).*

Proof of Lemma 6. To verify identifiability condition [C.3] of Theorem 3, we need to verify that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \boldsymbol{\Theta}_0$, there exist $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \setminus \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$,

$$\|\mathbb{E}_{\boldsymbol{\theta}^\star}\, b(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})\|_2 \quad \geq \quad \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2}^{\alpha} \quad \text{for some} \;\; 0 \leq \alpha \leq 1,$$

where $\mathbb{E}_{\boldsymbol{\theta}^\star}\, b(\boldsymbol{X})$ is the expectation of $b(\boldsymbol{X})$ under the data-generating exponential-family distribution with natural parameter vector $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta}^\star$ and $\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ is the expectation of $b(\boldsymbol{X})$ under the misspecified exponential-family distribution with natural parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$.

Since the expectations $\mathbb{E}_{\boldsymbol{\theta}^\star}\, b(\boldsymbol{X})$ and $\mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})$ are expectations of the number of within-neighborhood edges and transitive edges, an argument along the lines of

Lemma 5 shows that, for all $\epsilon > 0$ small enough so that $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon) \subseteq \boldsymbol{\Theta}_0$, there exist $\delta(\epsilon) > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \setminus \mathcal{B}(\boldsymbol{\theta}^\star,\, \epsilon)$,

$$\|\mathbb{E}_{\boldsymbol{\theta}^\star}\, b(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}}\, b(\boldsymbol{X})\|_2 \quad \geq \quad \delta(\epsilon) \sum_{k=1}^{K} \binom{|\mathcal{A}_k|}{2},$$

provided $|\mathcal{A}_k| \geq 3$ $(k = 1, \ldots, K)$. Therefore, identifiability condition [C.3] of Theorem 3 is satisfied with $\alpha = 1$ provided $|\mathcal{A}_k| \geq 3$ $(k = 1, \ldots, K)$.

REFERENCES

[1] Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: John Wiley & Sons.

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

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

[4] Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press.

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

[6] Butts, C. T. (2011), "Bernoulli graph bounds for general random graph models," *Sociological Methodology*, 41, 299–345.

[7] Butts, C. T., and Almquist, Z. W. (2015), "A Flexible Parameterization for Baseline Mean Degree in Multiple-Network ERGMs," *Journal of Mathematical Sociology*, 39, 163–167.

[8] Chatterjee, S. (2005), "Concentration Inequalities with Exchangeable Pairs," Ph.D. thesis, Department of Statistics, Stanford University.

[9] — (2007), "Estimation in spin glasses: A first step," *The Annals of Statistics*, 35, 1931–1946.

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

[11] Crane, H., and Dempsey, W. (2015), "A framework for statistical network modeling," Available at https://arxiv.org/abs/1509.08185.v4.

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

[13] Efron, B. (1975), "Defining the curvature of a statistical problem (with applications to second order efficiency)," *The Annals of Statistics*, 3, 1189–1242.

[14] — (1978), "The geometry of exponential families," *The Annals of Statistics*, 6, 362–376.

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

[16] Geyer, C. J. (2009), "Likelihood inference in exponential families and directions of recession," *Electronic Journal of Statistics*, 3, 259–289.

[17] Godambe, V. P. (1991), *Estimating Functions*, Oxford: Oxford University Press.

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

[19] Handcock, M. S., and Gile, K. (2010), "Modeling social networks from sampled data," *The Annals of Applied Statistics*, 4, 5–25.

[20] Harris, J. K. (2013), *An Introduction to Exponential Random Graph Modeling*, Thousand Oaks, California: Sage.

[21] Holland, P. W., and Leinhardt, S. (1976), "Local structure in social networks," *Sociological Methodology*, 1–45.

[22] Hollway, J., Lomi, A., Pallotti, F., and Stadtfeld, C. (2017), "Multilevel social spaces: The network dynamics of organizational fields," *Network Science*, 5, 187–212.

[23] Hunter, D. R. (2007), "Curved exponential family models for social networks," *Social Networks*, 29, 216–230.

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

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

[26] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), "Computational statistical methods for social network models," *Journal of Computational and Graphical Statistics*, 21, 856–882.

[27] Janson, S., and Rucinski, A. (2002), "The infamous upper tail," *Random Structures and Algorithms*, 20, 317–342.

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

[29] Kass, R., and Vos, P. (1997), *Geometrical foundations of asymptotic inference*, New York: Wiley.

[30] Kim, J. H., and Vu, V. H. (2004), "Divide and conquer martingales and the number of triangles in a random graph," *Random Structures & Algorithms*, 24, 166–174.

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

[32] Kontorovich, L., and Ramanan, K. (2008), "Concentration inequalities for dependent random variables via the martingale method," *The Annals of Probability*, 36, 2126–2158.

[33] Krivitsky, P. N. (2012), "Exponential-family models for valued networks," *Electronic Journal of Statistics*, 6, 1100–1128.

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

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

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

[37] Lazega, E., and Snijders, T. A. B. (eds.) (2016), *Multilevel Network Analysis for the Social Sciences*, Switzerland: Springer-Verlag.

[38] Lomi, A., Robins, G., and Tranmer, M. (2016), "Introduction to multilevel social networks," *Social Networks*, 266–268.

[39] Lusher, D., Koskinen, J., and Robins, G. (2013), *Exponential Random Graph Models for Social Networks*, Cambridge, UK: Cambridge University Press.

[40] Mukherjee, S. (2013), "Consistent estimation in the two star exponential random graph model," Tech. rep., Department of Statistics, Columbia University, arXiv:1310.4526v1.

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

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

[43] Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), "On the geometry of discrete exponential families with application to exponential random graph models," *Electronic Journal of Statistics*,

3, 446–484.
[44] Rinaldo, A., Petrovic, S., and Fienberg, S. E. (2013), "Maximum likelihood estimation in network models," *The Annals of Statistics*, 41, 1085–1110.
[45] Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.
[46] Samson, P. M. (2000), "Concentration of measure inequalities for Markov chains and Φ-mixing processes," *The Annals of Probability*, 28, 416–461.
[47] Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.
[48] Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society, Series B*, 77, 647–676.
[49] Schweinberger, M., and Luna, P. (2018), "HERGM: Hierarchical exponential-family random graph models," *Journal of Statistical Software*, 85, 1–39.
[50] Schweinberger, M., and Stewart, J. (2018), "Supplement: Finite-graph concentration and consistency results for canonical and curved exponential-family models of random graphs," Tech. rep., Department of Statistics, Rice University.
[51] Shalizi, C. R., and Rinaldo, A. (2013), "Consistency under sampling of exponential random graph models," *The Annals of Statistics*, 41, 508–535.
[52] Slaughter, A. J., and Koehly, L. M. (2016), "Multilevel models for social networks: hierarchical Bayesian approaches to exponential random graph modeling," *Social Networks*, 44, 334–345.
[53] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153.
[54] Stewart, J., Schweinberger, M., Bojanowski, M., and Morris, M. (2018), "Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms," *Social Networks*, to appear.
[55] Vu, V. H. (2002), "Concentration of non-Lipschitz functions and applications," *Random Structures & Algorithms*, 20, 262–316.
[56] Wang, P., Robins, G., Pattison, P., and Lazega, E. (2013), "Exponential random graph models for multilevel networks," *Social Networks*, 35, 96–115.
[57] Wasserman, S., and Pattison, P. (1996), "Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^*," *Psychometrika*, 61, 401–425.
[58] Xiang, R., and Neville, J. (2011), "Relational learning with one network: an asymptotic analysis," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1–10.
[59] Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2018), "Statistical inference in a directed network model with covariates," *Journal of the American Statistical Association*, 1–33, to appear.
[60] Yan, T., Leng, C., and Zhu, J. (2016), "Asymptotics in directed exponential random graph models with an increasing bi-degree sequence," *The Annals of Statistics*, 44, 31–57.
[61] Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015), "Graphical models via univariate exponential family distributions," *Journal of Machine Learning Research*, 16, 3813–3847.
[62] Zappa, P., and Lomi, A. (2015), "The analysis of multilevel networks in organizations: models and empirical tests," *Organizational Research Methods*, 18, 542–569.

MICHAEL SCHWEINBERGER, JONATHAN STEWART
DEPARTMENT OF STATISTICS
RICE UNIVERSITY
6100 MAIN ST, MS-138
HOUSTON, TX 77005-1827
E-MAIL: M.S@RICE.EDU
JONATHAN.STEWART@RICE.EDU