Statistical Science
2020, Vol. 35, No. 4, 627–662
https://doi.org/10.1214/19-STS743

Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios

Michael Schweinberger, Pavel N. Krivitsky, Carter T. Butts and Jonathan R. Stewart

Abstract. Exponential-family Random Graph Models (ERGMs) constitute a large statistical framework for modeling dense and sparse random graphs with short- or long-tailed degree distributions, covariate effects and a wide range of complex dependencies. Special cases of ERGMs include network equivalents of generalized linear models (GLMs), Bernoulli random graphs, β-models, p_1-models and models related to Markov random fields in spatial statistics and image processing. While ERGMs are widely used in practice, questions have been raised about their theoretical properties. These include concerns that some ERGMs are near-degenerate and that many ERGMs are non-projective. To address such questions, careful attention must be paid to model specifications and their underlying assumptions, and to the inferential settings in which models are employed. As we discuss, near-degeneracy can affect simplistic ERGMs lacking structure, but well-posed ERGMs with additional structure can be well-behaved. Likewise, lack of projectivity can affect non-likelihood-based inference, but likelihood-based inference does not require projectivity. Here, we review well-posed ERGMs along with likelihood-based inference. We first clarify the core statistical notions of "sample" and "population" in the ERGM framework, separating the process that generates the population graph from the observation process. We then review likelihood-based inference in finite, super and infinite population scenarios. We conclude with consistency results, and an application to human brain networks.

Key words and phrases: Social network, exponential-family random graph model, ERGM, model degeneracy, projectivity.

CONTENTS

Michael Schweinberger is Assistant Professor, Department of Statistics, Rice University, Houston, Texas 77005, USA (e-mail: m.s@rice.edu). Pavel N. Krivitsky is Lecturer, Department of Statistics, University of New South Wales, Sydney, New South Wales 2052, Australia (e-mail: p.krivitsky@unsw.edu.au). Carter T. Butts is Professor, Departments of Sociology, Statistics, Computer Science and Electrical Engineering and Computer Science, University of California, Irvine, California 92697, USA (e-mail: buttsc@uci.edu). Jonathan R. Stewart is graduate student, Department of Statistics, Rice University, Houston, Texas 77005, USA (e-mail: jrs6@rice.edu).

1. INTRODUCTION

The statistical analysis of network data is an emerging area of statistics with applications in epidemiology (e.g., disease-transmission networks), neuroscience (e.g., human brain networks), political science (e.g., insurgencies, terrorist networks), economics (e.g., financial markets), sociology (e.g., social networks) and computer science (e.g., Facebook, Twitter) (Kolaczyk, 2009). The key to the statistical analysis of network data is the development of random graph models. There is a large and growing body of literature on random graph models (Goldenberg et al., 2009, Fienberg, 2012, Hunter, Krivitsky and Schweinberger, 2012, Salter-Townshend et al., 2012), including exchangeable random graph models (e.g., Diaconis and Janson, 2008, Bickel and Chen, 2009, Cai, Campbell and Broderick, 2016, Caron and Fox, 2017, Janson, 2018, Crane and Dempsey 2018, 2020, Crane, 2018); stochastic block models (e.g., Nowicki and Snijders, 2001, Bickel and Chen, 2009, Rohe, Chatterjee and Yu, 2011); latent space models (e.g., Hoff, Raftery and Handcock, 2002, Schweinberger and Snijders, 2003, Handcock, Raftery and Tantrum, 2007, Sewell and Chen, 2015); and many other random graph models (e.g., Hoff 2005, 2020, Rastelli, Friel and Raftery, 2016, Fosdick et al., 2019). Most such models seek to address the twin challenges of modeling complex and dependent network data: the presence of *heterogeneity* (different subgraphs may have different properties) and *dependence* (the presence or absence of an edge may be dependent upon the presence or absence of other edges in the graph). The latter, dependence, has proven to be the more challenging of the two. Indeed, network data can exhibit a wide range of complex dependencies, the most famous of which is triadic closure or transitivity (e.g., Wasserman and Faust, 1994), which has been found in human brain networks (e.g., Simpson, Moussa and Laurienti, 2012, Sinke et al., 2016, Obando and De Vico Fallani, 2017), social networks (Holland and Leinhardt 1970, 1972, 1976), and other network data. Other examples include degree heterogeneity (Hunter, 2007) and suppression of short chordless cycles (Bearman, Moody and Stovel, 2004). We focus here on models designed to capture a wide range of complex dependencies, including—but not limited to—transitivity.

A broad statistical framework that has turned out to be useful for testing and modeling complex dependencies in areas as diverse as physics, neuroscience, artificial intelligence, spatial statistics and other areas of statistics are exponential-family models for dependent random variables (Barndorff-Nielsen, 1978, Brown, 1986). Such models have long been used in statistics and related areas, classic examples being Ising models in physics (Ising, 1925) and discrete Markov random fields in spatial statistics (Besag, 1974, Cressie, 1993, Stein, 1999) and machine learning (e.g., Ravikumar, Wainwright and Lafferty, 2010, Yang et al., 2015). In the network science literature, exponential-family models for dependent random variables are known as Exponential-family Random Graph Models (ERGMs) (Frank and Strauss, 1986, Wasserman and Pattison, 1996, Lusher, Koskinen and Robins, 2013, Harris, 2013). These models are versatile and, when properly specified, are capable of modeling dense and sparse random graphs with short- or long-tailed degree distributions, covariate effects and a wide range of complex dependencies (e.g., Lusher, Koskinen and Robins, 2013, Harris, 2013). Some notable special cases of ERGMs include:

- network equivalents of GLMs (McCullagh and Nelder, 1983), including network logistic regression and other forms of network regression (Krackhardt, 1988);
- Bernoulli random graphs (Gilbert, 1959, Erdős and Rényi 1959, 1960);
- categorical data models (e.g., U|MAN models, Wasserman and Faust, 1994);
- β-models (Chatterjee, Diaconis and Sly, 2011);
- p_1-models (Holland and Leinhardt, 1981);
- canonical exponential-family models with Markov dependence (Frank and Strauss, 1986), which are related to Ising models in physics (Ising, 1925), Markov logic networks in artificial intelligence (Richardson and Domingos, 2006), Markov random fields in spatial statistics (Besag, 1974, Cressie, 1993, Stein, 1999), and undirected graphical models (e.g., Lauritzen, 1996, Ravikumar, Wainwright and Lafferty, 2010);
- curved exponential-family models (Snijders et al., 2006, Hunter and Handcock, 2006, Hunter, Goodreau and Handcock, 2008), which are curved exponential families in the sense of Efron (1975, 1978).

The advantages of exponential-family representations of random graph models are many. Some of the more important ones are:

1. *Language.* Exponential families provide a convenient language for formulating ideas about network data and dependencies therein. In so doing, the language of exponential families facilitates the construction of models for complex and dependent network data.

2. *Unifying statistical framework.* The exponential-family framework is a unifying statistical framework that includes a wide range of random graph models. The unifying exponential-family framework has computational, theoretical and practical advantages.

3. *Computational advantages.* Exponential families have useful convexity properties: for example, the natural parameter space of exponential families is a convex set and the negative loglikelihood function is a strictly convex function on the interior of the natural parameter space (Brown, 1986). These convexity properties imply that maximum likelihood estimation of natural parameters is a convex minimization program with a unique solution, provided there exists a solution (Handcock, 2003, Rinaldo, Fienberg and Zhou, 2009). In addition, exponential families admit data reduction by sufficiency, and exponential-family likelihood functions depend on the data only through minimal sufficient statistics (Fisher, 1934). As a consequence, likelihood-based estimation algorithms are agnostic to the structure of the sample space of network data—which can be a large and discrete set—as long as observed and expected minimal sufficient statistics can be computed exactly or approximately (Krivitsky and Butts, 2017). Thus, the unifying statistical framework helps design statistical algorithms that can estimate a wide range of random graph models with exponential-family representations, and helps implement them in user-friendly statistical software. In fact, there are 20 ERGM-related `R` packages,[1] with tens of thousands of downloads a month among them. It is worth noting that likelihood-based methods require approximations of the likelihood function when it is intractable (a problem shared with discrete Markov random fields in spatial statistics and machine learning), but there are multiple general-purpose methods for approximate likelihood-based inference (see Section 1.2.1).

4. *Theoretical advantages.* Among the theoretical advantages is the fact that the exponential-family framework helps construct statistical theory for a wide range of random graph models that have exponential-family representations. The seminal work of Berk (1972) and Portnoy (1988) and others demonstrates that statistical theory is easier for exponential families than non-exponential families, and statistical theory for dependent network data is likewise facilitated by exponential families (as demonstrated in Section 3.1 of Schweinberger and Stewart, 2020). Network data are complex enough, and it therefore makes sense to keep statistical theory as simple as possible. The exponential-family framework enables theoreticians to do so.

[1]The 20 ERGM-related `R` packages found at `CRAN.R-project.org`: `Bergm`, `blkergm`, `btergm`, `dnr`, `EpiModel`, `ergm`, `ergm.count`, `ergm.ego`, `ergm.graphlets`, `ergm.rank`, `ergm.userterms`, `ergmharris`, `fergm`, `GERGM`, `gwdegree`, `hergm`, `mlergm`, `statnetWeb`, `tergm` and `xergm`.

5. *Practical advantages.* Among the practical advantages is the fact that ERGMs include GLMs for independent network data—for example, logistic regression models for network data—as special cases and may be viewed as GLMs for independent as well as dependent network data. Many network scientists are familiar with logistic regression models, which facilitates the interpretation of results, in that estimates of parameters may be interpreted in terms of conditional log odds and log odds ratios (e.g., Hunter, 2007, Lusher, Koskinen and Robins, 2013, Stewart et al., 2019).

In addition, exponential-family models can be used as building blocks to construct more complex models, for example, by using them as building blocks in stochastic block models to capture transitivity and other complex dependencies within blocks (Schweinberger and Handcock, 2015).

In contrast to latent space models (e.g., Hoff, Raftery and Handcock, 2002) and other latent variable models, ERGMs specify dependence by incorporating network features as sufficient statistics, providing a direct link between the structural properties on which network science theories are based (Wasserman and Faust, 1994) and the resulting models. By contrast, most latent variable models represent dependence among edges as the result of underlying heterogeneity in edge probabilities (represented by latent structure), in some cases making interpretation less direct. While both approaches have useful applications and can be combined, the ability to directly and parsimoniously model dependencies of scientific interest helps explain the popularity of Ising models in physics and ERGMs in network science. Indeed, well-posed ERGMs have found widespread application, ranging from the study of the human brain (e.g., Simpson, Moussa and Laurienti, 2012, Sinke et al., 2016, Obando and De Vico Fallani, 2017) and epidemics (e.g., Groendyke, Welch and Hunter, 2012) to differential privacy (e.g., Karwa, Krivitsky and Slavković, 2017), transient structure in intrinsically disordered proteins (Grazioli, Martin and Butts, 2019) and networks of radical environmentalists (Almquist and Bagozzi, 2019).

While ERGMs are widely used in practice, questions have been raised about their theoretical properties. To a large extent, these questions reflect a lack of clarity about the construction of ERGMs and their underlying assumptions, and the inferential settings in which ERGMs are employed. ERGMs may be used, among other things, to draw conclusions about the state of a single graph based on sampled subgraphs; the nature of a generative process that produces graphs of fixed size; or the properties of a generative process that produces graphs of arbitrary size. These scenarios are distinct, and have implications for both model specification and inference. The main goal of our paper is to clarify these distinct uses, with an eye

to facilitating more principled inference. We argue that concerns raised about the applications of ERGMs can be addressed by paying close attention to model specifications and their underlying assumptions, along with the inferential settings in which models are used. We review likelihood-based inference for well-posed ERGMs in finite, super and infinite population scenarios along with consistency and asymptotic normality results, and demonstrate likelihood-based inference in super population scenarios by an application to human brain networks.

1.1 Outline

To prepare the ground for the remainder of our paper, we first describe and address two questions that have been raised about the theoretical properties of ERGMs, stemming from the observation that some ERGMs are near-degenerate and that many ERGMs are non-projective. A close examination of these questions demonstrates the need for proper statistical language to clarify the core statistical notions of "sample" and "population" in the ERGM framework, and shows that a failure to use proper statistical language may result in misleading statistical conclusions. In the remainder of the paper, we therefore provide a review of ERGMs that:

- introduces exponential-family parameterizations of random graph models (Section 2);
- describes and addresses two questions that have been raised about the theoretical properties of ERGMs (Section 3), concerning model near-degeneracy (Section 3.1) and projectivity (Section 3.2);
- reviews ERGMs with additional structure, which helps construct ERGMs with attractive theoretical properties (Section 4);
- clarifies the core statistical notions of "sample" and "population" in the ERGM framework (Section 5), separating the complete-data generating process (the process that generates the population graph: Section 5.1) from the incomplete-data generating process (the observation process: Section 5.2);
- distinguishes statistical inference for graphs of fixed size, including finite and super population inference, and statistical inference for sequences of graphs of increasing size (Section 5);
- describes likelihood-based inference given incomplete data, generated by ignorable incomplete-data generating processes (Section 6);
- discusses consistency and asymptotic normality of likelihood-based estimators in finite, super and infinite population scenarios (Section 7);
- demonstrates likelihood-based inference for well-posed ERGMs in super population scenarios by an application to human brain networks (Section 8).

1.2 Topics Not Covered

The study of ERGMs is a rich area of statistical network analysis, with many computational and statistical topics that could be addressed. As we focus on the specification of well-posed ERGMs and accompanying inference scenarios, there are many other issues that space does not allow us to address. Here, we mention some of these topics, along with pointers to related work.

1.2.1 *Computational methods for ERGMs*. We do not cover computational methods for ERGMs, because computational issues are separable from the theoretical issues discussed here, and discussions can be found elsewhere in the literature. Many computational methods for ERGMs are closely related to computational methods for other discrete exponential-family models for dependent random variables, such as Ising models in physics and discrete Markov random fields in spatial statistics and machine learning. While discrete exponential-family models for dependent random variables come in countless forms and shapes, many of them pose similar computational challenges—for example, likelihood functions with intractable normalizing constants and posteriors with doubly intractable normalizing constants—so similar solutions are applicable to them. Without providing additional details, we mention here the main classes of computational approaches that have been explored to date:

- Pseudo-likelihood (Strauss and Ikeda, 1990, Leifeld, Cranmer and Desmarais, 2018) and composite-likelihood methods (Asuncion et al., 2010, Krivitsky, 2017), contrastive divergence (Asuncion et al., 2010) and approximations based on graph limits (He and Zheng, 2015); note that methods which are not based on the likelihood function need to be used with caution, as explained in Section 3.2.2.
- Stochastic approximation methods for maximum likelihood estimation (Snijders, 2002, Snijders and van Duijn, 2002, Jin and Liang, 2013).
- Monte Carlo maximization methods for maximum likelihood estimation (Geyer and Thompson, 1992, Handcock, 2003, Hunter and Handcock, 2006, Hummel, Hunter and Handcock, 2012, Okabayashi and Geyer, 2012, Yang, Rinaldo and Fienberg, 2014, Krivitsky, 2017, Byshkin et al., 2018).
- Bayesian Markov chain Monte Carlo methods (Koskinen, 2004, Møller et al., 2006, Murray, Ghahramani and MacKay, 2006, Caimo and Friel 2011, 2013, Everitt, 2012, Atchadé, Lartillot and Robert, 2013, Jin, Yuan and Liang, 2013, Liang and Jin, 2013, Wang and Atchadé, 2014, Caimo and Mira, 2015, Lyne et al., 2015, Liang et al., 2016, Park and Haran, 2018).
- Incomplete-data maximum likelihood and Bayesian estimation of ERGMs based on sampled data and missing data (Gile and Handcock, 2006, Handcock and Gile,

2010, Koskinen, Robins and Pattison, 2010, Snijders, 2010, Pattison et al., 2013, Krivitsky and Morris, 2017, Karwa, Krivitsky and Slavković, 2017).
- To generate draws of graphs from ERGMs, Markov chain Monte Carlo (Snijders, 2002, Morris, Handcock and Hunter, 2008) is the workhorse, though perfect sampling (Butts, 2018) and non-Markov chain Monte Carlo approximate sampling approaches (Butts, 2015) have also been developed.

1.2.2 *Mixtures of ERGMs*. As we focus on ERGMs per se, we do not cover random graph models that use ERGMs as building blocks, such as mixtures of ERGMs. Most mixtures of ERGMs are finite or infinite mixtures of Bernoulli random graphs. Examples are:

- Stochastic block models (e.g., Nowicki and Snijders, 2001, Airoldi et al., 2008, Bickel and Chen, 2009, Bickel, Chen and Levina, 2011, Choi, Wolfe and Airoldi, 2012, Rohe, Chatterjee and Yu, 2011, Zhao, Levina and Zhu, 2012, Amini et al., 2013, Lei and Rinaldo, 2015, Jin, 2015, Gao, Lu and Zhou, 2015, Zhang and Zhou, 2016, Binkiewicz, Vogelstein and Rohe, 2017, Sengupta and Chen, 2018).
- Latent space models (e.g., Hoff, Raftery and Handcock, 2002, Schweinberger and Snijders, 2003, Handcock, Raftery and Tantrum, 2007, Raftery et al., 2012, Salter-Townshend and Murphy, 2013, Tang, Sussman and Priebe, 2013, Sewell and Chen, 2015).
- Other latent variable models (e.g., van Duijn, 1995, van Duijn, Snijders and Zijlstra, 2004, Gill and Swartz, 2004, Hoff 2005, 2008, 2009, 2020, Fosdick and Hoff, 2015, Fosdick et al., 2019).

In many applications of such models, interest centers on unobserved structure underlying networks, for example, unobserved community structure in applications of stochastic block models. While we do not focus on such models, many of the issues we discuss—first and foremost the inference scenarios reviewed here—may inform treatment of ERGM mixtures as well.

1.2.3 *Exchangeable random graph models and other models*. Various other models have been proposed, such as exchangeable random graphs—random graphs invariant to the labeling of nodes (Diaconis and Janson, 2008, Bickel and Chen, 2009) or the labeling of edges (Diaconis and Janson, 2008, Cai, Campbell and Broderick, 2016, Janson, 2018, Crane and Dempsey 2018, 2020, Crane, 2018); scale-free networks (e.g., Barabási and Albert, 1999, Bollobás et al., 2001, Willinger, Alderson and Doyle, 2009); and other random graph models (e.g., Rapoport, 1979/80). The literature on such models is large and diverse, as are their motivations and properties, and their relationships to ERGMs have not been studied in depth. We hence do not attempt to treat these models here, but refer readers to the above-cited papers and other related literature (e.g., Diaconis and Janson, 2008, Lovász, 2012, Orbanz and Roy, 2015, Veitch, 2015, Caron and Fox, 2017, Janson, 2018, Lauritzen, Rinaldo and Sadeghi, 2018, Borgs et al., 2019, Veitch and Roy, 2019). However, we briefly discuss one notable example of exchangeable random graphs: the edge-exchangeable models of Crane and Dempsey (2018). Edge-exchangeable models assign equal probabilities to all edge-labeled graphs which are isomorphic up to relabeling (Definition 3.1, Crane and Dempsey, 2018). A special case of the generic modeling framework was elaborated by Crane and Dempsey (2018): the so-called "Hollywood model." The motivating example of the Hollywood model concerns actors linked by movies, although the generic modeling framework is not restricted to the motivating example. While such models hold promise, it is an open question whether, and how, those models can be used to address one of the key challenges of statistical network analysis: network data are dependent data, and testing and modeling dependencies in network data is of great interest. The range of dependencies captured by those models has not been characterized to date, and more research is needed to determine how such models can be used to test and model complex dependencies in network data, such as transitivity.

2. EXPONENTIAL-FAMILY PARAMETERIZATIONS OF RANDOM GRAPH MODELS

To introduce exponential-family parameterizations of random graph models, which are known as Exponential-family Random Graph Models (ERGMs), we consider a finite population of nodes $\mathcal{N} = \{1, \ldots, N\}$ $(N \geq 2)$. The nodes $i \in \mathcal{N}$ may have attributes (e.g., age). We denote the collection of attributes of population members by $\boldsymbol{x}_{\mathcal{N}} \in \mathcal{X}_{\mathcal{N}} \subseteq \mathbb{R}^q$. In addition to having attributes, nodes $i \in \mathcal{N}$ may be connected to other nodes $j \in \mathcal{N}$ by edges. In the simplest case, edges are undirected and self-edges are excluded, although many of the key ideas we discuss can be extended to directed edges and self-edges. Edges between nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ are considered random variables and can take on weights, denoted by $Y_{i,j}$. The weights of edges $Y_{i,j}$ can take on values in the following sets:

- $\mathcal{Y} = \{0, 1\}$, where 0 indicates the absence of an edge and 1 indicates the presence of an edge;
- $\mathcal{Y} = \{0, 1, \ldots\}$, where $0, 1, \ldots$ are counts of the number of interactions or other relational events;
- $\mathcal{Y} = \mathbb{R}$, indicating financial transactions or other relational events with real-valued outcomes.

The vast majority of ERGM-related publications focuses on random graphs $\boldsymbol{Y} = (Y_{i,j})_{i<j:i,j\in\mathcal{N}}$ with sample spaces of the form $\mathcal{Y}_{\mathcal{N}} = \{0, 1\}^{\binom{|\mathcal{N}|}{2}}$, but there are ERGMs for random graphs with sample spaces of the

form $\mathcal{Y}_\mathcal{N} = \{-1, 0, 1\}^{\binom{|\mathcal{N}|}{2}}$ (Huitsing et al., 2012), $\mathcal{Y}_\mathcal{N} = \{0, 1, \ldots\}^{\binom{|\mathcal{N}|}{2}}$ (Krivitsky, 2012), and $\mathcal{Y}_\mathcal{N} = \mathbb{R}^{\binom{|\mathcal{N}|}{2}}$ (Desmarais and Cranmer, 2012), as well as random graphs where edges are ranks (Krivitsky and Butts, 2017), categorical with unordered categories (Robins, Pattison and Wasserman, 1999) or ordered categories (Caimo and Gollini, 2020), and multivariate combinations thereof (Pattison and Wasserman, 1999, Lazega and Pattison, 1999, Krivitsky, Marcum and Koehly, 2019). In fact, there are entire R packages devoted to ERGMs for random graphs with sample spaces of the form $\mathcal{Y}_\mathcal{N} = \{0, 1, \ldots\}^{\binom{|\mathcal{N}|}{2}}$ and $\mathcal{Y}_\mathcal{N} = \mathbb{R}^{\binom{|\mathcal{N}|}{2}}$.

To cover random graphs with sample spaces of the form $\mathcal{Y}_\mathcal{N} = \{0, 1\}^{\binom{|\mathcal{N}|}{2}}$, $\mathcal{Y}_\mathcal{N} = \{0, 1, \ldots\}^{\binom{|\mathcal{N}|}{2}}$, $\mathcal{Y}_\mathcal{N} = \mathbb{R}^{\binom{|\mathcal{N}|}{2}}$, and other sample spaces that have been explored in the literature, we consider exponential families of densities with respect to a σ-finite reference measure ν with support $\mathcal{Y}_\mathcal{N}$, specified by a sufficient statistic $\boldsymbol{s} : \mathcal{X}_\mathcal{N} \times \mathcal{Y}_\mathcal{N} \mapsto \mathbb{R}^p$ and a map $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathcal{N} \mapsto \mathbb{R}^p$ with $\boldsymbol{\Theta} \subseteq \{\boldsymbol{\theta} \in \mathbb{R}^p : \psi(\boldsymbol{\theta}, \mathcal{N}) < \infty\}$:

$$\frac{\mathrm{d}\,\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}}{\mathrm{d}\,\nu}(\boldsymbol{y}_\mathcal{N}) = \exp\big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_\mathcal{N}, \boldsymbol{y}_\mathcal{N})\rangle - \psi(\boldsymbol{\theta}, \mathcal{N})\big),$$

where $\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_\mathcal{N}, \boldsymbol{y}_\mathcal{N})\rangle$ denotes the inner product of natural parameter $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N})$ and sufficient statistic $\boldsymbol{s}(\boldsymbol{x}_\mathcal{N}, \boldsymbol{y}_\mathcal{N})$ and

$$\psi(\boldsymbol{\theta}, \mathcal{N}) = \log \int_{\mathcal{Y}_\mathcal{N}} \exp\big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_\mathcal{N}, \boldsymbol{y}'_\mathcal{N})\rangle\big)\, \mathrm{d}\,\nu(\boldsymbol{y}'_\mathcal{N}).$$

To present the key ideas in the simplest possible setting, we focus henceforth on the simplest and most common case: exponential-family models of random graphs with sample spaces of the form $\mathcal{Y}_\mathcal{N} = \{0, 1\}^{\binom{|\mathcal{N}|}{2}}$, in which case

$$\begin{aligned}&\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_\mathcal{N} = \boldsymbol{y}_\mathcal{N})\\ &\quad= \exp\big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_\mathcal{N}, \boldsymbol{y}_\mathcal{N})\rangle - \psi(\boldsymbol{\theta}, \mathcal{N})\big)\, \nu(\boldsymbol{y}_\mathcal{N}),\end{aligned}$$

where

$$\psi(\boldsymbol{\theta}, \mathcal{N}) = \log \sum_{\boldsymbol{y}'_\mathcal{N} \in \mathcal{Y}_\mathcal{N}} \exp\big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_\mathcal{N}, \boldsymbol{y}'_\mathcal{N})\rangle\big)\nu(\boldsymbol{y}'_\mathcal{N}).$$

All of these quantities can depend on the population of nodes $\mathcal{N}$.

The generic exponential-family framework may be intimidating and the reader may question the high level of abstraction and generality of the above definition, not the least the fact that all quantities are allowed to depend on the population of nodes $\mathcal{N}$. A simple example may help demonstrate why all quantities can depend on $\mathcal{N}$, and why it is desirable to allow them to depend on $\mathcal{N}$. Consider the family of sparse Bernoulli$(\pi_{|\mathcal{N}|})$ random graphs with size-dependent edge probabilities $\pi_{|\mathcal{N}|} = \mathrm{logit}^{-1}(\theta - \log|\mathcal{N}|)$ $(\theta \in \mathbb{R})$, with probability mass function

$$\begin{aligned}&\mathbb{P}_{\mathcal{N},\eta(\theta,\mathcal{N})}(\boldsymbol{Y}_\mathcal{N} = \boldsymbol{y}_\mathcal{N})\\ &\quad= \exp\left(\eta(\theta, \mathcal{N}) \sum_{i<j:\, i,j\in\mathcal{N}} y_{i,j} - \psi(\theta, \mathcal{N})\right) \nu(\boldsymbol{y}_\mathcal{N}),\end{aligned}$$

where the support of the reference measure depends on $\mathcal{N}$:

$$\nu(\boldsymbol{y}_\mathcal{N}) = \begin{cases} 1 & \text{if } \boldsymbol{y}_\mathcal{N} \in \{0, 1\}^{\binom{|\mathcal{N}|}{2}} \\ 0 & \text{otherwise,} \end{cases}$$

as do the natural parameter $\eta(\theta, \mathcal{N}) = \theta - \log|\mathcal{N}|$ and $\psi(\theta, \mathcal{N}) = \log(1 + \exp(\theta - \log|\mathcal{N}|))$. In other words, all quantities depend on the population of nodes $\mathcal{N}$. Indeed, the dependence on $\mathcal{N}$ stems from the offset $\log|\mathcal{N}|$, which induces sparsity in Bernoulli random graphs. We motivate sparsity and sparsity-inducing Bernoulli$(\pi_{|\mathcal{N}|})$ random graphs in Section 3.2. Covariates can be included by using the logit link function $\mathrm{logit}(\pi_{|\mathcal{N}|}) = \theta - \log|\mathcal{N}|$ and adding covariate terms, as in logistic regression (McCullagh and Nelder, 1983). Covariate terms are reviewed in Morris, Handcock and Hunter (2008). Examples of other ERGMs with support $\mathcal{Y}_\mathcal{N} = \{0, 1\}^{\binom{|\mathcal{N}|}{2}}$ can be found throughout the paper. ERGMs with other forms of support can be found in the literature cited above.

3. TWO QUESTIONS RAISED ABOUT ERGMS AND CLARIFICATIONS

Here, we describe two questions that have been raised about the theoretical properties of ERGMs, discuss the historical and mathematical context in which those questions arose, and outline which lessons have been learned and how the associated issues have been addressed. The resulting discussion motivates a more careful look at the specification of ERGMs and inference for ERGMs, which are discussed in the following sections.

3.1 Question I: Are Non-Trivial ERGMs Near-Degenerate?

3.1.1 *Overview and history.* A common concern is that ERGMs with non-trivial dependence structure can be ill-behaved, in the sense that ERGMs can be either near-degenerate or indistinguishable from Bernoulli random graphs with, in some cases, a phase transition between these two regimes (Strauss, 1986, Jonasson, 1999, Häggström and Jonasson, 1999, Handcock, 2003, Bhamidi, Bresler and Sly, 2008, 2011, Rinaldo, Fienberg and Zhou, 2009, Schweinberger, 2011, Butts, 2011, Chatterjee and Diaconis, 2013, Mele, 2017, Bhamidi et al., 2018). Both near-degenerate ERGMs and near-Bernoulli models are problematic as models of network data: near-degenerate ERGMs concentrate probability mass on small subsets of graphs, such as near-complete graphs with almost all possible edges, whereas near-Bernoulli random graphs induce vanishing dependence.

These properties render them useless in most applications, although there are exceptions. One notable exception is the behavior of some physical systems (e.g., networks of crystal contacts or amlyoid fibrils), which resembles the behavior of near-degenerate ERGMs—including the existence of phase transitions. In such cases, the behavior of near-degenerate ERGMs can be both realistic and desirable (Grazioli et al., 2019). However, we follow here convention and treat such behavior as undesirable in most cases.

The fact that some ERGMs are ill-behaved was first discovered by Strauss (1986), Jonasson (1999) and Häggström and Jonasson (1999). We discuss the theoretical results of these pioneers in Section 3.1.2 along with more recent work. In practice, the undesirable properties of ill-behaved ERGMs went unnoticed at first, in part because pseudo-likelihood-based methods—which masked the undesirable properties of ill-behaved ERGMs—were used to estimate them (Strauss and Ikeda, 1990), and in part because model assessment tools were unavailable. The introduction of Markov chain Monte Carlo methods for generating draws of graphs from ERGMs in the 1990s and 2000s revealed that some models estimated by maximum pseudo-likelihood methods performed very poorly (generating, e.g., graphs with almost all possible edges when estimated from observed graphs with a moderate number of edges). At the time, this was believed to be due to poor estimation, and indeed maximum pseudo-likelihood estimators were shown to be inferior to maximum likelihood estimators (Dahmström and Dahmström 1993, 1999, Corander, Dahmström and Dahmström 1998, 2002). In time, Monte Carlo maximum likelihood estimators were developed (Snijders, 2002, Hunter and Handcock, 2006), along with simulation-based model assessment tools (Hunter, Goodreau and Handcock, 2008). These developments revealed that, while some of the bad behavior did stem from inferior estimators, the bad behavior in other cases was inherent to the specified models, as first pointed out by Snijders (2002) and Handcock (2003) and anticipated by the work of Strauss (1986), Jonasson (1999) and Häggström and Jonasson (1999). Since then, statistical theory has shed more light on those undesirable properties, and has led to the development of improved model specifications. We review the existing theoretical results in Section 3.1.2, and improved model specifications in Section 3.1.3.

3.1.2 *Clarification*: *Most theoretical results are limited to simplistic ERGMs that lack structure*. Since the 1980s (Strauss, 1986, Jonasson, 1999, Häggström and Jonasson, 1999), it has been known that some ERGMs are ill-behaved. However, two important points have been lost in more recent discussions of ill-behaved ERGMs. First, almost all theoretical results—discussed below—are limited to simplistic ERGMs which lack structure that could restrict interactions among edge variables, and which have the same number of natural parameters, regardless of how large $|\mathcal{N}|$ is. Such simplistic ERGMs resemble Ising models in physics without lattice structure or discrete Markov random fields in spatial statistics without spatial structure, and are of limited use in understanding large networks with complex dependence. Theoretical results based on such models are therefore limited in scope, and must be interpreted with careful attention to the underlying assumptions. As we shall discuss, most of those results do not generalize to ERGMs with additional structure. Indeed, well-posed ERGMs with additional structure can be well-behaved and are widely used in practice. We discuss the first point below and the second point in Section 3.1.3.

The first theoretical results on ill-behaved ERGMs were reported by Strauss (1986), Jonasson (1999) and Häggström and Jonasson (1999). Strauss (1986) pointed out that the Markov random graphs of Frank and Strauss (1986), an important class of ERGMs, induce long-range dependence by allowing each edge variable to interact with $2(|\mathcal{N}| - 2)$ other edge variables. The long-range dependence is rooted in the lack of structure of those models: without additional structure, it is difficult to constrain the range of interactions. If strong homogeneity assumptions are imposed on Markov random graphs, long-range dependence results in strong dependence and model near-degeneracy: that is, Markov random graphs concentrate probability mass on a small subset of graphs, for example, graphs with almost no edges or almost all possible edges (Strauss, 1986). Jonasson (1999) and Häggström and Jonasson (1999) studied the model near-degeneracy and phase transitions of the triangle model, a special case of Markov random graphs, and concluded that the model near-degeneracy of the triangle model is rooted in the lack of structure of the model. More work on model near-degeneracy can be found in Schweinberger (2011), Butts (2011), Chatterjee and Diaconis (2013), Mele (2017), and Bhamidi et al. (2018), and more work on phase transitions in Chatterjee and Diaconis (2013), Mukherjee (2013a), Radin and Yin (2013), Aristoff and Radin (2013), Yin, Rinaldo and Fadnavis (2016) and Kenyon and Yin (2017); some related work on phase transitions in the physics literature can be found in Park and Newman (2004, 2005). Handcock (2003) studied the implications of model near-degeneracy in terms of statistical inference, including the existence of maximum likelihood and Monte Carlo maximum likelihood estimators, and argued that Monte Carlo maximum likelihood estimators frequently do not exist due to model near-degeneracy, resulting in computational failure. Rinaldo, Fienberg and Zhou (2009) investigated these existence issues in more depth by studying the geometry of ERGMs. The fact that near-degenerate ERGMs have regimes that can be approximated by Bernoulli random graphs in large random graphs was first pointed out

by Bhamidi, Bresler and Sly (2008, 2011), with more work by Chatterjee and Diaconis (2013), Mele (2017) and Bhamidi et al. (2018).

Despite the insights gained by these theoretical results, it is important to keep in mind that these results are limited to simplistic ERGMs that lack structure and have the same number of natural parameters, regardless of how large $|\mathcal{N}|$ is (with one exception, Schweinberger, 2011, which we discuss in Section 3.1.3). In particular, these results do not cover ERGMs with additional structure and ERGMs for which the number of natural parameters increases with $|\mathcal{N}|$, which can be better-behaved. A related limitation of these results is that many real-world settings (e.g., families, school classrooms, local groups of insurgents and terrorist cells) involve bounded networks with small numbers of nodes (e.g., 5–50). In such situations, some "near-degenerate" model specifications may in fact be well-behaved. Thus, theoretical results suggesting problematic behavior must be considered within the context in which the model is employed (and should checked by simulating graphs from the model, as advised by Hunter, Goodreau and Handcock, 2008).

3.1.3 *Clarification*: *Well-posed ERGMs with additional structure can be well-behaved*. The most important lesson from the theoretical results on simplistic ERGMs is that ERGMs for large networks need additional structure. The pioneers (Strauss, 1986, Strauss and Ikeda, 1990, Jonasson, 1999, Häggström and Jonasson, 1999) understood full well that the undesirable behavior of simplistic ERGMs is rooted in the lack of structure of those models, compared with Ising models in physics, which have additional structure in the form of lattice structure; and discrete Markov random fields in spatial statistics, which have additional structure in the form of spatial structure. To address the lack of structure of simplistic ERGMs, the pioneers suggested to endow ERGMs with additional structure. For instance, Strauss and Ikeda (1990) introduced Markov random graphs with observed blocks—using categorical covariates to partition a set of nodes into subsets (blocks)—to constrain the dependence of Markov random graphs to sets of edge variables within blocks, noting that *"Markov models without blocks are unsuitable for large data sets because of the possibility of degeneracy"* (Strauss and Ikeda, 1990, p. 206). In the special case of the triangle model, Jonasson (1999) concluded: *"the random triangle model is explosive; depending on q we get nothing or everything. The important moral of this is that for any random graph model with transitivity not degenerate in this sense, the non-degeneracy relies on the extra, and perhaps unintended, structure imposed on the graph"* (Jonasson, 1999, p. 866).[2]

[2]The parameter q mentioned by Jonasson (1999) is equivalent to $q = \exp(\theta_2)$, where θ_2 is the triangle parameter of the triangle model stated above.

In the ERGM framework, many possible forms of additional structure exist, including block structure, multilevel structure, spatial structure or temporal structure. Additional structure can be used to construct well-behaved ERGMs. We review ERGMs with additional structure in Section 4. As a motivating example, however, we highlight one important instance of ERGMs with additional structure: curved ERGMs with geometrically weighted model terms (Snijders et al., 2006, Hunter and Handcock, 2006, Hunter, 2007). Curved ERGMs, which are curved exponential families in the sense of Efron (1975, 1978), impose additional structure in the form of nonlinear constraints on the natural parameter space of the exponential family (Barndorff-Nielsen, 1978, Brown, 1986). The additional structure helps construct better-behaved models. To demonstrate, first consider the ill-behaved triangle model studied by Strauss (1986), Jonasson (1999), Häggström and Jonasson (1999) and others. The triangle model assumes that the probability mass function of a population graph $\boldsymbol{Y}_{\mathcal{N}} \in \{0,1\}^{\binom{|\mathcal{N}|}{2}}$ is of the form

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \propto \exp\left(\theta_1 \sum_{i<j:\, i,j\in\mathcal{N}} y_{i,j} + \theta_2 \sum_{i<j<k:\, i,j,k\in\mathcal{N}} y_{i,j}\, y_{j,k}\, y_{i,k}\right),$$

where $\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}) = \boldsymbol{\theta} \in \mathbb{R}^2$, $\sum_{i<j:\, i,j\in\mathcal{N}} y_{i,j}$ is the number of edges, and $\sum_{i<j<k:\, i,j,k\in\mathcal{N}} y_{i,j}\, y_{j,k}\, y_{i,k}$ is the number of triangles. The triangle model with $\theta_2 > 0$ rewards transitivity by rewarding triangles. While transitivity is an important feature of many real-world networks (Holland and Leinhardt 1970, 1972, 1976), the triangle model assumes that the added value of additional triangles does not decrease: the log odds of a graph with a edges and b triangles relative to a graph with a edges and $b-1$ triangles is

$$\log \frac{\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \text{graph with } a \text{ edges}, b \text{ triangles})}{\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \text{graph with } a \text{ edges}, b-1 \text{ triangles})} = \theta_2.$$

Note that the log odds does not depend on the number of triangles in the graph. Indeed, for each pair of nodes, each additional triangle contributes the same amount to the log odds of the conditional probability of an edge, regardless of the number of triangles in which the two nodes are already involved. This, upon reflection, is counterintuitive: it makes little sense for the tenth shared partner to carry as much weight as the first. Indeed, the assumption that the added value of additional triangles does not decrease is not without consequences, giving rise to the undesirable behavior described in Section 3.1.2. To ensure that the added value of additional triangles decreases, curved ERGMs with Geometrically Weighted Edgewise Shared Partner (GWESP) terms and other model terms have been developed (Snijders et al., 2006, Hunter and

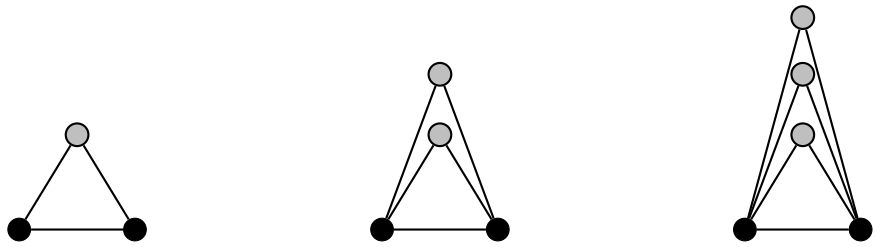

FIG. 1. *A graphical representation of a connected pair of nodes (represented by two black circles connected by a line), which has 1 edgewise shared partner (left), 2 edgewise shared partners (middle), and 3 edgewise shared partners (right).*

Handcock, 2006, Hunter, 2007). These models assume that the probability mass function of a population graph $\boldsymbol{Y}_{\mathcal{N}} \in \{0,1\}^{\binom{|\mathcal{N}|}{2}}$ is of the form

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \propto \exp\left(\eta_1(\boldsymbol{\theta},\mathcal{N}) \sum_{i<j:\, i,j\in\mathcal{N}} y_{i,j} + \sum_{m=1}^{|\mathcal{N}|-2} \eta_{1+m}(\boldsymbol{\theta},\mathcal{N})\, s_m(\boldsymbol{y}_{\mathcal{N}})\right),$$

where the sufficient statistics of the exponential family are the number of edges and the number of connected pairs of nodes with m edgewise shared partners, $s_m(\boldsymbol{y}_{\mathcal{N}})$, and the natural parameters of the exponential family are

$$\eta_1(\boldsymbol{\theta},\mathcal{N}) = \theta_1,$$
$$\eta_{1+m}(\boldsymbol{\theta},\mathcal{N}) = \theta_2 \exp(\theta_3)\big[1 - \big(1 - \exp(-\theta_3)\big)^m\big],$$

where $m = 1, \ldots, |\mathcal{N}| - 2$. A graphical representation of connected pairs of nodes with 1, 2 and 3 edgewise shared partners can be found in Figure 1.

If $\theta_2 > 0$ and $\theta_3 > 0$, the model rewards triangles, but ensures that the added value of additional triangles decreases. To see that, consider a connected pair of nodes $\{i, j\}$. The log odds of a graph where $\{i, j\}$ has m shared partners and hence m triangles relative to a graph where $\{i, j\}$ has $m - 1$ triangles is, assuming everything else is the same, given by

$$\theta_2\big(1 - \exp(-\theta_3)\big)^{m-1}, \quad m = 1, \ldots, |\mathcal{N}| - 2.$$

In other words, the added value of additional triangles decays at a geometric rate, provided $\theta_2 > 0$ and $\theta_3 > 0$. A graphical representation of the added value of additional triangles is shown in Figure 2. Curved ERGMs with geometrically weighted model terms are well-posed as long as $\theta_3 \geq 0$; note that $\theta_3 \in [-\log 2, 0)$ implies that the added value of the m-th triangle either decreases or increases, depending on the sign of θ_2 and whether m is even or odd, and that $\theta_3 \in (-\infty, -\log 2)$ implies a form of model near-degeneracy when $|\mathcal{N}|$ is large (Schweinberger, 2011). In practice, curved ERGMs with GWESP terms and other geometrically weighted model terms have turned out to be well-behaved in a wide range of settings; selected applications can be found in Snijders et al. (2006), Hunter and Handcock (2006), Hunter (2007), Hunter, Goodreau and Handcock (2008), Goodreau, Kitts and Morris (2009), Gile and Handcock (2006), Handcock and Gile (2010), Koskinen, Robins and Pattison (2010), Simpson, Hayasaka and Laurienti (2011), Suesse (2012), Rolls et al. (2013), Wang et al. (2013), Obando and De Vico Fallani (2017), Gondal (2018), Almquist and Bagozzi (2019), and Stewart et al. (2019). We apply curved ERGMs to human brain network data in Section 8 to demonstrate some of the benefits that curved ERGMs can provide in applications, compared with Bernoulli random graphs and latent space models.

In addition to improved model behavior, there exist consistency results for curved ERGMs with GWESP

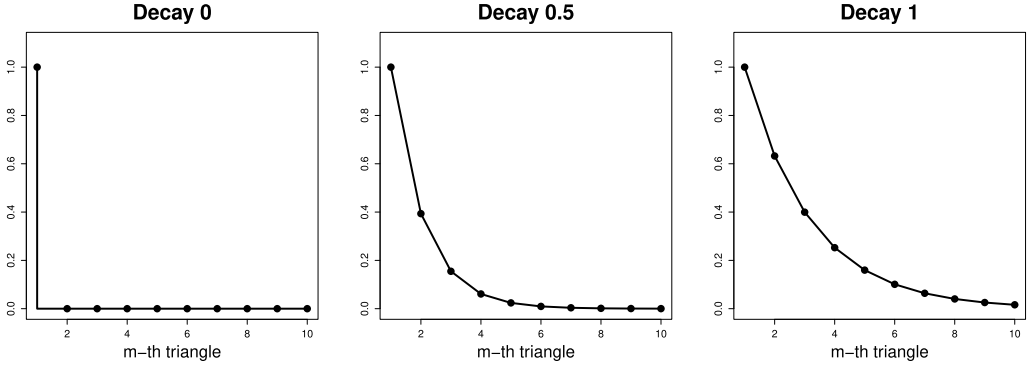

FIG. 2. *The added value of the m-th triangle in terms of $\theta_2(1-\exp(-\theta_3))^{m-1}$, $m = 1, 2, \ldots$. The plots are based on the so-called base parameter $\theta_2 = 1$ and the so-called decay parameter $\theta_3 = 0$ (left), $\theta_3 = 0.5$ (middle) and $\theta_3 = 1$ (right).*

terms and other curved ERGMs. We review consistency results for curved ERGMs in Section 7. It is worth noting that the results of Chatterjee and Diaconis (2013) and others do not preclude consistency results for curved ERGMs: the results of Chatterjee and Diaconis (2013) are limited to ERGMs which are canonical exponential families with a fixed number of natural parameters of the form $\eta_i(\boldsymbol{\theta}) = \theta_i$, and those results do not cover curved exponential families with an increasing number of natural parameters subject to nonlinear constraints, such as curved ERGMs with GWESP terms and $1 + |\mathcal{N}| - 2$ natural parameters.

3.2 Question II: Must ERGMs Be Projective for Well-Posed Inference?

3.2.1 *Overview and history.* A second question that has been raised is whether ERGMs must satisfy a property called projectivity for statistical inference to be well-posed. This question arises from the observation that many ERGMs are not projective, as first pointed out by Lauritzen (2008) and Snijders (2010). Necessary and sufficient conditions for projectivity of ERGMs with size-invariant natural parameters of fixed dimension were provided by Shalizi and Rinaldo (2013); some follow-up work can be found in Lauritzen, Rinaldo and Sadeghi (2018). Other notable work related to the topic of projectivity—outside of the ERGM framework—can be found in, for example, Caron and Fox (2017), Crane (2018) and Crane and Dempsey (2020).

In the ERGM framework, projectivity may be defined as follows. Consider a subgraph $\boldsymbol{y}_{\mathcal{N}'}$ of a population graph $\boldsymbol{y}_{\mathcal{N}}$ induced by a subset of nodes $\mathcal{N}' \subset \mathcal{N}$, that is, the subgraph with the set of nodes $\mathcal{N}'$ and all edges among nodes in $\mathcal{N}'$ contained in the population graph $\boldsymbol{y}_{\mathcal{N}}$. An ERGM is called projective if

$$\boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}') = \boldsymbol{\theta} \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and all } \mathcal{N}' \subseteq \mathcal{N}$$

and

$$\mathbb{P}_{\mathcal{N}',\boldsymbol{\theta}}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}) = \mathbb{P}_{\mathcal{N},\boldsymbol{\theta}}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}, \boldsymbol{Y}_{\mathcal{N}\setminus\mathcal{N}'} \in \mathcal{Y}_{\mathcal{N}\setminus\mathcal{N}'}),$$

where $\boldsymbol{y}_{\mathcal{N}\setminus\mathcal{N}'} \in \mathcal{Y}_{\mathcal{N}\setminus\mathcal{N}'}$ denotes the subset of possible edges of the population graph that are not contained in the subgraph induced by $\mathcal{N}'$. The notion of projectivity above was advanced by Shalizi and Rinaldo (2013), who focused on ERGMs with counting measure as reference measure, although Shalizi and Rinaldo considered more general reference measures in a supplement. We consider here the more general definition, covering all σ-finite reference measures. It is worth noting that there are weaker forms of projectivity in the ERGM framework, for example, conditional projectivity (Snijders, 2010) and block projectivity (Schweinberger and Handcock, 2015). Here, we restrict attention to the above definition of projectivity.

Projectivity is a form of closure under marginalization, and implies that the same parameters govern the population graph and the marginal distributions of all of its subgraphs. While closure under marginalization is convenient on mathematical grounds, it embodies the strong assumption that no subgraph of the population graph is affected by its embeddedness in the population graph. For example, consider 3 people attending a party of 30. To be projective, the probabilities of interactions among those 3 people must neither be affected by interactions of those 3 with the 27 others nor by the interactions among the 27 others. In other words, the probabilities of interactions among the 3 people must be the same, regardless of whether 27 other people attend the party. Such an assumption may not be satisfied by many networks (human or otherwise). Not surprisingly, many ERGMs are not projective, including some of the simplest and most classic random graph models.

For instance, sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs with size-dependent edge probabilities $\pi_{|\mathcal{N}|}$ (Gilbert, 1959, Erdős and Rényi 1959, 1960) are not projective, because the edge probabilities $\pi_{|\mathcal{N}|}$ decrease with the size $|\mathcal{N}|$ of $\mathcal{N}$. Sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs assume that the edge variables $Y_{i,j}$ are independent Bernoulli($\pi_{|\mathcal{N}|}$) random variables, and that the expected number of edges $\binom{|\mathcal{N}|}{2}\pi_{|\mathcal{N}|}$ grows slower than $\binom{|\mathcal{N}|}{2}$, which implies that $\pi_{|\mathcal{N}|}$ must decrease with $|\mathcal{N}|$. Consider the parameterization $\pi_{|\mathcal{N}|} = \text{logit}^{-1}(\theta - \log|\mathcal{N}|)$ ($\theta \in \mathbb{R}$), which implies

$$\begin{aligned}
&\mathbb{P}_{\mathcal{N},\eta(\theta,\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \\
&\quad = \prod_{i<j:\, i,j\in\mathcal{N}} \pi_{|\mathcal{N}|}^{y_{i,j}}(1-\pi_{|\mathcal{N}|})^{1-y_{i,j}} \\
&\quad \propto \exp\left(\eta(\theta,\mathcal{N}) \sum_{i<j:\, i,j\in\mathcal{N}} y_{i,j}\right) \nu(\boldsymbol{y}_{\mathcal{N}}),
\end{aligned}$$

where the natural parameter is

$$\eta(\theta, \mathcal{N}') = \theta \quad \text{for all } \theta \in \mathbb{R} \text{ and all } \mathcal{N}' \subseteq \mathcal{N}$$

and the offset $\log|\mathcal{N}|$ has been absorbed into the reference measure:

$$\nu(\boldsymbol{y}_{\mathcal{N}}) = \begin{cases} \exp\left(-\log|\mathcal{N}| \sum_{i<j:\, i,j\in\mathcal{N}} y_{i,j}\right) & \text{if } \boldsymbol{y}_{\mathcal{N}} \in \{0,1\}^{\binom{|\mathcal{N}|}{2}} \\ 0 \quad \text{otherwise.} \end{cases}$$

Sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs are not projective, because $|\mathcal{N}'| \neq |\mathcal{N}|$ implies $\pi_{|\mathcal{N}'|} \neq \pi_{|\mathcal{N}|}$, so

$$\begin{aligned}
&\prod_{i<j:\, i,j\in\mathcal{N}'} \pi_{|\mathcal{N}'|}^{y_{i,j}}(1-\pi_{|\mathcal{N}'|})^{1-y_{i,j}} \\
&\quad \neq \prod_{i<j:\, i,j\in\mathcal{N}'} \pi_{|\mathcal{N}|}^{y_{i,j}}(1-\pi_{|\mathcal{N}|})^{1-y_{i,j}},
\end{aligned}$$

and hence

$$\mathbb{P}_{\mathcal{N}',\theta}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}) \neq \mathbb{P}_{\mathcal{N},\theta}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}, \boldsymbol{Y}_{\mathcal{N}\setminus\mathcal{N}'} \in \mathcal{Y}_{\mathcal{N}\setminus\mathcal{N}'}).$$

It is worth noting that, despite the lack of projectivity, sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs have meaningful asymptotic behavior: for example, the expected number of edges of each node tends to the constant $\exp(\theta)$ as $|\mathcal{N}| \to \infty$ (Krivitsky, Handcock and Morris, 2011). Thus, lack of projectivity does not rule out meaningful asymptotic behavior. Indeed, sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs have many interesting asymptotic properties and have been studied in random graph theory since the seminal work of Gilbert (1959) and Erdős and Rényi (1959, 1960): see, for example, the classic monograph of Bollobás (1985) and the more recent books of Janson, Łuczak and Rucinski (2000) and Frieze and Karoński (2016). Moreover, the sparsity-inducing reference measure $\nu(\boldsymbol{y}_{\mathcal{N}})$ can be shown to arise from a simple stochastic process with an attractive interpretation (Butts, 2019), providing a substantive motivation for its use.

The above suggests that projectivity is too restrictive to be a basis for building plausible models of complex and dependent network data. Nonetheless, projectivity is convenient, in that it provides a simple basis for extending an ERGM from a subgraph to the population graph. This raises the question of whether, in the absence of projectivity, population probability models can be inferred from a subgraph of the population graph.

3.2.2 *Clarification*: *The likelihood function is not affected by lack of projectivity.* While many ERGMs are not projective, lack of projectivity does not preclude likelihood-based inference, for at least two reasons. First, the likelihood function is not affected by lack of projectivity. Second, lack of projectivity does not imply that likelihood-based estimators are inconsistent. In fact, consistency results for likelihood-based estimators of non-projective ERGMs do exist. Taken together, these two points imply that likelihood-based inference for well-posed ERGMs is possible despite lack of projectivity. We discuss the first point below and the second point in Section 3.2.3.

The first important point is that the likelihood function is not affected by lack of projectivity. Consider the motivating example of Shalizi and Rinaldo (2013): there is a finite population of nodes $\mathcal{N}$ and a population graph $\boldsymbol{y}_{\mathcal{N}}$ is generated by a population probability model $\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})$. This motivating example is representative of other ERGM applications: for example, in the human brain network example in Section 8, the population of nodes $\mathcal{N}$ corresponds to 56 regions of the human brain.

To derive the likelihood function in the motivating example, it is imperative to separate the process that generates the population graph from the process that generates observations of edges in the population graph. We follow here the principled approach of Fisher (1922) and Rubin (1976) to likelihood-based inference in complete- and incomplete-data scenarios, which was adapted to ERGMs by Gile and Handcock (2006), Handcock and Gile (2010) and Koskinen, Robins and Pattison (2010). If the whole population graph $\boldsymbol{y}_{\mathcal{N}}$ is observed, the likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{N}}) \propto \mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}).$$

If a subgraph $\boldsymbol{y}_{\mathcal{N}'}$ of $\boldsymbol{y}_{\mathcal{N}}$ induced by a subset of nodes $\mathcal{N}' \subset \mathcal{N}$ is observed, generated by an incomplete-data generating process that is ignorable in the sense of Rubin (1976), the likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{N}'}) \propto \mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}, \boldsymbol{Y}_{\mathcal{N}\setminus\mathcal{N}'} \in \mathcal{Y}_{\mathcal{N}\setminus\mathcal{N}'}).$$

In other words, the likelihood function can be obtained by summing the population probability mass function $\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})$ with respect to the unobserved edges.

While deriving the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{N}'})$ is trivial from a mathematical point of view—although computing it may not be trivial, as discussed below—the form of the likelihood function has important statistical implications. First, as Shalizi and Rinaldo (2013) point out, it is problematic to base inference on $\mathbb{P}_{\mathcal{N}',\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}')}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'})$, because $\mathbb{P}_{\mathcal{N}',\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}')}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'})$ may not be relatable to $\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})$ when the model is not projective. It is therefore comforting to know that likelihood-based inference is not based on $\mathbb{P}_{\mathcal{N}',\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}')}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'})$, but is based on the marginal probability mass function $\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}, \boldsymbol{Y}_{\mathcal{N}\setminus\mathcal{N}'} \in \mathcal{Y}_{\mathcal{N}\setminus\mathcal{N}'})$ induced by $\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})$. Indeed, the marginal probability mass function is related to $\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})$ by marginalization, regardless of whether the model is projective. As a result, the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{N}'})$ is not affected by lack of projectivity. Projectivity does facilitate the evaluation of the likelihood function, but the likelihood function does not require it.

Statistical implications. The results of Shalizi and Rinaldo (2013) underscore the importance of likelihood-based inference: statistical inference for ERGMs should be based on the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{N}'})$, which is not affected by lack of projectivity. Indeed, since the 1990s, ERGM experts have known that non-likelihood-based inference is problematic. Dahmström and Dahmström (1993, 1999), Corander, Dahmström and Dahmström (1998, 2002), Lubbers and Snijders (2007), and Van Duijn, Gile and Handcock (2009) compared likelihood-based and non-likelihood-based estimators by using exact computations, based on complete enumeration of all possible graphs of small sizes, along with simulation studies and data analyses for graphs of larger sizes. All of them concluded that non-likelihood-based inference,

in particular pseudo-likelihood-based inference (Strauss and Ikeda, 1990), tends to be inferior to likelihood-based inference—at least when the dependence induced by the model is strong and can propagate throughout the population graph. We review likelihood-based inference given incomplete data in Section 6.

Computational implications. While projectivity is not necessary, it is convenient for the purpose of evaluating the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathbb{N}'})$. If the model is projective, summing $\mathbb{P}_{\mathbb{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathbb{N})}(\boldsymbol{Y}_{\mathbb{N}} = \boldsymbol{y}_{\mathbb{N}})$ with respect to the unobserved edges by using computational methods is unnecessary, because the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathbb{N}'})$ reduces to $\mathbb{P}_{\mathbb{N}',\boldsymbol{\eta}(\boldsymbol{\theta},\mathbb{N}')}(\boldsymbol{Y}_{\mathbb{N}'} = \boldsymbol{y}_{\mathbb{N}'})$. Otherwise, one needs to sum $\mathbb{P}_{\mathbb{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathbb{N})}(\boldsymbol{Y}_{\mathbb{N}} = \boldsymbol{y}_{\mathbb{N}})$ with respect to the unobserved edges by using computational methods, either exactly or approximately, by using Markov chain Monte Carlo methods (Gile and Handcock, 2006, Handcock and Gile, 2010, Koskinen, Robins and Pattison, 2010). These computational challenges are the same as in other discrete exponential-family models for dependent random variables, such as discrete Markov random fields in spatial statistics (Besag, 1974, Cressie, 1993, Stein, 1999) and machine learning (e.g., Ravikumar, Wainwright and Lafferty, 2010, Yang et al., 2015). In many applications of ERGMs, the required computations are feasible, because either the population of interest is not too large or the population has additional structure that facilitates computations.

An example of small populations are the human brain networks used in Section 8: each of the 108 human brain networks has 56 nodes and is therefore small enough to approximate the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathbb{N}'})$ by using Markov chain Monte Carlo methods when network data are sampled or missing.

Examples of populations with additional structure are populations with block structure, multilevel structure, spatial structure and temporal structure. Some ERGMs with additional structure have factorization properties. Factorization properties facilitate likelihood-based computations by allowing to break down likelihood functions into parts, and the parts may be computed by using parallel computing on multi-core computers and computing clusters. We review ERGMs with additional structure, including ERGMs with factorization properties, in Section 4.

Such likelihood-based methods are implemented in many ERGM-related statistical software packages—including 20 ERGM-related `R` packages and the program `pnet` (Wang, Robins and Pattison, 2006)—and the computational burden has not prevented network scientists from applying ERGMs to a substantial number of real-world problems (with some examples given in Section 1).

3.2.3 *Clarification*: *Consistency results for likelihood-based estimators of non-projective ERGMs do exist*. The second important point regarding projectivity is that lack of projectivity does not imply that estimators of $\boldsymbol{\theta}$ are inconsistent. Shalizi and Rinaldo (2013) showed that projectivity is a *sufficient* condition for consistency of maximum likelihood estimators for size-invariant natural parameters of fixed dimension. However, projectivity is not a *necessary* condition for consistency of maximum likelihood estimators: there do exist consistency results for maximum likelihood estimators of non-projective ERGMs. We review them in Section 7.

3.2.4 *Clarification*: *Typically, ERGMs are not applied to large populations, unless there is additional structure*. An implicit assumption, underlying both questions described above, is that ERGMs are applied to large populations without additional structure. However, in most applications to date, ERGMs have been applied to either small populations or large populations with additional structure, as pointed out in Section 3.2.2. Statistical theory should take advantage of additional structure rather than working under the assumption that ERGMs are applied without it. Advances in the development of concentration inequalities (e.g., Talagrand, 1996, Boucheron, Lugosi and Massart, 2013) enable statistical theory to do so. We review an example in Section 7.2.

4. ERGMS WITH ADDITIONAL STRUCTURE

Well-posed ERGMs with additional structure address the lack of structure of simplistic ERGMs and therefore have an important place in the ERGM framework, as pointed out in Section 3.1.3. Here, "additional structure" is understood as additional mathematical structure imposed on:

(a) the support or reference measure of ERGMs;
(b) the dependence structure of ERGMs;
(c) the parameter space of ERGMs;
(d) combinations of (a), (b) and (c).

For example, additional structure may come in the form of constraints on the number of edges or other functions of the random graph; constraints on the parameter space of ERGMs, as imposed by curved ERGMs; or constraints on the dependence structure of ERGMs imposed by block structure, multilevel structure, spatial structure or temporal structure.

We review here some of the more established classes of ERGMs with additional structure. Other classes, not discussed here, are measurement models with an underlying latent graph generated by an ERGM (Wyatt, Choudhury and Bilmes, 2008); ERGMs with restrictions on the support (Karwa, Petrović and Bajić, 2016); nonparametric ERGMs with constraints (Thiemichen and Kauermann, 2017); and ERGMs constraining variances of sufficient statistics (Fellows and Handcock, 2017).

4.1 Dyad-Independent ERGMs

A simple form of additional structure comes in the form of constraints on the dependence structure of ERGMs. In the simplest case, one can assume that all edge variables $Y_{i,j}$ in undirected random graphs are independent or all dyads $(Y_{i,j}, Y_{j,i})$ in directed random graphs are independent.

In the undirected case, assuming edge variables $Y_{i,j}$ are independent, the population probability mass function satisfies the following factorization property:

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) = \prod_{i<j:\, i,j\in\mathcal{N}} \mathbb{P}_{\{i,j\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(Y_{i,j} = y_{i,j}). \tag{1}$$

In the directed case, assuming dyads $(Y_{i,j}, Y_{j,i})$ are independent, the population probability mass function satisfies the following factorization property:

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) = \prod_{i<j:\, i,j\in\mathcal{N}} \mathbb{P}_{\{i,j\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(Y_{i,j} = y_{i,j}, Y_{j,i} = y_{j,i}). \tag{2}$$

Classic examples are Bernoulli random graphs and β-models for undirected random graphs (Gilbert, 1959, Erdős and Rényi 1959, 1960, Chatterjee, Diaconis and Sly, 2011, Rinaldo, Petrović and Fienberg, 2013, Karwa and Slavković, 2016) and p_1-models for directed random graphs (Holland and Leinhardt, 1981, Yan, Zhao and Qin, 2015, Yan, Leng and Zhu, 2016, Yan, Qin and Wang, 2016, Yan et al., 2019). These models can capture heterogeneity in the propensities of nodes to form edges. For example, β-models assume that edge variables $Y_{i,j}$ are independent Bernoulli$(\pi_{i,j})$ random variables with node-dependent edge probabilities $\pi_{i,j} = \text{logit}^{-1}(\theta_i + \theta_j)$ $(\theta_i \in \mathbb{R},\ \theta_j \in \mathbb{R})$, which are equivalent to ERGMs with population probability mass functions of the form

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \propto \prod_{i<j:\, i,j\in\mathcal{N}} \exp\big(\eta_{i,j}(\boldsymbol{\theta},\mathcal{N})\, y_{i,j}\big),$$

where

$$\eta_{i,j}(\boldsymbol{\theta},\mathcal{N}) = \theta_i + \theta_j, \quad i \in \mathcal{N},\ j \in \mathcal{N}.$$

The parameters θ_i and θ_j can be interpreted as the propensities of nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ to form edges, respectively. These propensities can vary from node to node, so the model can capture heterogeneity in the propensities of nodes to form edges. In addition, when the edges in the population graph are directed and the population graph is generated by p_1-models (Holland and Leinhardt, 1981), such models can capture reciprocity. Reciprocity refers to the tendency of nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ to reciprocate edges, that is, $Y_{i,j} = 1$ and $Y_{j,i} = 1$ are observed more frequently than would be expected when the edges $Y_{i,j}$ and $Y_{j,i}$ were independent. Note that reciprocity induces dependence between edge variables $Y_{i,j}$ and $Y_{j,i}$, but leaves dyads $(Y_{i,j}, Y_{j,i})$ independent.

Dyad-independent ERGMs constrain the range of dependence and hence do not have the undesirable properties of the triangle model and other simplistic ERGMs with complex dependence, which are rooted in the lack of structure and the strong dependence induced by those models (as explained in Section 3.1). In addition, the factorization properties of probability mass functions (1) and (2) have computational advantages, facilitating the evaluation of likelihood functions.

4.2 Curved ERGMs

While dyad-independent ERGMs do not have the undesirable properties of the triangle model and other simplistic ERGMs, such models do not capture dependencies among edge variables, other than reciprocity in directed random graphs. To model transitivity and other network phenomena inducing dependence, ERGMs with less restrictive forms of additional structure have to be considered. Curved ERGMs are an important example.

Curved ERGMs were developed by Snijders et al. (2006) and Hunter and Handcock (2006) to address the flaws of simplistic ERGMs lacking structure, such as the triangle model (see, e.g., Snijders et al., 2006, Hunter and Handcock, 2006, Hunter, 2007, Hunter, Goodreau and Handcock, 2008, Goodreau, Kitts and Morris, 2009, Robins, Pattison and Wang, 2009). Curved ERGMs impose additional structure in the form of nonlinear constraints on the natural parameter space of the exponential family (Barndorff-Nielsen, 1978, Brown, 1986) and are curved exponential families in the sense of Efron (1975, 1978). We presented one example of curved ERGMs in Section 3.1.3: curved ERGMs with edge and GWESP terms, which ensure that the added value of additional triangles decreases, in contrast to the triangle model. To describe a large class of curved ERGMs with geometrically weighted model terms, consider ERGMs with population probability mass functions of the form

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \propto \exp\left(\sum_{m=1}^{M} \eta_m(\boldsymbol{\theta},\mathcal{N})\, s_m(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}}) + \cdots\right),$$

where the dots refer to additional model terms, such as edge terms. Here, the sufficient statistics $s_m(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ count the number of units of a specified type with m subgraph configurations of a specified type. The units may refer to nodes; unconnected pairs of nodes; or connected pairs of nodes, as in GWESP terms. The subgraph configurations may refer to shared partners, as in GWESP terms, but many other subgraph configurations are possible. Some examples can be found in Hunter, Goodreau

and Handcock (2008). In addition, the sufficient statistic may depend on the attributes $\boldsymbol{x}_{\mathcal{N}}$ of the population of nodes $\mathcal{N}$. The natural parameters are of the form

$$\eta_m(\boldsymbol{\theta}, \mathcal{N}) = \theta_1 \exp(\theta_2)\big[1 - \big(1 - \exp(-\theta_2)\big)^m\big],$$

where $m = 1, \ldots, M$.

The additional structure imposed by these curved ERGMs comes in the form of nonlinear constraints on the natural parameter space of the exponential family. In the example used in Section 3.1.3, curved ERGMs with edge and GWESP terms, the natural parameter space is $\mathbb{R}^{1+|\mathcal{N}|-2}$, and the curved ERGM imposes nonlinear constraints on $\mathbb{R}^{1+|\mathcal{N}|-2}$. If $\theta_1 > 0$ and $\theta_2 > 0$, these nonlinear constraints ensure that the added value of additional subgraph configurations of the specified type decreases. To see this, consider a single unit, for example, a connected pair of nodes $\{i, j\}$. The log odds of a graph where the unit has m configurations of the specified type relative to a graph where it has $m - 1$ configurations is, assuming everything else is the same, given by

$$\theta_1\big(1 - \exp(-\theta_2)\big)^{m-1},$$

where $m = 1, \ldots, M$. In other words, the added value of additional configurations of the specified type decreases at a geometric rate. Figure 2 in Section 3.1.3 demonstrates how fast the added value of additional triangles decays in the special case of GWESP.

The main ERGM-related statistical software packages, including the 20 ERGM-related `R` packages mentioned in Section 1 and the program `pnet` (Wang, Robins and Pattison, 2006), implement many geometrically weighted model terms and related model terms (e.g., Butts, 2008, Robins, Pattison and Wang, 2009). In practice, curved ERGMs with geometrically weighted terms have been found useful: selected examples are cited in Section 3.2.2. We illustrate in Section 8 that curved ERGMs with edge, GWESP, and other terms can outperform both Bernoulli random graphs and latent space models.

4.3 ERGMs with Block Structure

A simple form of additional structure is block structure, which is popular in the literature on stochastic block models (e.g., Nowicki and Snijders, 2001, Bickel and Chen, 2009, Rohe, Chatterjee and Yu, 2011). In the simplest case, a block structure corresponds to a partition of a population of nodes $\mathcal{N}$ into K subpopulations $\mathcal{A}_1, \ldots, \mathcal{A}_K$, called blocks. While stochastic block models assume that edges within and between blocks are independent conditional on the block structure, ERGMs with block structure allow edges to be dependent within and between blocks. As a result, ERGMs with block structure can be viewed as generalizations of stochastic block models. As in stochastic block models, the block structure may be observed or unobserved.

We present here two classes of ERGMs with block structure: one class of ERGMs that exploits block structure to constrain the range of dependence and another class of ERGMs that exploits block structure to capture unobserved heterogeneity.

4.3.1 *Constraining the range of dependence*. ERGMs that exploit observed block structure to constrain the range of dependence were introduced by Strauss and Ikeda (1990) in the special case of Markov random graphs, using categorical covariates to partition a population of nodes into blocks. A more general class, with observed and unobserved blocks, was developed by Schweinberger and Handcock (2015) and Schweinberger (2020). For simplicity, we focus henceforth on observed block structure.

Given observed block structure, these models are characterized by the following factorization property:

$$\begin{aligned}(3)\qquad &\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})\\ &\quad= \prod_{k=1}^{K}\prod_{l=1}^{k} \mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_l\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_l} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_l}),\end{aligned}$$

where $\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_l}$ denotes the set of edge variables corresponding to possible edges between nodes in subpopulation $\mathcal{N}_k$ and nodes in subpopulation $\mathcal{N}_l$. To constrain the range of dependence to blocks, these models assume that between-block edges are independent,

$$\begin{aligned}&\mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_l\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_l} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_l})\\ &\quad= \prod_{i\in\mathcal{N}_k, j\in\mathcal{N}_l} \mathbb{P}_{\{i,j\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(Y_{i,j} = y_{i,j}),\end{aligned}$$

whereas within-block edges may be dependent:

$$\begin{aligned}&\mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_k\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_k} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})\\ &\quad\neq \prod_{i<j:\, i,j\in\mathcal{N}_k} \mathbb{P}_{\{i,j\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(Y_{i,j} = y_{i,j}).\end{aligned}$$

If ERGMs are used as within- and between-block models, then the population probability model is an ERGM: an ERGM with additional structure in the form of block structure and local dependence within blocks. A concrete example is given by the curved ERGMs with block-dependent edge and GWESP terms described in Section 7.2.

Such ERGMs, exploiting block structure to constrain the range of dependence, address the lack of structure of simplistic ERGMs and have at least three advantages. First, local dependence often makes substantive sense, because many real-world networks are local in nature (e.g., Homans, 1950, Wasserman and Faust, 1994, Pattison and Robins, 2002). Second, these models retain the main advantage of ERGMs, the flexibility to model a wide range of network features, because the within-block ERGMs

can model a wide range of network features within blocks. At the same time, the models address the main disadvantage of simplistic ERGMs, the lack of structure, by using block structure to constrain the range of dependence. As long as the blocks are not too large, the overall dependence induced by the model is weak and the model does not suffer from model near-degeneracy, which results from strong long-range dependence (as discussed in Sections 3.1.2 and 3.1.3). Third, these models have statistical advantages, because weak dependence can be exploited to derive concentration results, which in turn can be used to establish consistency results for likelihood-based estimators. We discuss them in Section 7. Last, but not least, ERGMs with block structure have computational advantages, because the factorization of probability mass function (3) facilitates the computation of likelihood functions.

4.3.2 *Capturing unobserved heterogeneity.* We discuss here two classes of ERGMs that take advantage of block structure to capture unobserved heterogeneity.

The ERGMs of Koskinen (2009) use block-dependent edge terms to capture unobserved heterogeneity in the propensities of nodes to form edges, along with alternating k-triangle terms to capture transitivity; note that alternating k-triangle terms are related to GWESP terms, as explained by Hunter (2007). Wang et al. (2018) used ERGMs with local dependence within blocks (Schweinberger and Handcock, 2015) to capture heterogeneity among communities in terms of the propensities to form edges as well as other network features.

Salter-Townshend and Murphy (2015) developed mixtures of ERGMs for network data that are collected by sampling nodes from a population of nodes (called egos), recording which nodes are connected to egos (called alters), and which of the alters are connected. Salter-Townshend and Murphy assumed that these ego-centric networks were generated by a finite mixture of ERGMs. The resulting mixture of ERGMs can be viewed as a model with block structure, where each ego belongs to one block, each block has a block-dependent ERGM, and the ego-centric networks of all egos in the same block are generated by the same block-dependent ERGM. It is worth noting though that the resulting model is a model of ego-centric networks rather than the population graph, and each possible edge shows up in two ego-centric networks and is hence governed by two block-dependent ERGMs (see Salter-Townshend and Murphy, 2015).

4.4 ERGMs with Multilevel Structure

There is a large and growing body of work on multilevel network data and models (e.g., Lubbers, 2003, Wang et al. 2013, 2016a, Zappa and Lomi, 2015, Lomi, Robins and Tranmer, 2016, Slaughter and Koehly, 2016, Hollway and Koskinen, 2016, Lazega and Snijders, 2016, Brailly et al., 2016, Meredith et al., 2017, Hollway et al., 2017, Gondal, 2018, Stewart et al., 2019).

Multilevel network data are network data with hierarchical structure, in the sense that level-1 units (nodes) are nested within level-2 units (subsets of nodes), which in turn may be nested within level-3 units (subsets of subsets of nodes), and so forth. A simple form of multilevel structure is observed block structure, as described in Section 4.3, where nodes correspond to level-1 units and blocks correspond to level-2 units. More general forms of multilevel structure exist: for example, in universities, faculty members (level-1 units) are nested within departments (level-2 units); departments are nested within schools (level-3 units); and schools are nested within universities (level-4 units). It is worth noting that multilevel structure is observed: for instance, it can be observed which faculty member belongs to which department, and which department belongs to which school.

To demonstrate multilevel ERGMs, consider two-level networks, with nodes (level-1 units) nested within subsets of nodes $\mathcal{N}_1, \ldots, \mathcal{N}_K$ (level-2 units). An example of two-level networks is given by the human brain networks used in Section 8, where the level-1 units correspond to 56 regions of the brain and the level-2 units correspond to 108 brains. A simple form of two-level ERGM assumes that

$$\begin{aligned}&\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})\\ &\quad= \prod_{k=1}^{K} \mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_k\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_k} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}).\end{aligned} \tag{4}$$

An ERGM with two-level structure (4) is a special case of an ERGM with block structure (3) when edges between subpopulations do not exist with probability 1. A specific example of a two-level ERGM with edge, GWESP, and other terms can be found in Section 8. Well-posed ERGMs with multilevel structure share the advantages of ERGMs with block structure, as discussed in Section 4.3.

4.5 ERGMs with Spatial Structure

In some applications of ERGMs, the population of nodes is embedded in a space. The space in question may be a geographical space (Butts and Acton, 2011) or a social space (McPherson, 1983), constructed from observed attributes of population members (e.g., race). Spatial structure can be exploited to construct more realistic and better-behaved ERGMs. For instance, researchers may hypothesize that the probability of an edge decreases as the distance between population members increases. To test such hypotheses and estimate the strength of the effect of distance on the population graph, ERGMs can be used, with model terms that are functions of the distances between population members. Such ERGMs can induce sparsity by penalizing edges between pairs of nodes that are separated by large distances, and can help control the

dependence of edge variables, resulting in better-behaved ERGMs.

To give a simple example, assume that the spatial locations of nodes are observed, for example, the positions are based on geographical or other observed attributes of nodes. We do not consider unobserved spatial structure, but note that ERGMs with unobserved spatial structure can be viewed as generalizations of latent space models (Hoff, Raftery and Handcock, 2002). A simple example of an ERGM with observed spatial structure is an ERGM with population probability mass function

$$\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N},d)}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \propto \exp\left(\sum_{i<j:\, i,j\in\mathcal{N}} \eta_{i,j}(\boldsymbol{\theta},\mathcal{N},d)\, y_{i,j}\right),$$

where

$$\eta_{i,j}(\boldsymbol{\theta},\mathcal{N},d) = \theta_1 - \theta_2\, f\big(d(i,j)\big).$$

Here, $d : \mathcal{N}\times\mathcal{N} \mapsto \mathbb{R}_0^+$ is a distance function and $f : \mathbb{R}_0^+ \mapsto \mathbb{R}_0^+$ is a function of distance, where $\mathbb{R}_0^+ = \mathbb{R}^+ \cup \{0\}$ denotes the set of positive real numbers $\mathbb{R}^+$ and 0. The function f specifies how the distance affects the log odds of the probability of an edge:

$$\log \frac{\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N},d)}(Y_{i,j}=1)}{1-\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N},d)}(Y_{i,j}=1)} = \eta_{i,j}(\boldsymbol{\theta},\mathcal{N},d). \tag{5}$$

The function $f : \mathbb{R}_0^+ \mapsto \mathbb{R}_0^+$ can take many forms, for example, if $f(d(i,j)) = d(i,j)$ and $d(i,j)$ is the Euclidean distance between the positions of $i \in \mathcal{N}$ and $j \in \mathcal{N}$ in $\mathbb{R}^d$, the model is equivalent to the latent space model of Hoff, Raftery and Handcock (2002) with $\theta_2 = 1$ and observed distances. But other choices of $f : \mathbb{R}_0^+ \mapsto \mathbb{R}_0^+$ are possible, allowing the log odds to decay slower or faster as a function of distance. Butts and Acton (2011) showed that the rate of decay can have a considerable impact on the structure of the population graph, so care must be taken when specifying ERGMs with spatial structure. In addition, ERGMs can contain more model terms depending on distance, although the log odds of the probability of an edge (5) needs to be replaced by the log odds of the conditional probability of an edge when the model terms induce dependence among edge variables.

ERGMs with spatial structure have at least three advantages. First, ERGMs with spatial structure offer many opportunities for testing and modeling the impact of spatial structure on the population graph. Second, ERGMs assuming that the probability of an edge decreases as a function of distance can induce sparsity by penalizing edges between pairs of nodes that are separated by large distances. Last, but not least, forcing the dependence of edge variables to decay as a function of distance can help control dependence, resulting in better-behaved ERGMs. For instance, Butts (2011) showed that even a local triangle term based on triangles within a specified radius can be well-behaved, provided that the radius is not too large relative to the population density. The same idea can be applied to GWESP terms and other geometrically weighted model terms in curved ERGMs, which are expected to behave even better than local triangle terms.

4.6 ERGMs with Temporal Structure

Many networks change over time, and temporal structure can help construct well-behaved ERGMs: modeling a sequence of small changes of a large population graph may be easier than modeling the whole population graph at once, because edges are interdependent. Hanneke, Fu and Xing (2010) introduced discrete-time Markov models to do so, with transition probabilities parameterized by ERGMs. Krivitsky and Handcock (2014) elaborated a separable parameterization, separating the edge formation and dissolution process. We do not discuss such models here, because ERGMs with temporal structure are a complex class of models that deserves a separate treatment elsewhere, and because many of the associated statistical issues are special cases of classes treated elsewhere in our paper. We refer to the cited literature for details, and to Robins and Pattison (2001) and Ouzienko, Guo and Obradovic (2011) for earlier work on temporal ERGMs.

4.7 ERGMs with Non-Random and Random Attributes

In addition to edges, nodes may have attributes, which may be non-random or random, observed or unobserved. Attributes may not impose much structure, but help capture heterogeneity, observed or unobserved, in the propensities of nodes to form edges and other subgraph configurations.

We first discuss ERGMs with observed non-random and random attributes and then ERGMs with unobserved, random attributes.

4.7.1 *Observed non-random and random attributes.* The attributes of nodes may either be exogenous, non-random (e.g., race) or endogenous, random (e.g., political preference).

Incorporating non-random attributes as predictors of edges makes sense, is straightforward and has a long tradition in the ERGM literature. Some examples are provided by Morris, Handcock and Hunter (2008) and Hunter, Goodreau and Handcock (2008): for example, homophily or similarity with respect to categorical attributes of nodes can be captured by including sufficient statistics of the form $\sum_{i<j:\, i,j\in\mathcal{N}} I(x_i = x_j)\, y_{i,j}$, where x_i and x_j are categorical attributes of nodes i and j, respectively, and $I(x_i = x_j) = 1$ if $x_i = x_j$ and $I(x_i = x_j) = 0$ otherwise.

A special case of interest is when the attributes are random, governed by a joint probability model for both the

random attributes and the random graph. Models for both random attributes and random graphs were explored by Fellows and Handcock (2012) in the exponential-family framework. The resulting models are complex, and we refer to Fellows and Handcock (2012) for details.

4.7.2 *Unobserved random attributes*. We distinguish ERGMs with discrete and continuous unobserved random attributes.

The discrete case was considered by Koskinen (2009), Schweinberger and Handcock (2015) and Wang et al. (2018), as discussed in Section 4.3. The continuous case was considered by Thiemichen et al. (2016), who developed a class of ERGMs with random effects. In the special case of dyad-independent ERGMs, there has been a long tradition of using random effects models, dating back to the p_2-models of van Duijn (1995) and van Duijn, Snijders and Zijlstra (2004) and the related models of Gill and Swartz (2004), which are random effects versions of the p_1-models of Holland and Leinhardt (1981) and related to the random effects and mixed effects models of Hoff (2003, 2005). Thiemichen et al. (2016) considered more general ERGMs with the number of edges $f_i(\boldsymbol{y}_{\mathcal{N}}) = \sum_{j\neq i: j\in\mathcal{N}} y_{i,j}$ of nodes $i \in \mathcal{N}$ as sufficient statistics, along with other sufficient statistics. Thiemichen et al. assumed that the weights $\eta_i(\boldsymbol{\theta}, \mathcal{N})$ of the sufficient statistics $f_i(\boldsymbol{y}_{\mathcal{N}})$ are random coefficients of the form $\eta_i(\boldsymbol{\theta}, \mathcal{N}) = \theta_1 + \epsilon_i$, where $\theta_1 \in \mathbb{R}$ can be interpreted as the overall propensity to form edges in the population, $\epsilon_i \in \mathbb{R}$ can be interpreted as the deviation of node $i \in \mathcal{N}$ from the overall propensity and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ $(\sigma^2 > 0)$. The resulting random effects ERGMs can capture unobserved heterogeneity in the propensities of nodes to form edges, along with other network features.

5. COMPLETE- AND INCOMPLETE-DATA GENERATING PROCESS

The discussion of the likelihood function in Section 3.2.2 demonstrates that likelihood-based inference requires proper statistical language to clarify the core statistical notions of "sample" and "population" in the ERGM framework, and to separate the process that generates the population graph from the observation process.

We follow here the principled approach of Rubin (1976) and distinguish the complete-data generating process (generating the population graph) from the incomplete-data generating process (the observation process). A failure to take both of these processes into account can lead to misleading statistical conclusions, as discussed by Rubin (1976), Dawid and Dickey (1977), Thompson and Frank (2000), Gile and Handcock (2006, 2017), Handcock and Gile (2010), Koskinen, Robins and Pattison (2010), Crane (2018) and Crane and Dempsey (2020). We discuss complete- and incomplete-data generating processes in Sections 5.1 and 5.2, respectively.

The specification of the complete-data generating process serves at least two additional purposes. First, the parameters of the complete-data generating process constitute the natural target of statistical inference. Second, the population graph or super population of population graphs generated by the complete-data generating process is the population or super population to which statistical inferences generalize.

As a consequence, the specification of the complete-data generating process is coupled with the goal of statistical inference. We distinguish three broad goals of statistical inference: finite, super and infinite population inference. These notions are inspired by the corresponding notions in classical statistics (e.g., Hartley and Sielken, 1975). We adapt them here to the statistical analysis of network data.

DEFINITION. *Finite population inference* is concerned with a finite population of nodes $\mathcal{N}$ and a fixed population graph $\boldsymbol{y}_{\mathcal{N}}$ defined on $\mathcal{N}$. It does not assume that the population graph was generated by a population probability model. The goal is to estimate functions of the population graph (e.g., the total number of edges in the population graph or the presence or absence of specific edges).

DEFINITION. *Super population inference* is concerned with a finite population of nodes $\mathcal{N}$ and a population graph $\boldsymbol{y}_{\mathcal{N}}$ defined on $\mathcal{N}$. In contrast to finite population inference, it assumes that the population graph was generated by a population probability model. The goal is to estimate the population probability model governing the super population of possible population graphs.

DEFINITION. *Infinite population inference* is concerned with an infinite population of nodes $\mathcal{N}$ and a population graph $\boldsymbol{y}_{\mathcal{N}}$ defined on $\mathcal{N}$ generated by a population probability model. The goal of statistical inference is to estimate the population probability model based on a subgraph $\boldsymbol{y}_{\mathcal{N}}$ induced by a subset of nodes $\mathcal{N}' \subset \mathcal{N}$.

5.1 Complete-Data Generating Process

The complete-data generating process is the process that generates the complete data, that is, the population graph of interest.

It is possible to make no assumptions about the complete-data generating process, leading to finite population inference (Section 5.1.1). If the process that generates the population graph is of substantive interest, one may specify a super population of possible population graphs along with a population probability model that generates population graphs. The specification of the super population of population graphs may assume that the sizes of graphs are either fixed or limited to a finite range of possible sizes, leading to super population inference for population probability models of graphs of the same size or similar sizes (Section 5.1.2). An alternative is to make

assumptions about how the model behaves as the size and composition of the set of nodes $\mathcal{N}$ changes, leading to infinite population inference on models of sequences of graphs of increasing size (Section 5.1.3). We discuss these cases in turn.

5.1.1 *Finite graphs*: *Finite population inference*. In some applications, it is neither necessary nor desirable to make assumptions about the complete-data generating process. For example, consider the network of sexual relationships between HIV-infected residents and non-infected residents of New York City (NYC) during a specified period of time, where the goal is to estimate the number of sexual contacts between HIV-infected and non-infected residents. The population of interest $\mathcal{N}$ consists of the residents of NYC and the population graph $\boldsymbol{y}_{\mathcal{N}}$ consists of sexual relationships between residents of NYC. If the whole population graph $\boldsymbol{y}_{\mathcal{N}}$ is observed, the population graph can be used to answer the question of interest by counting the number of sexual relationships between HIV-infected and non-infected residents. If it is not possible to observe the whole population graph $\boldsymbol{y}_{\mathcal{N}}$ but a sample of sexual relationships is generated (as discussed in Section 5.2), then the sample can be used to construct an estimator of the number of sexual relationships between HIV-infected and non-infected residents. But, regardless of whether the whole population graph $\boldsymbol{y}_{\mathcal{N}}$ is observed, answering the question of interest does not require any assumption about the complete-data generating process. In such situations, finite population inference is all that is needed to answer the question of interest.

Target of statistical inference. In finite population inference, any function of the population graph $\boldsymbol{y}_{\mathcal{N}}$ is a legitimate target of statistical inference: for example, in the sexual network example described above, researchers may be interested in estimating the number of sexual relationships between HIV-infected and non-infected residents of NYC. Here, model-based inference may neither be necessary nor desirable and design-based inference is all that is needed (Kurant et al., 2012, Gjoka, Smith and Butts 2014, 2015).

A special case where model-based inference based on ERGMs is useful for finite population inference was considered by Krivitsky and Morris (2017). Krivitsky and Morris (2017) used ego-centric sampling to estimate population-level network features of interest, then used the estimated population-level network features as sufficient statistics of an ERGM to simulate graphs that are similar to the estimated population-level network features. To elaborate, define

$$\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}}) = \underset{\boldsymbol{\theta}' \in \boldsymbol{\Theta}}{\arg\max} \big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}', \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}}) \rangle - \psi(\boldsymbol{\theta}', \mathcal{N}) \big),$$

and note that the maximizer $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ exists and is unique as long as $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ falls into the relative interior of the convex hull of the set $\{\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}}) : \boldsymbol{y}_{\mathcal{N}} \in \mathcal{Y}_{\mathcal{N}}\}$ (Barndorff-Nielsen, 1978, p. 151). The maximizer $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is a function of the attributes of population members $\boldsymbol{x}_{\mathcal{N}}$ and the population graph $\boldsymbol{y}_{\mathcal{N}}$ and is hence a legitimate target of finite population inference. We note that the maximizer $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is equivalent to the maximum likelihood estimate, but $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is not random, because neither $\boldsymbol{x}_{\mathcal{N}}$ nor $\boldsymbol{y}_{\mathcal{N}}$ are random. In fact, if the whole population graph $\boldsymbol{y}_{\mathcal{N}}$ is observed, then the maximizer can in principle be computed without error, though in practice one may have to approximate the maximizer by using Monte Carlo maximum likelihood estimates as described by Krivitsky and Morris (2017). The function $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is of interest, because it can be used to simulate graphs that are similar to the population graph: by well-known exponential-family properties (Brown, 1986, Theorem 5.5, p. 148), the expected sufficient statistic $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{Y}_{\mathcal{N}})$ matches the sufficient statistic $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ of the population $\mathcal{N}$ under $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$. Thus, graphs simulated from the ERGM with parameter $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ will have sufficient statistics that are similar to the population graph in terms of the sufficient statistic $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$.

An example of a situation in which one may wish to simulate similar graphs are data privacy settings involving network data. In other words, researchers may wish to share network data with others, while protecting the privacy of population members. To do so, researchers can simulate a graph that is similar—but not identical—to the population graph $\boldsymbol{y}_{\mathcal{N}}$ in terms of network features $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ (Fienberg and Slavkovic, 2010, Karwa, Krivitsky and Slavković, 2017). Goodreau et al. (2008) used the described procedure to create synthetic school networks based on the National Longitudinal Study of Adolescent Health, some of which are included in `R` package `ergm` (Hunter et al., 2008). These networks are used for educational purposes, such as tutorials and workshops, and software testing. In such applications, it is useful to have network data with realistic structure, but it is not essential to have exact replications of the original network data.

Last, but not least, an example that combines both motivations is model-based imputation, where one seeks to impute the states of unobserved edge variables in a fixed population graph. Model-based imputation (Gile and Handcock, 2006, Handcock and Gile, 2010, Koskinen, Robins and Pattison, 2010) can be performed by estimating an ERGM from an incomplete observation of the population graph—using the likelihood function $\mathcal{L}(\boldsymbol{\theta};\ \boldsymbol{y}_{\mathcal{N}'})$ described in Section 3.2.2—and then simulating draws from the estimated model conditional on the observed network data. Note that here the model is employed to (a) leverage information from observed edge variables to predict the states of unobserved edge variables and (2) ensures that the imputed population graphs have properties that are compatible with the observed data (as discussed above). The ERGM used to make the model-based imputations need not be the data-generating model of the

population graph, as long as it helps impute the states of unobserved edge variables. In such settings, it is natural to assess model performance via prediction of held-out data, as proposed by Wang et al. (2016b).

5.1.2 *Finite graphs*: *Super population inference*. While in some applications it may neither be necessary nor desirable to make assumptions about the complete-data generating process, in other applications the complete-data generating process is of substantive interest. For example, neuroscientists may be interested in the probability law that governs connections between regions of the human brain (e.g., Simpson, Moussa and Laurienti, 2012, Sinke et al., 2016, Obando and De Vico Fallani, 2017). Here, interest centers on a population probability model that generates finite graphs of the same size or similar sizes. The neuroscience application in Section 8 serves as an example: the population of interest consists of 56 regions of the human brain, and the goal of statistical inference is to infer the probability law that governs connections between these 56 regions based on 108 brain networks (i.e., 108 replications).

Target of statistical inference. In super population inference, the target of statistical inference is the parameter $\boldsymbol{\theta}$ of the population probability model that generated the population graph and governs the super population of all possible population graphs of the same size or a finite range of possible sizes. We note that even when the whole population graph is observed, uncertainty arises from the fact that the parameter $\boldsymbol{\theta}$ is unknown.

5.1.3 *Sequences of graphs of increasing size*: *Infinite population inference*. In both statistical practice and theory, it is sometimes convenient to consider sequences of graphs of increasing size. In many such situations, there is an explicit or implicit assumption that there exists a graph limit—that is, an infinite population graph defined on an infinite population of nodes—to which sequences of graphs converge (Lovász, 2012). We therefore refer to statistical inference based on sequences of graphs of increasing size as infinite population inference, despite the fact that researchers in practice may be more interested in subsequences of graphs of finite sizes rather than the graph limit itself.

In statistical practice, sequences of graphs of increasing size may be meaningful when, for example, one observes two or more graphs of different sizes and wishes to formulate a model that is invariant in a well-defined sense. Consider residents of NYC and Seattle, where two residents are connected by an edge if the residents meet at least twice a month to work out together. While NYC has more than 10 times as many residents as Seattle, it is not credible that the expected number of workout partners of NYC residents is more than 10 times higher than the expected number of workout partners of Seattle residents, as dense Bernoulli(π) random graphs assume. In such situations, it is convenient to formulate a model of sequences of graphs of increasing size such that the expected number of edges of each node is invariant to network size and consider the two observed graphs—the large NYC exercise network and the small Seattle exercise network—as two observations taken from a sequence of graphs generated by the model. ERGMs that respect such desiderata have been developed by Krivitsky, Handcock and Morris (2011), Krivitsky and Kolaczyk (2015), and Butts and Almquist (2015).

In statistical theory, it is convenient to embed observed data (e.g., an observed graph) into a sequence of data sets of increasing size (e.g., a sequence of graphs of increasing size), which is a classic approach in statistical theory: for example, Lehmann (1999) suggested

> *"to embed the actual situation in a sequence of situations, the limit of which serves as the desired approximation"* (Lehmann, 1999, p. 1).

Sequences of graphs of increasing size can be constructed in many ways: for example, graphs can grow by adding nodes or subsets of nodes along with edges. To cover a wide range of sequences of graphs of increasing size, including cumulative and non-cumulative sequences, let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of sets of nodes and $\mathcal{N}_1, \mathcal{N}_2, \ldots$ be a sequence of sets of nodes satisfying $\mathcal{N}_k \subseteq \bigcup_{l=1}^{k} \mathcal{A}_l$. Suppose that the sequence of random graphs $\boldsymbol{Y}_{\mathcal{N}_1}, \boldsymbol{Y}_{\mathcal{N}_2}, \ldots$ is generated by a sequence of models of the form $\mathbb{P}_{\mathcal{N}_1,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_1)}, \mathbb{P}_{\mathcal{N}_2,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_2)}, \ldots$, where the natural parameter $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}_k)$ may depend on the set of nodes $\mathcal{N}_k$ and the dimension of parameter $\boldsymbol{\theta}$ may grow with the size $|\mathcal{N}_k|$ of $\mathcal{N}_k$. Then the generating processes can be described by a sequence of the form

$$\begin{aligned}&\left(\mathcal{N}_1, \boldsymbol{x}_{\mathcal{N}_1}, \boldsymbol{Y}_{\mathcal{N}_1}, \mathbb{P}_{\mathcal{N}_1,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_1)}\right),\\ &\left(\mathcal{N}_2, \boldsymbol{x}_{\mathcal{N}_2}, \boldsymbol{Y}_{\mathcal{N}_2}, \mathbb{P}_{\mathcal{N}_2,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_2)}\right), \ \ldots\end{aligned}$$

Target of statistical inference. In infinite population inference, the target of statistical inference is the parameter $\boldsymbol{\theta}$ of the population probability model; note that $\boldsymbol{\theta}$ may not be the natural parameter of the exponential family and the dimension of $\boldsymbol{\theta}$ may depend on the number of nodes, as the β-models in Section 4.1 demonstrates.

5.2 Incomplete-Data Generating Process

The incomplete-data generating process is the process that, conditional on the population graph generated by the complete-data generating process, determines which subgraphs of the population graph are observed. In the best-case scenario, the whole population graph is observed, but in more common scenarios, some of the edges in the population graph are unobserved. The two most common reasons for incomplete data are sampling and

missing data. We discuss selected incomplete-data generating processes, with an emphasis on sampling designs (Sections 5.2.1, 5.2.2 and 5.2.3) and missing data (Section 5.2.4). We conclude with some comments on the fundamental concept of ignorability of incomplete-data generating processes for the purpose of likelihood-based super and infinite population inference (Section 5.2.5).

5.2.1 *Sampling nodes*: *Ego-centric sampling and link-tracing*. If a population of nodes $\mathcal{N}$ is large, it may not be possible to observe the whole population graph. A popular solution is to sample edges by using ego-centric sampling (Krivitsky and Morris, 2017) or link-tracing (Thompson and Frank, 2000, Gile and Handcock, 2006, Handcock and Gile, 2010). Both sample a subset of nodes $\mathcal{N}' \subseteq \mathcal{N}$ and record edges from nodes in $\mathcal{N}'$ to nodes in $\mathcal{N}$.

An ego-centric sampling design generates a sample of nodes along with edges as follows (Krivitsky and Morris, 2017):

1. Generate a probability sample of nodes, called egos.
2. For each sampled ego, record edges to connected nodes, called alters.

A probability sample of nodes can be generated by any sampling design for sampling from finite populations (e.g., Thompson, 2012).

A number of variations of ego-centric sampling designs are possible. First, some ego-centric sampling designs identify alters, so that it is known whether two egos nominated the same alter. Second, other ego-centric sampling designs ask egos to report which pairs of alters have edges (Smith et al., 1972–2016). Third, an important extension of ego-centric sampling is link-tracing. Link-tracing exploits the observed edges of sampled nodes to include additional nodes into the sample, provided that the identities of the egos and alters of sampled nodes are known. One specific form of k-wave link-tracing samples nodes and edges as follows (Thompson and Frank, 2000):

1. Wave $l = 0$: Generate an ego-centric sample.
2. Wave $l = 1, \ldots, k$:
   (a) Add the nodes who are linked to the population members of wave $l - 1$ to the sample.
   (b) For each added node, record edges.

Ego-centric sampling can be considered to be a special case of k-wave link-tracing with $k = 0$. Additional examples of link-tracing are snowball sampling (Goodman, 1961) and respondent-driven sampling (Heckathorn, 1997, Salganik and Heckathorn, 2004, Gile and Handcock, 2010, Gile, 2011). Some link-tracing sampling designs, such as respondent-driven sampling, may not generate probability samples, but approximate probability samples when suitable sampling designs are used (Kurant, Markopoulou and Thiran, 2011, Gile, 2011).

5.2.2 *Sampling pairs of nodes*: *Edge sampling*. While ego-centric sampling and link-tracing sample edges indirectly by first sampling nodes and then recording edges of sampled nodes, one can sample edges directly. One example is a sampling design that samples spouses from a sampling frame of married couples, that is, which samples pairs of nodes connected by an edge (here, marriage). A theoretical treatment of edge sampling can be found in Crane and Dempsey (2018, 2020) and Crane (2018).

5.2.3 *Sampling subgraphs*. An alternative approach is based on sampling a subset of nodes $\mathcal{N}' \subseteq \mathcal{N}$ and collecting information about the whole subgraph $\boldsymbol{y}_{\mathcal{N}'}$ of $\boldsymbol{y}_{\mathcal{N}}$ induced by $\mathcal{N}' \subseteq \mathcal{N}$. Sampling subgraphs is distinct from ego-centric sampling and link-tracing, because subgraph sampling collects information about all edges among nodes in $\mathcal{N}'$ but does not collect information about edges between nodes in $\mathcal{N}'$ and nodes in $\mathcal{N} \setminus \mathcal{N}'$, which ego-centric sampling and link-tracing do. The most widely used form of subgraph sampling is multilevel sampling (Snijders and Bosker, 2012, Lazega and Snijders, 2016). Consider a population of nodes $\mathcal{N}$ partitioned into subpopulations $\mathcal{A}_1, \ldots, \mathcal{A}_K$. Suppose that a subset of subpopulations $\mathcal{S} \subseteq \{1, \ldots, K\}$ is sampled and that the subgraphs $\boldsymbol{y}_{\mathcal{A}_k}$ induced by the sampled subpopulations $\mathcal{A}_k$ with $k \in \mathcal{S}$ are observed. A simple example of a multilevel sample is a sample of school classes from a population of school classes, generated by any sampling design for sampling from finite populations (e.g., Thompson, 2012). If all students in the sampled school classes are asked to report edges to other students in the same school class, the subgraphs induced by the sampled school classes are observed.

5.2.4 *Missing data*. In addition to design-based missingness due to sampling, there may be out-of-design missingness due to, among other things, non-response of respondents in network surveys (Gile and Handcock, 2006, Handcock and Gile, 2010, Koskinen, Robins and Pattison, 2010). Out-of-design missingness is not under the control of researchers, but is ignorable for the purpose of likelihood-based super and infinite population inference under certain conditions, detailed in Section 5.2.5.

5.2.5 *Ignorable incomplete-data generating processes*. An important concept in likelihood-based super and infinite population inference given incomplete data is the notion of ignorability due to Rubin (1976). An incomplete-data generating process is ignorable for the purpose of estimating the parameters of the population probability model provided:

(a) the probability of not observing the value of an edge variable $Y_{i,j}$ does not depend on the value of $Y_{i,j}$;

(b) the parameters of the complete- and incomplete-data generating process are variation-independent (Gile and Handcock, 2006, Handcock and Gile, 2010, Koskinen, Robins and Pattison, 2010).

A more formal description of ignorable incomplete-data generating processes can be found in Section 6.

Examples of ignorable incomplete-data generating processes include ego-centric sampling and link-tracing, edge sampling, subgraph sampling and data missing at random, but exclude respondent-driven sampling (Lunagomez and Airoldi, 2014). We refer to Gile and Handcock (2006), Handcock and Gile (2010), and Koskinen, Robins and Pattison (2010) for likelihood-based inference with ignorable incomplete-data generating processes and Lunagomez and Airoldi (2014) for likelihood-based inference with non-ignorable incomplete-data generating processes. We discuss likelihood-based inference given incomplete data, generated by ignorable incomplete-data generating processes, in Section 6.

6. LIKELIHOOD-BASED INFERENCE GIVEN INCOMPLETE DATA GENERATED BY IGNORABLE INCOMPLETE-DATA PROCESSES

We describe likelihood-based inference for well-posed ERGMs, based on incomplete data generated by ignorable incomplete-data processes. We focus here on the maximum likelihood approach of Handcock and Gile (2010), and note that Koskinen, Robins and Pattison (2010) describe a Bayesian approach. Both of them are based on the principled approach of Rubin (1976) to likelihood-based inference in incomplete-data scenarios.

To describe the likelihood-based approach of Handcock and Gile (2010), denote by $\boldsymbol{A}$ the $|\mathcal{N}| \times |\mathcal{N}|$-matrix with elements $A_{i,j} \in \{0, 1\}$, where $A_{i,j} = 1$ if the value $y_{i,j}$ of $Y_{i,j}$ is observed and $A_{i,j} = 0$ otherwise ($i < j,\ i, j \in \mathcal{N}$; elements on the main diagonal of $\boldsymbol{A}$, and below the main diagonal, are undefined). The matrix $\boldsymbol{A}$ can deal with all forms of incomplete observations of the population graph, whether data are unobserved due to node sampling, edge sampling, subgraph sampling, missing data or any combination of the aforementioned incomplete-data generating processes. Let $\mathcal{S} = \{i < j : i, j \in \mathcal{N}, A_{i,j} = 1\}$ be the set of pairs of nodes with observed data and $\boldsymbol{y}_{\mathcal{S}} = \{y_{i,j} : i < j,\ i, j \in \mathcal{S}\}$ be the observed data. The incomplete-data generating process is called ignorable for the purpose of likelihood-based inference for the parameter $\boldsymbol{\theta}$ of the complete-data generating process provided:

(a) $\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) = \mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{Y}_{\mathcal{S}} = \boldsymbol{y}_{\mathcal{S}})$, where $\boldsymbol{\alpha}$ is the parameter of the incomplete-data generating process (e.g., the elements of $\boldsymbol{\alpha}$ may be sample inclusion probabilities);

(b) the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ of the complete- and incomplete-data generating process are variation-independent in the sense that the parameter space is a product space.

In other words, the incomplete-data generating process is ignorable as long as the probability of being unobserved does not depend on the nature of the unobserved data. Handcock and Gile (2010) demonstrated that many sampling designs are ignorable, including ego-centric and link-tracing sampling designs.

If the incomplete-data generating process is ignorable, the likelihood function factorizes as follows:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{S}}) \\
&\quad \propto \sum_{\boldsymbol{y}_{\mathcal{N}} \in \mathcal{Y}_{\mathcal{N}}(\boldsymbol{y}_{\mathcal{S}})} \mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{Y}_{\mathcal{S}} = \boldsymbol{y}_{\mathcal{S}})\, \mathbb{P}_{\mathcal{N}, \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) \\
&\quad \propto \underbrace{\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{Y}_{\mathcal{S}} = \boldsymbol{y}_{\mathcal{S}})}_{} \times \underbrace{\sum_{\boldsymbol{y}_{\mathcal{N}} \in \mathcal{Y}_{\mathcal{N}}(\boldsymbol{y}_{\mathcal{S}})} \mathbb{P}_{\mathcal{N}, \boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})}_{} \\
&\quad \propto \qquad \mathcal{L}(\boldsymbol{\alpha}; \boldsymbol{y}_{\mathcal{S}}) \qquad \times \qquad \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{S}}),
\end{aligned}
$$

where $\mathcal{Y}_{\mathcal{N}}(\boldsymbol{y}_{\mathcal{S}})$ is the subset of graphs $\boldsymbol{y}_{\mathcal{N}} \in \mathcal{Y}_{\mathcal{N}}$ compatible with the observed data $\boldsymbol{y}_{\mathcal{S}}$.

As a consequence, as long as the incomplete-data generating process is ignorable, likelihood-based inference for

- the parameter $\boldsymbol{\alpha}$ can be based on the likelihood function $\mathcal{L}(\boldsymbol{\alpha}; \boldsymbol{y}_{\mathcal{S}})$;
- the parameter $\boldsymbol{\theta}$ can be based on the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{S}})$.

As pointed out in Section 3.2.2, the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{S}})$ is based on marginalizations of the population probability mass function, regardless of whether the model is projective. Therefore, the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathcal{S}})$ is not affected by lack of projectivity. In other words, statistical inference that respects both the complete- and incomplete-data generating process and is based on the likelihood function is not affected by lack of projectivity.

Computational methods for likelihood-based inference given incomplete data are described by Handcock and Gile (2010) and Koskinen, Robins and Pattison (2010). Other work on statistical inference given incomplete data can be found in Snijders (2010), Pattison et al. (2013), Krivitsky and Morris (2017), Karwa, Krivitsky and Slavković (2017) and Gile and Handcock (2017).

7. CONSISTENCY AND ASYMPTOTIC NORMALITY OF ESTIMATORS

We review consistency and asymptotic normality results for likelihood-based estimators of well-posed ERGMs in finite, super and infinite population scenarios. These results demonstrate that likelihood-based inference for well-posed ERGMs is possible. We do not consider statistical inference for ill-posed ERGMs, because inferring models which are known to be ill-posed and which are not believed to have generated observed network data is not meaningful.

7.1 Finite Population Inference

Finite population inference focuses on functions of the population graph, such as the number of edges in the population graph, and does not assume that the population graph was generated by a population probability model. If the whole population graph is observed, there is no uncertainty. However, when a sample from the population graph is generated—as described in Section 5.2—there is uncertainty due to the unobserved edges in the population graph. In such situations, two forms of consistency are available for estimators of population quantities based on sample quantities: Fisher-consistency and consistency and asymptotic normality under sampling.

First, many estimators of population quantities are Fisher-consistent (Fisher, 1922). In other words, when the whole population graph is observed, the estimator of the population quantity of interest is equal to the population quantity. An example is an estimator of the proportion of edges in the population graph based on the proportion of edges in a sample.

Second, it is often possible to write functions of the population graph of interest in terms of weighted population totals. In such settings, one can construct classical Horvitz–Thompson estimators for the weighted population total of interest, whose properties follow from the sampling design and often include consistency and asymptotic normality under sampling (Gjoka, Smith and Butts, 2015).

Last, but not least, consider the following function of the attributes of the population of nodes, $\boldsymbol{x}_{\mathcal{N}}$, and the population graph, $\boldsymbol{y}_{\mathcal{N}}$, which we reviewed in Section 5.1.1:

$$\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}}) = \arg\max_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}', \mathcal{N}), \boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}}) \rangle - \psi(\boldsymbol{\theta}', \mathcal{N}) \big).$$

Consider the case where the population quantity $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is unknown, which implies that $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is unknown. If an ego-centric sample is generated—as described in Section 5.2.1—and $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is estimated from the ego-centric sample, then the resulting estimator of $\boldsymbol{\theta}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ is consistent and asymptotically normal, provided $\boldsymbol{s}(\boldsymbol{x}_{\mathcal{N}}, \boldsymbol{y}_{\mathcal{N}})$ can be reconstructed from ego-centric observations of all members of the population $\mathcal{N}$ (Krivitsky and Morris, 2017).

7.2 Super Population Inference

Super population inference is concerned with a finite population of nodes $\mathcal{N}$ and a population graph defined on $\mathcal{N}$, generated by a population probability model. We review here concentration and consistency results for likelihood-based estimators of well-posed ERGMs with block structure in super population scenarios. We consider both complete-data scenarios, where the whole population graph is observed, and incomplete-data scenarios, where subgraphs are sampled by ignorable sampling designs. These concentration and consistency results respect the fact that the population of nodes is finite and cover all finite populations with $K \geq 2$ blocks, although the results are most interesting when K is large.

We assume that the population graph was generated by an ERGM with observed block structure, as described in Section 4.3, consisting of within-block ERGMs with block-dependent edge and GWESP terms and between-block ERGMs with block-dependent edge terms. In other words, the sufficient statistics of the within-block ERGMs count the number of edges and the number of connected pairs of nodes with $1, \ldots, |\mathcal{A}_k| - 2$ shared partners in block $\mathcal{A}_k$, and the natural parameters of the within-block ERGMs are

$$\eta_{k,1}(\boldsymbol{\theta}, \mathcal{N}) = \theta_1,$$
$$\eta_{k,1+m}(\boldsymbol{\theta}, \mathcal{N}) = \exp(\vartheta)\big[1 - \big(1 - \exp(-\vartheta)\big)^m\big],$$

where $m = 1, \ldots, |\mathcal{A}_k| - 2$ and $k = 1, \ldots, K$. Here, $\theta_2 = \exp(\vartheta) \in (0, 1)$, so $\boldsymbol{\Theta} = \mathbb{R} \times (0, 1)$. We assume henceforth that parameters of the within- and between-block ERGMs are variation-independent, that is, the parameter space is a product space.

The following finite population concentration and consistency results are taken from Corollaries 1 and 2 of Schweinberger and Stewart (2020). The first result assumes that the whole population graph is observed, whereas the second result assumes that a sample of blocks is generated by an ignorable sampling design and the subgraphs induced by the sampled blocks are observed.

THEOREM 1. *Suppose that a finite population of nodes $\mathcal{N}$ is partitioned into K blocks $\mathcal{A}_1, \ldots, \mathcal{A}_K$, where the size of the smallest block is at least 4 and the size of the largest block is a constant multiple of the smallest block, and is bounded above by a finite constant. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the data-generating parameter and $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator based on a complete observation of the population graph $\boldsymbol{Y}_{\mathcal{N}}$. Then, for all $\epsilon > 0$, there exist $\delta(\epsilon) > 0$ and $C_1 > 0$ such that, for all $K \geq 2$,*

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 < \epsilon\big) \geq 1 - 4\exp\big(-\delta(\epsilon)^2 C_1 K\big),$$

where $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2$ denotes the ℓ_2-distance between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.

More refined, and more general results on maximum likelihood and M-estimators, covering full and non-full, curved exponential-family models of random graph with correct and incorrect model specifications, can be found in Schweinberger and Stewart (2020).

Theorem 1 is a finite population concentration and consistency result in the sense that it applies to all finite populations with $K \geq 2$ blocks and shows that the probability mass of maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ concentrates around the data-generating parameter $\boldsymbol{\theta}$, provided that K is sufficiently large. Note that these results extend to many other ERGMs with block structure (Schweinberger and

Stewart, 2020). An important special case is given by K independent graphs $\boldsymbol{Y}_{\mathcal{A}_1}, \ldots, \boldsymbol{Y}_{\mathcal{A}_K}$ defined on non-empty, disjoint sets of nodes $\mathcal{A}_1, \ldots, \mathcal{A}_K$, where edges between sets of nodes $\mathcal{A}_1, \ldots, \mathcal{A}_K$ are absent with probability 1. An example is the $K = 108$ human brain networks used in Section 8, where connections between brains are impossible. Theorem 1 provides a lower bound on the probability of event $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 < \epsilon$.

Theorem 1 assumes that the whole population graph can be observed. When it is infeasible to observe the whole population graph, but it is feasible to sample blocks by using an ignorable sampling design and observing the subgraphs induced by the sampled blocks, then the following finite-sample concentration result can be obtained.

THEOREM 2. *Suppose that a sample of blocks $\mathcal{L} \subseteq \{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ is generated by an ignorable sampling design and that the subgraphs of the population graph induced by the sampled blocks are observed. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the data-generating parameter and $\widehat{\boldsymbol{\theta}}_{\mathcal{L}}$ be the maximum likelihood estimator based on the subgraphs induced by $\mathcal{L} \subseteq \{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$. Then, under the assumptions of Theorem 1, for all $\epsilon > 0$, there exist $\delta(\epsilon) > 0$ and $C_2 > 0$ such that, for all $|\mathcal{L}| \geq 2$,*

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\theta}}_{\mathcal{L}} - \boldsymbol{\theta}\|_2 < \epsilon\big) \geq 1 - 4\exp\big(-\delta(\epsilon)^2\, C_2\, |\mathcal{L}|\big).$$

The difference between Theorems 1 and 2 is that the total number of subpopulations $|\mathcal{K}|$ is replaced by the number of sampled blocks $|\mathcal{L}|$. If the population is finite but large, in the sense that the number of blocks K and the number of sampled blocks $|\mathcal{L}|$ are sufficiently large, then the probability of event $\|\widehat{\boldsymbol{\theta}}_{\mathcal{L}} - \boldsymbol{\theta}\|_2 < \epsilon$ is close to 1.

It is worth comparing these results to Chatterjee and Diaconis (2013). Chatterjee and Diaconis (2013) considered infinite populations without additional structure, resembling Ising models in physics without lattice structure and discrete Markov random fields in spatial statistics without spatial structure, and allowed edges to depend on many other edges. Many of the resulting models are ill-posed—as discussed in Section 3.1—and consistent estimation of such ill-posed models may not be possible. In contrast, the concentration and consistency results stated above are based on:

- finite populations;
- populations with additional structure in the form of subpopulations;
- short-range dependence, in the sense that dependence is restricted to subpopulations;
- within subpopulations, curved exponential-family parameterizations ensure that the added value of additional triangles decreases.

Indeed, the most important implication of Theorems 1 and 2 is that sensible assumptions give rise to sensible concentration and consistency results for well-posed ERGMs with additional structure in super and infinite population scenarios. The infinite population case is reviewed in Section 7.3.3.

7.3 Infinite Population Inference

We turn to consistency and asymptotic normality results for likelihood-based estimators of well-posed ERGMs in infinite population scenarios, including projective ERGMs (Section 7.3.1), dyad-independent ERGMs (Section 7.3.2) and dyad-dependent ERGMs (Section 7.3.3). These results cover both projective and non-projective ERGMs, showing that consistency and asymptotic normality results can be obtained for likelihood-based estimators of well-posed ERGMs despite lack of projectivity.

In addition to the work discussed below, Xiang and Neville (2011) showed that consistency results can be obtained under weak dependence assumptions, but did not give any example of an ERGM with non-trivial dependence that satisfies those weak dependence assumptions; and Mukherjee (2020) established consistency results for models with functions of degrees as sufficient statistics.

7.3.1 *Projective ERGMs.* The first set of consistency results concerns projective ERGMs.

Examples of projective ERGMs are Bernoulli(π) random graphs with size-invariant edge probability π and other dyad-independent ERGMs with size-invariant natural parameters of fixed dimension satisfying

$$\boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}') = \boldsymbol{\theta} \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and all } \mathcal{N}' \subseteq \mathcal{N}$$

and

$$\mathbb{P}_{\mathcal{N}',\boldsymbol{\theta}}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}) = \mathbb{P}_{\mathcal{N},\boldsymbol{\theta}}(\boldsymbol{Y}_{\mathcal{N}'} = \boldsymbol{y}_{\mathcal{N}'}, \boldsymbol{Y}_{\mathcal{N}\setminus\mathcal{N}'} \in \mathcal{Y}_{\mathcal{N}\setminus\mathcal{N}'}).$$

Assuming projectivity, Shalizi and Rinaldo (2013) proved:

THEOREM 3. *Let $\mathcal{N}_1, \mathcal{N}_2, \ldots$ be a sequence of sets of nodes and $\boldsymbol{Y}_{\mathcal{N}_1}, \boldsymbol{Y}_{\mathcal{N}_2}, \ldots$ be a sequence of random graphs governed by a sequence of projective ERGMs $\mathbb{P}_{\mathcal{N}_1,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_1)}, \mathbb{P}_{\mathcal{N}_2,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_2)}, \ldots$, where $\mathcal{N}_k = \{1, \ldots, k\}$ and $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathcal{N}'_k) = \boldsymbol{\theta}$ for all $\mathcal{N}'_k \subseteq \mathcal{N}_k$ $(k = 1, 2, \ldots)$. Then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{|\mathcal{N}|}$ based on $\boldsymbol{Y}_{\mathcal{N}}$ is a strongly consistent estimator of $\boldsymbol{\theta}$ as $|\mathcal{N}| \to \infty$.*

The results of Shalizi and Rinaldo (2013) extend to dyad-independent ERGMs with covariates.

7.3.2 *Dyad-independent ERGMs.* Most existing consistency and asymptotic normality results concern dyad-independent ERGMs. Examples are consistency and asymptotic normality results for β-models and p_1-models (Chatterjee, Diaconis and Sly, 2011, Rinaldo, Petrović and Fienberg, 2013, Krivitsky and Kolaczyk, 2015, Yan, Zhao and Qin, 2015, Yan, Leng and Zhu, 2016, Yan, Qin and Wang, 2016, Yan et al., 2019, Mukherjee, Mukherjee and Sen, 2018). We present here two interesting examples, one with node-dependent parameters and one with size-dependent parameters.

The first example concerns p_1-models for directed random graphs with node-dependent parameters (Yan, Leng and Zhu, 2016). Under p_1-models without reciprocity, the directed edges are independent Bernoulli($\pi_{i,j}$) random variables with edge probabilities $\pi_{i,j} = \text{logit}^{-1}(\alpha_i + \beta_j)$ and natural parameters $\eta_{i,j}(\boldsymbol{\theta}, \mathcal{N}) = \alpha_i + \beta_j$, where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_{|\mathcal{N}|}, \beta_1, \ldots, \beta_{|\mathcal{N}|})$. To make the model identifiable, Yan, Leng and Zhu (2016) set $\beta_{|\mathcal{N}|} = 0$, so $\boldsymbol{\theta} \in \mathbb{R}^{2|\mathcal{N}|-1}$. The following result follows from Theorems 1 and 2 of Yan, Leng and Zhu (2016).

THEOREM 4. *Let $\mathcal{N}_1, \mathcal{N}_2, \ldots$ be a sequence of sets of nodes and $\boldsymbol{Y}_{\mathcal{N}_1}, \boldsymbol{Y}_{\mathcal{N}_2}, \ldots$ be a sequence of random graphs governed by a sequence of p_1-models without reciprocity $\mathbb{P}_{\mathcal{N}_1,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_1)}$, $\mathbb{P}_{\mathcal{N}_2,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_2)}$, $\ldots$, where $\mathcal{N}_k = \{1, \ldots, k\}$ ($k = 1, 2, \ldots$). Assume that $\|\boldsymbol{\theta}\|_\infty \leq \tau \log|\mathcal{N}|$, where $0 < \tau < 1/44$ and $\|\boldsymbol{\theta}\|_\infty = \max_{1\leq i\leq 2|\mathcal{N}|-1} |\theta_i|$. Then*

- *with a probability approaching 1, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathcal{N}}$ based on $\boldsymbol{Y}_{\mathcal{N}}$ exists, is unique, and $\|\widehat{\boldsymbol{\theta}}_{\mathcal{N}} - \boldsymbol{\theta}\|_\infty \xrightarrow{p} 0$ as $|\mathcal{N}| \to \infty$.*
- *for any fixed $k \geq 1$, the vector consisting of the first k elements of $\widehat{\boldsymbol{\theta}}_{\mathcal{N}} - \boldsymbol{\theta}$ is asymptotically multivariate normal with mean vector zero and variance-covariance matrix given by the corresponding $k \times k$ block of the inverse Fisher information matrix as $|\mathcal{N}| \to \infty$.*

It may be surprising that consistent estimation of the parameter $\boldsymbol{\theta}$ of dimension $2|\mathcal{N}| - 1$ is possible. Note, however, that the number of independent observations from the p_1-model without reciprocity is $|\mathcal{N}|(|\mathcal{N}| - 1)$, so the number of independent observations (which is quadratic in $|\mathcal{N}|$) grows faster than the number of parameters (which is linear in $|\mathcal{N}|$). We note that additional results on p_1-models with reciprocity and covariates exist (Chatterjee, Diaconis and Sly, 2011, Rinaldo, Petrović and Fienberg, 2013, Krivitsky and Kolaczyk, 2015, Yan, Zhao and Qin, 2015, Yan, Leng and Zhu, 2016, Yan, Qin and Wang, 2016, Yan et al., 2019).

The second example concerns sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs with size-dependent edge probabilities $\pi_{|\mathcal{N}|} = \text{logit}^{-1}(\theta - \log|\mathcal{N}|)$ and natural parameters $\eta(\theta, \mathcal{N}) = \theta - \log|\mathcal{N}|$. The following result is based on Theorem 3.1 of Krivitsky and Kolaczyk (2015).

THEOREM 5. *Let $\mathcal{N}_1, \mathcal{N}_2, \ldots$ be a sequence of sets of nodes and $\boldsymbol{Y}_{\mathcal{N}_1}, \boldsymbol{Y}_{\mathcal{N}_2}, \ldots$ be a sequence of random graphs governed by a sequence of sparse Bernoulli random graph models $\mathbb{P}_{\mathcal{N}_1,\eta(\theta,\mathcal{N}_1)}, \mathbb{P}_{\mathcal{N}_2,\eta(\theta,\mathcal{N}_2)}, \ldots$, where $\mathcal{N}_k = \{1, \ldots, k\}$ ($k = 1, 2, \ldots$). Then the maximum likelihood estimator $\widehat{\theta}_{|\mathcal{N}|}$ based on $\boldsymbol{Y}_{\mathcal{N}}$ is consistent and $\sqrt{|\mathcal{N}|}(\widehat{\theta}_{|\mathcal{N}|} - \theta) \xrightarrow{d} N(0, \exp(-\theta))$ as $|\mathcal{N}| \to \infty$.*

Other consistency and asymptotic normality results for sparse and dense ERGMs with dyad-independence can be found in Krivitsky and Kolaczyk (2015). Sparse ERGMs with dyad-independence are not projective: for example, sparse Bernoulli($\pi_{|\mathcal{N}|}$) random graphs with size-dependent edge probabilities $\pi_{|\mathcal{N}|} = \text{logit}^{-1}(\theta - \log|\mathcal{N}|)$ are not projective, as shown in Section 3.2. Therefore, these consistency and asymptotic normality results demonstrate that, when meaningful sequences of random graph models are specified and larger graphs contain more information than smaller graphs, consistency and asymptotic normality results for size-invariant parameters are possible despite lack of projectivity.

7.3.3 *Dyad-dependent ERGMs*. The following result shows that maximum likelihood estimators of curved ERGMs with block structure are consistent. The result follows from Theorem 1 in Section 7.2.

THEOREM 6. *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of blocks, $\mathcal{N}_1, \mathcal{N}_2, \ldots$ be a sequence of sets of nodes defined by $\mathcal{N}_K = \bigcup_{k=1}^{K} \mathcal{A}_k$, $K = 1, 2, \ldots$, and $\boldsymbol{Y}_{\mathcal{N}_1}, \boldsymbol{Y}_{\mathcal{N}_2}, \ldots$ be a sequence of random graphs governed by a sequence of curved ERGMs with block-dependent edge and GWESP terms $\mathbb{P}_{\mathcal{N}_1,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_1)}$, $\mathbb{P}_{\mathcal{N}_2,\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N}_2)}$, $\ldots$ Under the assumptions of Theorem 1, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_K$ based on $\boldsymbol{Y}_{\mathcal{N}_K}$ is a consistent estimator of the data-generating parameter $\boldsymbol{\theta}$ as $K \to \infty$.*

Theorem 6 shows that likelihood-based inference for well-posed ERGMs with additional structure and nontrivial dependence structure is possible despite lack of projectivity. As noted in Section 7.2, these results extend to other well-posed ERGMs with block structure.

8. APPLICATION TO HUMAN BRAIN NETWORKS

To demonstrate likelihood-based inference for well-posed ERGMs in super population scenarios, we use human brain network data. A short discussion of how ERGMs can be used in neuroscience applications can be found in the survey paper of Simpson, Bowman and Laurienti (2013). Some recent applications of ERGMs to human brain network data can be found in Simpson, Hayasaka and Laurienti (2011), Simpson, Moussa and Laurienti (2012), Sinke et al. (2016), and Obando and De Vico Fallani (2017). We use here the human brain network data of Obando and De Vico Fallani (2017).

Obando and De Vico Fallani (2017) extracted data from the online PhysioNet BCI data base (Goldberger et al., 2000, Schalk et al., 2004), consisting of EEG recordings from 108 human subjects. The EEG recordings cover 56 regions within each subject's brain, over four frequency bands in two states, "eyes closed" and "eyes open." We use the same data as Obando and De Vico Fallani (2017), focusing on the beta-frequency band in the "eyes open" state. The data, thresholded by Obando and De Vico Fallani (2017), are binary, that is, $Y_{i,j} \in \{0, 1\}$, and undirected, that is, $Y_{i,j} = Y_{j,i}$ with probability 1. Here, $Y_{i,j} =$

1 can be interpreted as an indicator of a strong connection between brain regions i and j. A more detailed description of the data can be found in Obando and De Vico Fallani (2017). Last, but not least, note that the 108 brain networks are fully observed, although we subsample the 108 brain networks in Section 8.4 to demonstrate incomplete-data maximum likelihood estimation.

Goal of statistical inference. The population of interest consists of 56 regions of the human brain. The goal of statistical inference is super population inference: we want to infer the probability law that governs connections between these 56 regions of the human brain based on 108 replications. Note that the size of the human brain is bounded above by the size of the human skull and, therefore, brain networks cannot grow without bound, so infinite population inference based on sequences of graphs of increasing size is not interesting here.

Population probability models. As a population probability model, we use a curved ERGM capturing connectivity and transitivity in the 108 human brain networks. We compare the curved ERGM to latent space cluster models, which capture a stochastic tendency towards transitivity and are the main competitors of curved ERGMs for the purpose of capturing transitivity. In addition, we use the Bernoulli(π) random graph model with size-invariant edge probability π as the primary example of a projective ERGM.

Curved ERGM. We use a curved ERGM with two levels of multilevel structure, with level-1 units corresponding to the 56 brain regions and level-2 units corresponding to the 108 brains. The probability mass function of the curved ERGM is of the form

$$\begin{aligned}&\mathbb{P}_{\mathcal{N},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}})\\ &\quad= \prod_{k=1}^{108} \mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_k\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_k} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}),\end{aligned}$$

where $\mathcal{N}_k = \{1,\dots,56\}$ and $\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}$ is the observed network of brain k, and

$$\begin{aligned}&\mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_k\},\boldsymbol{\eta}(\boldsymbol{\theta},\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_k} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})\\ &\quad\propto \exp\left(\sum_{m=1}^{63} \eta_m(\boldsymbol{\theta},\mathcal{N}_k)\, s_m(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})\right).\end{aligned}$$

The sufficient statistics of the model are:

- $s_1(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})$ is the number of edges in brain k;
- $s_2(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}),\dots,s_8(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})$ are the number of nodes with $0,\dots,6$ edges in brain k, respectively;
- $s_9(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})$ is the number of paths of length two in brain k;
- $s_{10}(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}),\dots,s_{63}(\boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k})$ are the number of connected pairs of nodes with $1,\dots,56-2$ shared partners in brain k, respectively.

The natural parameters of the model are:

$$\begin{aligned}\eta_m(\boldsymbol{\theta},\mathcal{N}_k) &= \theta_m, \quad m = 1,\dots,9,\\ \eta_{9+1}(\boldsymbol{\theta},\mathcal{N}_k) &= \theta_{10}\\ &\quad + \theta_{12}\exp(\theta_{13})\left[1-\left(1-\exp(-\theta_{13})\right)\right],\\ \eta_{9+2}(\boldsymbol{\theta},\mathcal{N}_k) &= \theta_{11}\\ &\quad + \theta_{12}\exp(\theta_{13})\left[1-\left(1-\exp(-\theta_{13})\right)^2\right],\\ \eta_{9+m}(\boldsymbol{\theta},\mathcal{N}_k) &= \theta_{12}\exp(\theta_{13})\left[1-\left(1-\exp(-\theta_{13})\right)^m\right],\end{aligned}$$

where $m = 3,\dots,56-2$ and $\boldsymbol{\Theta} = \mathbb{R}^{13}$. The resulting model is a curved ERGM with a shifted GWESP term, shifted in the sense that the natural parameters of the numbers of connected pairs of nodes with 1 and 2 shared partners are shifted by θ_{10} and θ_{11}, respectively. If $\theta_{10} = 0$ and $\theta_{11} = 0$, the shifted GWESP term reduces to the unshifted GWESP term. The shifted GWESP term offers more flexibility than the unshifted GWESP term and we found that the shifted GWESP term improves in-sample and out-of-sample performance relative to the unshifted GWESP term. The model is identifiable as long as $\theta_{12} \neq 0$ and the number of nodes is at least 5, so that with positive probability there are connected pairs of nodes with 3 or more shared partners; note that $\theta_{12} = 0$ implies that θ_{13} cannot be estimated, because $\theta_{12} = 0$ eliminates the GWESP term. The number of nodes must be at least 5, because $\eta_{10}(\boldsymbol{\theta},\mathcal{N}_k) = \theta_{10} + \theta_{12}$ and $\eta_{11}(\boldsymbol{\theta},\mathcal{N}_k) = \theta_{11} + \theta_{12}$ when $\theta_{13} = 0$, so adding a constant $c \neq 0$ to θ_{10} and θ_{11} and subtracting c from θ_{12} does not change the likelihood function when the number of nodes is smaller than 5. Note that values $\theta_{13} < -\log 2$ are identifiable, but induce a form of model near-degeneracy when $|\mathcal{N}_k|$ is large, as explained in Section 3.1.3. It is possible to constrain the maximization of the likelihood function to $\theta_{13} \geq -\log 2$, but it is rarely worth enforcing the constraint, in part because $|\mathcal{N}_k| = 56$ is small and in part because unconstrained Monte Carlo maximum likelihood algorithms typically do not venture into $(-\infty, -\log 2)$. A possible explanation is that the probability of network data is higher on $[-\log 2, +\infty)$ than $(-\infty, -\log 2)$, where the model is near-degenerate and places low probability mass on graphs that resemble real-world networks, so the likelihood function is lower on $(-\infty, -\log 2)$ than $[-\log 2, +\infty)$.

We used `R` package `hergm` (Schweinberger and Luna, 2018) to estimate the curved ERGM by Monte Carlo maximum likelihood methods. The estimates, along with standard errors, are shown in Table 1.

Latent space models. To compare curved ERGMs to other models capturing transitivity, we use latent space cluster models with node-dependent propensities to form edges, which generalize β-models, stochastic block models and latent space models. Suppose that each node i has a latent position $\boldsymbol{z}_i \in \mathbb{R}^3$, edges are independent conditional on the positions of nodes, and the log odds of the

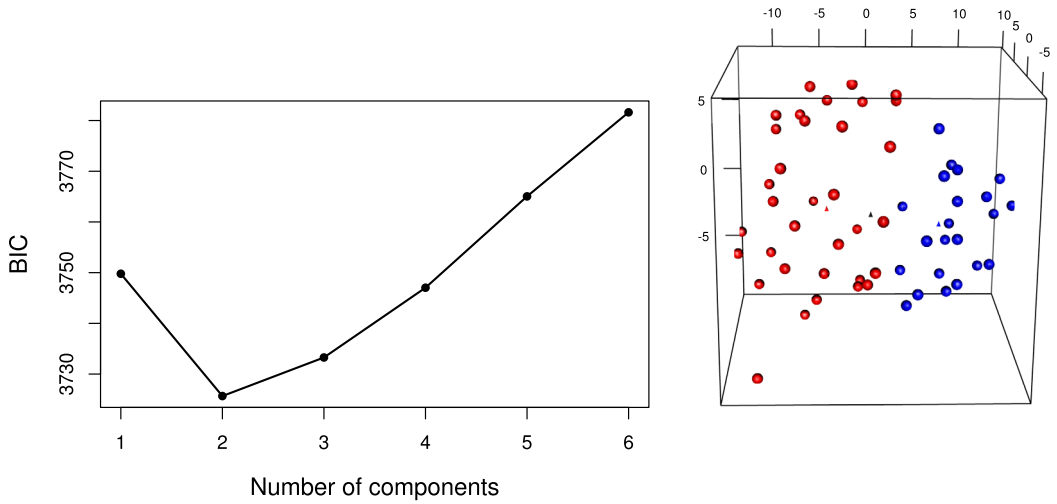

FIG. 3. *Latent space model*: *BIC based on* $q = 1, \ldots, 6$ *components* (*left*) *and estimated positions of the* 56 *nodes in* $\mathbb{R}^3$ *based on* $q = 2$ *components* (*right*). *The edges are not shown, because the three-dimensional plots produced by* R *package* `latentnet` *can represent binary edges, but cannot represent counts of the number of edges between brain regions based on* 108 *replications*.

TABLE 1
Monte Carlo maximum likelihood estimates, including standard errors, of all parameters in the curved ERGM, with the exception of the (*nuisance*) *parameters* $\theta_2, \ldots, \theta_8$.

Parameter		Estimate	Standard error
θ_1	Edge parameter	−4.972	0.560
θ_9	Two-path parameter	−0.091	0.033
θ_{10}	GWESP shift parameter	0.198	0.024
θ_{11}	GWESP shift parameter	0.305	0.022
θ_{12}	GWESP base parameter	1.061	0.018
θ_{13}	GWESP decay parameter	1.565	0.028

conditional probability of an edge between nodes i and j in brain k given the positions of nodes i and j is

$$\log \frac{\mathbb{P}_{\mathcal{N},\alpha,\boldsymbol{\beta}}(Y_{i,j} = 1 \mid \boldsymbol{z}_i, \boldsymbol{z}_j)}{1 - \mathbb{P}_{\mathcal{N},\alpha,\boldsymbol{\beta}}(Y_{i,j} = 1 \mid \boldsymbol{z}_i, \boldsymbol{z}_j)} = \alpha + \beta_i + \beta_j - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2,$$

where $\alpha \in \mathbb{R}$ can be interpreted as the overall propensity to form edges in the population and $\beta_i \in \mathbb{R}$ and $\beta_j \in \mathbb{R}$ can be interpreted as the deviations of nodes i and j, respectively. Since adding a constant $c \neq 0$ to α and subtracting $c/2$ from parameters $\beta_1, \ldots, \beta_{56}$ does not change the conditional probability of an edge, we set $\beta_1 = 0$. The positions of nodes are drawn from a multivariate Gaussian mixture distribution with q multivariate Gaussian component distributions, as in Handcock, Raftery and Tantrum (2007). The resulting models can be considered as generalizations of the β-models described in Section 4.1, which assume that the log odds of the probability of an edge is $\beta_i + \beta_j$. The additional term $-||\boldsymbol{z}_i - \boldsymbol{z}_j||_2$ can be interpreted as a penalty, which discourages edges between nodes separated by large distances. The fact that the positions of nodes are generated from a multivariate Gaussian mixture distribution with q components implies that the set of nodes is partitioned into q subsets, so one can view the resulting model as a generalization of stochastic block models. Such latent space models—generalizing β-models (e.g., Chatterjee, Diaconis and Sly, 2011), stochastic block models (e.g., Nowicki and Snijders, 2001), latent space models (e.g., Hoff, Raftery and Handcock, 2002) and latent space cluster models (Handcock, Raftery and Tantrum, 2007)—were proposed by Krivitsky et al. (2009).

We used R package `latentnet` (Krivitsky and Handcock, 2008) to estimate the latent space model. The number of components q was selected by BIC, as recommended by Handcock, Raftery and Tantrum (2007). The BIC shown in Figure 3 suggests to choose $q = 2$ components. All of the following results are based on $q = 2$ components. The estimated positions of the 56 nodes in $\mathbb{R}^3$ based on $q = 2$ components can be seen in Figure 3.

Bernoulli random graph model. We use the Bernoulli(π) random graph model with size-invariant edge probability π as the primary example of a projective ERGM. The Bernoulli(π) random graph model has population probability mass function

$$\mathbb{P}_{\mathcal{N},\eta(\theta,\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}} = \boldsymbol{y}_{\mathcal{N}}) = \prod_{k=1}^{108} \mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_k\},\eta(\theta,\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_k} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}),$$

where

$$\mathbb{P}_{\{\mathcal{N}_k,\mathcal{N}_k\},\eta(\theta,\mathcal{N})}(\boldsymbol{Y}_{\mathcal{N}_k,\mathcal{N}_k} = \boldsymbol{y}_{\mathcal{N}_k,\mathcal{N}_k}) \propto \exp\left(\eta(\theta,\mathcal{N}) \sum_{i<j:\, i,j\in\mathcal{N}_k} y_{i,j}\right)$$

with natural parameter $\eta(\theta, \mathcal{N}) = \operatorname{logit}(\pi) = \theta \in \mathbb{R}$.

Other random graph models. We do not use β-models and stochastic block models (e.g., Nowicki and Snijders, 2001), because both can be viewed as special cases of the latent space model described above. In addition, these and other models are not designed to capture transitivity, so comparing them to the curved ERGM and latent space model would be unfair to them.

8.1 In-Sample Performance

We first compare the in-sample performance of the Bernoulli random graph model, the latent space model and the curved ERGM in terms of geodesic distances (the length of the shortest path between dyads); the number of nodes with degree m (the number of nodes with m edges); the number of connected dyads with m triangles; and the total number of triangles; note that a dyad refers to a pair of nodes. All statistics are aggregated over the 108 brain networks. The first two statistics help assess the in-sample performance of models in terms of connectivity and reachability, whereas the other two help assess the in-sample performance in terms of transitivity.

The in-sample performance of the three models in terms of these statistics is shown in Figure 4. At least two interesting observations can be made. First, while some simplistic ERGMs lacking structure are indistinguishable from Bernoulli random graphs in the large-graph limit—as discussed in Section 3.1—the in-sample performance of the curved ERGM is very different from the in-sample performance of the Bernoulli random graph model. Indeed, the in-sample performance of the curved ERGM is far superior to the Bernoulli random graph model in terms of both connectivity and transitivity. Second, Figure 4 highlights a limitation of latent space models: while the latent space model induces a stochastic tendency towards transitivity, it is not a flexible model of transitivity, and it is not able to match the observed tendency towards transitivity. In fact, the latent space model has 56 unrestricted parameters $\alpha, \beta_2, \ldots, \beta_{56}$ and 56 latent variables $z_1, \ldots, z_{56}$, but it is outperformed by the curved ERGM with 13 unrestricted parameters $\theta_1, \ldots, \theta_{13}$, both in terms of connectivity and transitivity. Last, but not least, it is worth noting that the latent space model is not able to match the observed numbers of nodes with degrees $0, 1, \ldots$ While the reasons are unclear, it is crystal-clear why the curved ERGM is able to match the observed numbers of nodes with degrees $0, \ldots, 6$: the numbers of nodes with degrees $0, \ldots, 6$ are sufficient statistics of the curved ERGM and, under the maximum likelihood estimate, the expected and observed numbers of nodes with degrees $0, \ldots, 6$ are equal. Note that the numbers of nodes with degrees $7, 8, \ldots$ are not sufficient statistics of the curved ERGM, but the tail of the degree distribution seems to be captured by the other model terms.

8.2 Out-of-Sample Performance

We assess the out-of-sample performance of the best-fitting model, the curved ERGM, by sampling 75% of the 108 brains at random and estimating the curved ERGM from the sampled brain networks. We then generated model-based predictions of the 25% non-sampled brain networks based on the estimated curved ERGM.

Figure 5 suggests that the curved ERGM has high predictive power in terms of connectivity and transitivity: the model-based predictions are close to the observed statistics. The strong out-of-sample performance lends credence to the assumption that the 108 brain networks were generated by a common data-generating mechanism.

8.3 Sampling Brains: More Data Helps Estimate Parameters

To show that more data helps estimate parameters, we sampled 25%, 50% and 75% of the brains at random and observed the whole brain network of each sampled brain. We performed the described procedure 50 times.

Figure 6 shows that the Monte Carlo maximum likelihood estimates based on samples of brain networks approach the corresponding Monte Carlo maximum likelihood estimates based on all 108 brain networks, demonstrating that observing more networks does indeed improve parameter estimates.

8.4 Subsampling Brains: Incomplete-Data Maximum Likelihood Estimation

To illustrate likelihood-based inference based on incomplete data generated by ignorable incomplete-data processes, we sampled 50% and 75% of the nodes in each of the 108 brain networks at random and observed the edges of all sampled nodes. We used the incomplete-data Monte Carlo maximum likelihood procedure described in Section 6 to estimate the parameters from sampled subgraphs. We performed the procedure 50 times.

The results in Figure 7 show that incomplete-data Monte Carlo maximum likelihood estimates approach the Monte Carlo maximum likelihood estimates based on observing 100% of the nodes in the 108 brain networks. These results underscore that statistical inference from subgraphs to population graphs is possible despite lack of projectivity, as long as statistical inference is based on the likelihood function.

8.5 How to Deal with Graphs of Different Sizes

The human brain network application in Sections 8.1–8.4 demonstrates that curved ERGMs can outperform latent space models in super population scenarios where a population probability model generates graphs of the same size. A legitimate question to ask is how one can deal with super population scenarios where a population probability model generates graphs of different sizes.

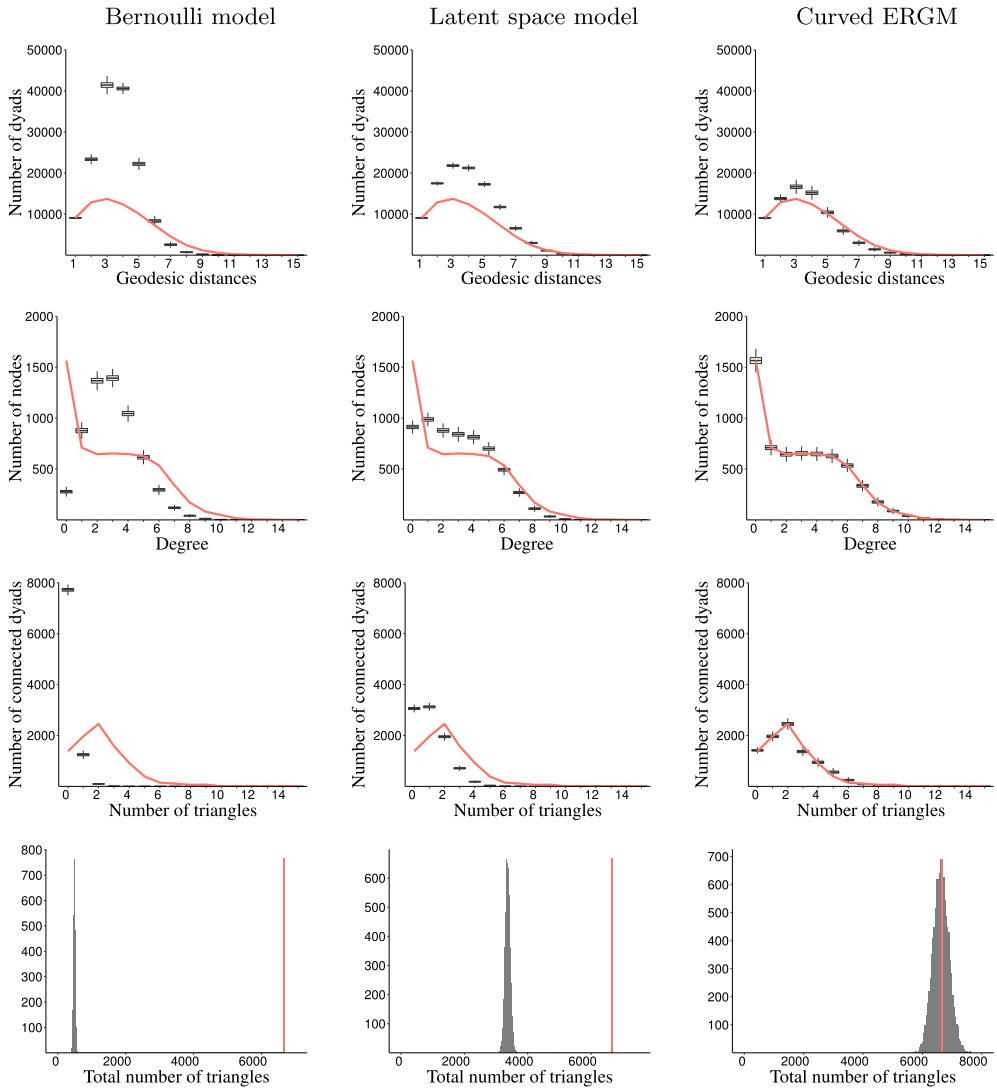

FIG. 4. *In-sample performance of the Bernoulli model, the latent space model, and the curved ERGM. The red lines and curves indicate the observed values of the statistics.*

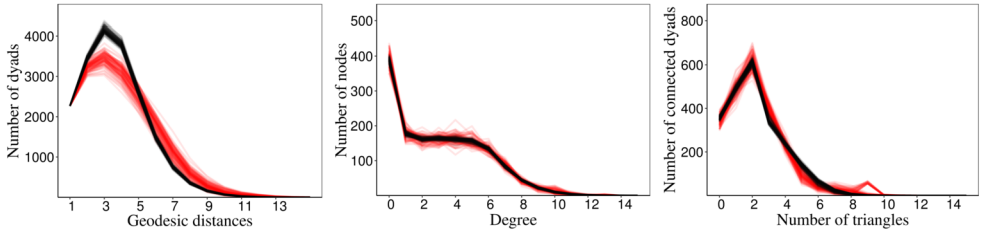

FIG. 5. *Out-of-sample performance of the curved ERGM, using* 75% *of the brain networks to estimate the curved ERGM and* 25% *of the brain networks to generate model-based predictions. The black curves are the model-based predictions, whereas the red curves are the observations.*

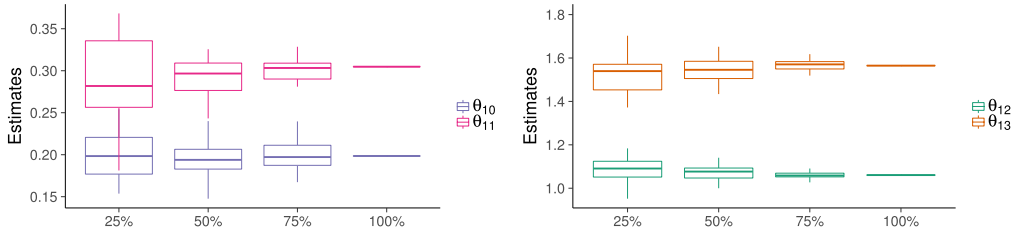

FIG. 6. *Sampling brains: Monte Carlo maximum likelihood estimates of the most interesting parameters, the parameters* $\theta_{10}, \ldots, \theta_{13}$ *of the shifted GWESP term capturing transitivity, based on observing* 25%, 50%, 75%, *and* 100% *of the* 108 *brain networks.*

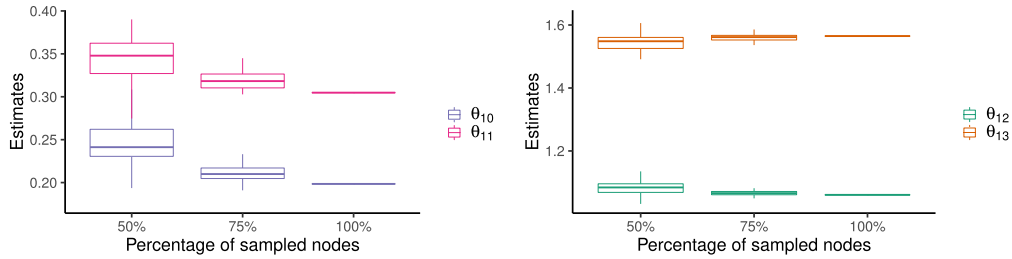

FIG. 7. *Subsampling brains: incomplete-data Monte Carlo maximum likelihood estimates of the most interesting parameters, the parameters* $\theta_{10}, \ldots, \theta_{13}$ *of the shifted GWESP term capturing transitivity, based on observing* 50%, 75%, *and* 100% *of the nodes in the* 108 *brain networks.*

Such scenarios arise in applications of ERGMs with block structure (Section 4.3) and multilevel structure (Section 4.4).

There are at least three approaches to accounting for different network sizes:

(a) size-dependent natural parameters using size-dependent offsets (Krivitsky, Handcock and Morris, 2011, Krivitsky and Kolaczyk, 2015, Butts and Almquist, 2015, Stewart et al., 2019);
(b) size-dependent natural parameters using network size as a covariate (Slaughter and Koehly, 2016);
(c) network-specific natural parameters with common mean and network-specific deviations (random effects) (Schweinberger and Handcock, 2015).

All of them assume that the natural parameters have the form

$$\textit{natural parameter} = \textit{size-invariant parameter} + \textit{size-dependent deviation},$$

so that the resulting natural parameters are size-dependent, and the goal of statistical inference is to estimate the size-invariant parameters.

9. CONCLUSIONS

The ERGM framework is widely used in practice, ranging from the study of the human brain and epidemics to differential privacy and social networks (see Section 1). We believe that the ERGM framework is most useful in super population scenarios (Section 5.1.2), although it can be useful in finite population scenarios (Section 5.1.1) and infinite population scenarios (Section 5.1.3), provided well-posed ERGMs are used and appropriate statistical procedures are employed (e.g., likelihood-based procedures). The consistency and asymptotic normality results for likelihood-based estimators in finite, super and infinite population scenarios reviewed in Section 7 confirm that statistical inference for ERGMs is possible, provided the language of exponential families is used to ask well-posed questions. It goes without saying that the language can be abused to ask ill-posed questions by specifying ill-posed models. But every language can be abused, and potential for abuse does not invalidate its potential for eloquent and effective communication when properly employed.

There is no denying that ERGMs are complex models and give rise to non-trivial computational challenges, challenges that are shared with other discrete exponential-family models for dependent random variables, such as discrete Markov random fields in spatial statistics (Besag, 1974, Cressie, 1993, Stein, 1999) and machine learning (e.g., Ravikumar, Wainwright and Lafferty, 2010, Yang et al., 2015). However, there is no such thing as a free lunch: ERGMs model complex dependence, and modeling complex dependence comes at a price. Stochastic block models and projective ERGMs are simpler models and more attractive on computational grounds, but are not capable of capturing the complex dependencies encountered in network data. Latent space models do capture a stochastic tendency towards transitivity (e.g., Hoff, Raftery and Handcock, 2002, Handcock, Raftery and Tantrum, 2007), but there may be more transitivity in network data than expected under latent space models, as we found in the human brain network data in Section 8. In addition, latent space models are not flexible models of other forms of complex dependence. Last, but not least, likelihood-based inference for latent space models is likewise expensive in terms of computing time, even when approximate procedures are used (e.g., Raftery et al., 2012, Salter-Townshend and Murphy, 2013). In the end, all of these approaches have useful applications and belong to an ever-growing arsenal of statistical tools to understand the structure of complex and dependent network data.

ACKNOWLEDGMENTS

We are indebted to Catalina Obando and Fabrizio De Vico Fallani for sharing the human brain network data used in Section 8. In addition, we are grateful to Johan Koskinen and the Department of Statistics at the University of Stockholm, Sweden for making the seminal works on pseudo-likelihood-based versus likelihood-based inference from the 1990s accessible. Last, but not least, we would like to thank an associate editor and three anonymous referees for helpful suggestions, and Noel Cressie, Mathias Drton, Mark S. Handcock,

Eric D. Kolaczyk, Martina Morris and Thomas Richardson for insightful discussions. Michael Schweinberger and Jonathan R. Stewart were partially supported by NSF awards DMS-1513644 and DMS-1812119. Carter T. Butts was partially supported by ARO award W911NF-14-1-0552 and NSF award DMS-136142.

REFERENCES

AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

ALMQUIST, Z. W. and BAGOZZI, B. E. (2019). Using radical environmentalist texts to uncover network structure and network features. *Sociol. Methods Res.* **48** 905–960. MR4003633 https://doi.org/10.1177/0049124117729696

AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859 https://doi.org/10.1214/13-AOS1138

ARISTOFF, D. and RADIN, C. (2013). Emergent structures in large networks. *J. Appl. Probab.* **50** 883–888. MR3102521 https://doi.org/10.1239/jap/1378401243

ASUNCION, A., LIU, Q., IHLER, A. T. and SMYTH, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In *Thirtheenth International Conference on AI and Statistics* 33–40.

ATCHADÉ, Y. F., LARTILLOT, N. and ROBERT, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Braz. J. Probab. Stat.* **27** 416–436. MR3105037 https://doi.org/10.1214/11-BJPS174

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634 https://doi.org/10.1126/science.286.5439.509

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. MR0489333

BEARMAN, P. S., MOODY, J. and STOVEL, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *Amer. J. Sociol.* **110** 44–91.

BERK, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Stat.* **43** 193–204. MR0298810

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BHAMIDI, S., BRESLER, G. and SLY, A. (2008). Mixing time of exponential random graphs. In 2008 *IEEE 49th Annual IEEE Symposium on Foundations of Computer Science* 803–812.

BHAMIDI, S., BRESLER, G. and SLY, A. (2011). Mixing time of exponential random graphs. *Ann. Appl. Probab.* **21** 2146–2170. MR2895412 https://doi.org/10.1214/10-AAP740

BHAMIDI, S., CHAKRABORTY, S., CRANMER, S. and DESMARAIS, B. (2018). Weighted exponential random graph models: Scope and large network limits. *J. Stat. Phys.* **173** 704–735. MR3876904 https://doi.org/10.1007/s10955-018-2103-0

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. In *Proceedings of the National Academy of Sciences* **106** 21068–21073.

BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

BINKIEWICZ, N., VOGELSTEIN, J. T. and ROHE, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104** 361–377. MR3698259 https://doi.org/10.1093/biomet/asx008

BOLLOBÁS, B. (1985). *Random Graphs*. Academic Press [Harcourt Brace Jovanovich, Publishers], London. MR0809996

BOLLOBÁS, B., RIORDAN, O., SPENCER, J. and TUSNÁDY, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures Algorithms* **18** 279–290. MR1824277 https://doi.org/10.1002/rsa.1009

BORGS, C., CHAYES, J. T., COHN, H. and VEITCH, V. (2019). Sampling perspectives on sparse exchangeable graphs. *Ann. Probab.* **47** 2754–2800. MR4021236 https://doi.org/10.1214/18-AOP1320

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

BRAILLY, J., FAVRE, G., CHATELLET, J. and LAZEGA, E. (2016). Embeddedness as a multilevel problem: A case study in economic sociology. *Soc. Netw.* **44** 319–333.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

BUTTS, C. T. (2008). A relational event framework for social action. *Sociol. Method.* **38** 155–200.

BUTTS, C. T. (2011). Bernoulli graph bounds for general random graph models. *Sociol. Method.* **41** 299–345.

BUTTS, C. T. (2015). A novel simulation method for binary discrete exponential families, with application to social networks. *J. Math. Sociol.* **39** 174–202. MR3367717 https://doi.org/10.1080/0022250X.2015.1022279

BUTTS, C. T. (2018). A perfect sampling method for exponential family random graph models. *J. Math. Sociol.* **42** 17–36. MR3764794 https://doi.org/10.1080/0022250X.2017.1396985

BUTTS, C. T. (2019). A dynamic process interpretation of the sparse ERGM reference model. *J. Math. Sociol.* **43** 40–57. MR3891856 https://doi.org/10.1080/0022250X.2018.1490737

BUTTS, C. T. and ACTON, R. M. (2011). Spatial modeling of social networks. In *The SAGE Handbook of GIS and Society Research* (T. Nyerges, H. Couclelis and R. McMaster, eds.) 222–250 12. SAGE, Thousand Oaks.

BUTTS, C. T. and ALMQUIST, Z. W. (2015). A flexible parameterization for baseline mean degree in multiple-network ERGMs. *J. Math. Sociol.* **39** 163–167. MR3367715 https://doi.org/10.1080/0022250X.2014.967851

BYSHKIN, M., STIVALA, A., MIRA, A., ROBINS, G. and LOMI, A. (2018). Fast maximum likelihood estimation via equilibrium expectation for large network data. *Sci. Rep.* **8** 11509. https://doi.org/10.1038/s41598-018-29725-8

CAI, D., CAMPBELL, T. and BRODERICK, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 4249–4257.

CAIMO, A. and FRIEL, N. (2011). Bayesian inference for exponential random graph models. *Soc. Netw.* **33** 41–55.

CAIMO, A. and FRIEL, N. (2013). Bayesian model selection for exponential random graph models. *Soc. Netw.* **35** 11–24.

CAIMO, A. and GOLLINI, I. (2020). A multilayer exponential random graph modelling approach for weighted networks. *Comput. Statist. Data Anal.* **142** 106825, 18. MR3995271 https://doi.org/10.1016/j.csda.2019.106825

CAIMO, A. and MIRA, A. (2015). Efficient computational strategies for doubly intractable problems with applications to Bayesian social networks. *Stat. Comput.* **25** 113–125. MR3304912 https://doi.org/10.1007/s11222-014-9516-7

CARON, F. and FOX, E. B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

CHATTERJEE, S. and DIACONIS, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461. MR3127871 https://doi.org/10.1214/13-AOS1155

CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435. MR2857452 https://doi.org/10.1214/10-AAP728

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253 https://doi.org/10.1093/biomet/asr053

CORANDER, J., DAHMSTRÖM, K. and DAHMSTRÖM, P. (1998). Maximum likelihood estimation for Markov graphs Technical Report Department of Statistics, Univ. Stockholm.

CORANDER, J., DAHMSTRÖM, K. and DAHMSTRÖM, P. (2002). Maximum likelihood estimation for exponential random graph models. In *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics*; *A Festschrift in Honour of Ove Frank* (J. Hagberg, ed.) 1–17. Dept. Statistics, Univ. Stockholm.

CRANE, H. (2018). *Probabilistic Foundations of Statistical Network Analysis*. *Monographs on Statistics and Applied Probability* **157**. CRC Press, Boca Raton, FL. MR3791467

CRANE, H. and DEMPSEY, W. (2018). Edge exchangeable models for interaction networks. *J. Amer. Statist. Assoc.* **113** 1311–1326. MR3862359 https://doi.org/10.1080/01621459.2017.1341413

CRANE, H. and DEMPSEY, W. (2020). A statistical framework for modern network science. *Statist. Sci*. To appear.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1239641 https://doi.org/10.1002/9781119115151

DAHMSTRÖM, K. and DAHMSTRÖM, P. (1993). ML-estimation of the clustering parameter in a Markov graph model Technical Report Univ. Stockholm, Department of Statistics.

DAHMSTRÖM, K. and DAHMSTRÖM, P. (1999). Properties of different estimators of the parameters in Markov graphs. In *Bulletin of the International Statistical Institute* 1–2. International Statistical Institute. Available at https://tilastokeskus.fi/isi99/proceedings/arkisto/varasto/dahm0777.pdf.

DAWID, A. P. and DICKEY, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Assoc.* **72** 845–850. MR0471124

DESMARAIS, B. A. and CRANMER, S. J. (2012). Statistical inference for valued-edge networks: The generalized exponential random graph model. *PLoS ONE* **7** 1–12.

DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242. MR0428531

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376. MR0471152

ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. MR0120167

ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. MR0125031

EVERITT, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comput. Graph. Statist.* **21** 940–960. MR3005805 https://doi.org/10.1080/10618600.2012.687493

FELLOWS, I. and HANDCOCK, M. S. (2012). Exponential-family random network models. Available at arXiv:1208.0121.

FELLOWS, I. and HANDCOCK, M. S. (2017). Removing phase transitions from Gibbs measures. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* (A. Singh and J. Zhu, eds.) **54** 289–297. Proceedings of Machine Learning Research.

FIENBERG, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Comput. Graph. Statist.* **21** 825–839. MR3005799 https://doi.org/10.1080/10618600.2012.738106

FIENBERG, S. E. and SLAVKOVIC, A. (2010). Data privacy and confidentiality. In *International Encyclopedia of Statistical Science* 342–345. Springer, Berlin.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A* **222** 309–368.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society A* **144** 285–307.

FOSDICK, B. K. and HOFF, P. D. (2015). Testing and modeling dependencies between a network and nodal attributes. *J. Amer. Statist. Assoc.* **110** 1047–1056. MR3420683 https://doi.org/10.1080/01621459.2015.1008697

FOSDICK, B. K., MCCORMICK, T. H., MURPHY, T. B., NG, T. L. J. and WESTLING, T. (2019). Multiresolution network models. *J. Comput. Graph. Statist.* **28** 185–196. MR3939381 https://doi.org/10.1080/10618600.2018.1505633

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

FRIEZE, A. and KAROŃSKI, M. (2016). *Introduction to Random Graphs*. Cambridge Univ. Press, Cambridge. MR3675279 https://doi.org/10.1017/CBO9781316339831

GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699. MR1185217

GILBERT, E. N. (1959). Random graphs. *Ann. Math. Stat.* **30** 1141–1144. MR0108839 https://doi.org/10.1214/aoms/1177706098

GILE, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Statist. Assoc.* **106** 135–146. MR2816708 https://doi.org/10.1198/jasa.2011.ap09475

GILE, K. and HANDCOCK, M. S. (2006). Model-based assessment of the impact of missing data on inference for networks Technical Report Center for Statistics and the Social Sciences, Univ. Washington, Seattle. Available at https://www.csss.washington.edu/Papers/wp66.pdf.

GILE, K. and HANDCOCK, M. H. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327.

GILE, K. J. and HANDCOCK, M. S. (2017). Analysis of networks with missing data with application to the National Longitudinal Study of Adolescent Health. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 501–519. MR3632339 https://doi.org/10.1111/rssc.12184

GILL, P. S. and SWARTZ, T. B. (2004). Bayesian analysis of directed graphs data with applications to social networks. *J. Roy. Statist. Soc. Ser. C* **53** 249–260. MR2055619 https://doi.org/10.1046/j.1467-9876.2003.05215.x

GJOKA, M., SMITH, E. J. and BUTTS, C. T. (2014). Estimating clique composition and size distributions from sampled network data. *Proceedings of the Sixth IEEE Workshop on Network Science for Communication Networks* (*NetSciCom* 2014).

GJOKA, M., SMITH, E. and BUTTS, C. T. (2015). Estimating subgraph frequencies with or without attributes from egocentrically sampled data. Available at arxiv.org/abs/1510.08119.

GOLDBERGER, A. L., AMARAL, L. A., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E.,

MOODY, G. B., PENG, C.-K. et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101** e215–e220.

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2009). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

GONDAL, N. (2018). Duality of departmental specializations and PhD exchange: A Weberian analysis of status in interaction using multilevel exponential random graph models (mERGM). *Soc. Netw.* **55** 202–212.

GOODMAN, L. A. (1961). Snowball sampling. *Ann. Math. Stat.* **32** 148–170. MR0124140 https://doi.org/10.1214/aoms/1177705148

GOODREAU, S. M., KITTS, J. A. and MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46** 103–125.

GOODREAU, S. M., HANDCOCK, M. S., HUNTER, D. R., BUTTS, C. T. and MORRIS, M. (2008). A statnet tutorial. *J. Stat. Softw.* **24** 1–27.

GRAZIOLI, G., MARTIN, R. W. and BUTTS, C. T. (2019). Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Frontiers in Molecular Biosciences*, *Biological Modeling and Simulation* **6**.

GRAZIOLI, G., YU, Y., UNHELKAR, M. H., MARTIN, R. W. and BUTTS, C. T. (2019). Network-based classification and modeling of amyloid fibrils. *J Phys Chem B* **123** 5452–5462. https://doi.org/10.1021/acs.jpcb.9b03494

GROENDYKE, C., WELCH, D. and HUNTER, D. R. (2012). A network-based analysis of the 1861 Hagelloch measles data. *Biometrics* **68** 755–765. MR3055180 https://doi.org/10.1111/j.1541-0420.2012.01748.x

HÄGGSTRÖM, O. and JONASSON, J. (1999). Phase transition in the random triangle model. *J. Appl. Probab.* **36** 1101–1115. MR1742153 https://doi.org/10.1017/s0021900200017897

HANDCOCK, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In *Dynamic Social Network Modeling and Analysis*: *Workshop Summary and Papers* (R. Breiger, K. Carley and P. Pattison, eds.) 1–12. National Academies Press, Washington, DC.

HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. MR2758082 https://doi.org/10.1214/08-AOAS221

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300 https://doi.org/10.1111/j.1467-985X.2007.00471.x

HANNEKE, S., FU, W. and XING, E. P. (2010). Discrete temporal models of social networks. *Electron. J. Stat.* **4** 585–605. MR2660534 https://doi.org/10.1214/09-EJS548

HARRIS, J. K. (2013). *An Introduction to Exponential Random Graph Modeling*. Sage, Thousand Oaks.

HARTLEY, H. O. and SIELKEN, R. L. JR. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics* **31** 411–422. MR0386084 https://doi.org/10.2307/2529429

HE, R. and ZHENG, T. (2015). GLMLE: Graph-limit enabled fast computation for fitting exponential random graph models to large social networks. *Soc. Netw. Anal. Min.* **5** 1–19.

HECKATHORN, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.

HOFF, P. D. (2003). Random effects models for network data. In *Dynamic Social Network Modeling and Analysis*: *Workshop Summary and Papers* (R. Breiger, K. Carley and P. Pattison, eds.) 303–312. National Academies Press, Washington, DC.

HOFF, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* **100** 286–295. MR2156838 https://doi.org/10.1198/016214504000001015

HOFF, P. D. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 657–664. MIT Press, Cambridge, MA.

HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 971–992. MR2750253 https://doi.org/10.1111/j.1467-9868.2009.00716.x

HOFF, P. D. (2020). Additive and multiplicative effects network models. *Statist. Sci*. To appear.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

HOLLAND, P. W. and LEINHARDT, S. (1970). A method for detecting structure in sociometric data. *Amer. J. Sociol.* **76** 492–513.

HOLLAND, P. W. and LEINHARDT, S. (1972). Some evidence on the transitivity of positive interpersonal sentiment. *Amer. J. Sociol.* **77** 1205–1209.

HOLLAND, P. W. and LEINHARDT, S. (1976). Local structure in social networks. *Sociol. Method.* 1–45.

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

HOLLWAY, J. and KOSKINEN, J. (2016). Multilevel embeddedness: The case of the global fisheries governance complex. *Soc. Netw.* **44** 281–294.

HOLLWAY, J., LOMI, A., PALLOTTI, F. and STADTFELD, C. (2017). Multilevel social spaces: The network dynamics of organizational fields. *Netw. Sci.* **5** 187–212.

HOMANS, G. C. (1950). *The Human Group*. Harcourt, Brace, New York.

HUITSING, G., VAN DUIJN, M. A. J., SNIJDERS, T. A. B., WANG, P., SAINIO, M., SALMIVALLI, C. and VEENSTRA, R. (2012). Univariate and multivariate models of positive and negative networks: Liking, disliking, and bully–victim relationships. *Soc. Netw.* **34** 645–657.

HUMMEL, R. M., HUNTER, D. R. and HANDCOCK, M. S. (2012). Improving simulation-based algorithms for fitting ERGMs. *J. Comput. Graph. Statist.* **21** 920–939. MR3005804 https://doi.org/10.1080/10618600.2012.679224

HUNTER, D. R. (2007). Curved exponential family models for social networks. *Soc. Netw.* **29** 216–230. https://doi.org/10.1016/j.socnet.2006.08.005

HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258. MR2394635 https://doi.org/10.1198/016214507000000446

HUNTER, D. R. and HANDCOCK, M. S. (2006). Inference in curved exponential family models for networks. *J. Comput. Graph. Statist.* **15** 565–583. MR2291264 https://doi.org/10.1198/106186006X133069

HUNTER, D. R., KRIVITSKY, P. N. and SCHWEINBERGER, M. (2012). Computational statistical methods for social network models. *J. Comput. Graph. Statist.* **21** 856–882. MR3005801 https://doi.org/10.1080/10618600.2012.732921

HUNTER, D. R., HANDCOCK, M. S., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24** 1–29.

ISING, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Z. Phys.* **31** 253–258.

JANSON, S. (2018). On edge exchangeable random graphs. *J. Stat. Phys.* **173** 448–484. MR3876897 https://doi.org/10.1007/s10955-017-1832-9

JANSON, S., ŁUCZAK, T. and RUCINSKI, A. (2000). *Random Graphs. Wiley-Interscience Series in Discrete Mathematics and Optimization*. Wiley Interscience, New York. MR1782847 https://doi.org/10.1002/9781118032718

JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

JIN, I. H. and LIANG, F. (2013). Fitting social network models using varying truncation stochastic approximation MCMC algorithm. *J. Comput. Graph. Statist.* **22** 927–952. MR3173750 https://doi.org/10.1080/10618600.2012.680851

JIN, I. H., YUAN, Y. and LIANG, F. (2013). Bayesian analysis for exponential random graph models using the adaptive exchange sampler. *Stat. Interface* **6** 559–576. MR3164659 https://doi.org/10.4310/SII.2013.v6.n4.a13

JONASSON, J. (1999). The random triangle model. *J. Appl. Probab.* **36** 852–867. MR1737058 https://doi.org/10.1239/jap/1032374639

KARWA, V., KRIVITSKY, P. N. and SLAVKOVIĆ, A. B. (2017). Sharing social network data: Differentially private estimation of exponential family random-graph models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 481–500. MR3632338 https://doi.org/10.1111/rssc.12185

KARWA, V., PETROVIĆ, S. and BAJIĆ, D. (2016). DERGMs: Degeneracy-restricted exponential random graph models. Preprint. Available at arXiv:1612.03054.

KARWA, V. and SLAVKOVIĆ, A. (2016). Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.* **44** 87–112. MR3449763 https://doi.org/10.1214/15-AOS1358

KENYON, R. and YIN, M. (2017). On the asymptotics of constrained exponential random graphs. *J. Appl. Probab.* **54** 165–180. MR3632612 https://doi.org/10.1017/jpr.2016.93

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models. Springer Series in Statistics*. Springer, New York. MR2724362 https://doi.org/10.1007/978-0-387-88146-1

KOSKINEN, J. (2004). Essays on Bayesian inference for social networks. Ph.D. thesis Stockholm Univ., Dept. of Statistics, Sweden.

KOSKINEN, J. H. (2009). Using latent variables to account for heterogeneity in exponential family random graph models. In *Proceedings of the 6th St. Petersburg Workshop on Simulation* (S. M. Ermakov, V. B. Melas and A. N. Pepelyshev, eds.) **2** 845–849. St. Petersburg State Univ., St. Petersburg, Russia.

KOSKINEN, J. H., ROBINS, G. L. and PATTISON, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Stat. Methodol.* **7** 366–384. MR2643608 https://doi.org/10.1016/j.stamet.2009.09.007

KRACKHARDT, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Soc. Netw.* **10** 359–381. MR0984597 https://doi.org/10.1016/0378-8733(88)90004-4

KRIVITSKY, P. N. (2012). Exponential-family random graph models for valued networks. *Electron. J. Stat.* **6** 1100–1128. MR2988440 https://doi.org/10.1214/12-EJS696

KRIVITSKY, P. N. (2017). Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. *Comput. Statist. Data Anal.* **107** 149–161. MR3575065 https://doi.org/10.1016/j.csda.2016.10.015

KRIVITSKY, P. N. and BUTTS, C. T. (2017). Exponential-family random graph models for rank-order relational data. *Sociol. Method.* **47** 68–112.

KRIVITSKY, P. N. and HANDCOCK, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *J. Stat. Softw.* **24**. https://doi.org/10.18637/jss.v024.i05

KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46. MR3153932 https://doi.org/10.1111/rssb.12014

KRIVITSKY, P. N., HANDCOCK, M. S. and MORRIS, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Stat. Methodol.* **8** 319–339. MR2800354 https://doi.org/10.1016/j.stamet.2011.01.005

KRIVITSKY, P. N. and KOLACZYK, E. D. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statist. Sci.* **30** 184–198. MR3353102 https://doi.org/10.1214/14-STS502

KRIVITSKY, P. N., MARCUM, C. S. and KOEHLY, L. (2019). Exponential-family random graph models for multi-layer networks.

KRIVITSKY, P. N. and MORRIS, M. (2017). Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. *Ann. Appl. Stat.* **11** 427–455. MR3634330 https://doi.org/10.1214/16-AOAS1010

KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E. and HOFF, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* **31** 204–213. https://doi.org/10.1016/j.socnet.2009.04.001

KURANT, M., MARKOPOULOU, A. and THIRAN, P. (2011). Towards unbiased BFS sampling. *IEEE J. Sel. Areas Commun.* **29** 1799–1809.

KURANT, M., GJOKA, M., WANG, Y., ALMQUIST, Z. W., BUTTS, C. T. and MARKOPOULOU, A. (2012). Coarse-grained topology estimation via graph sampling. In *Proceedings of ACM SIGCOMM Workshop on Online Social Networks* (*WOSN*) '12.

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York. MR1419991

LAURITZEN, S. L. (2008). Exchangeable Rasch matrices. *Rend. Mat. Appl.* (7) **28** 83–95. MR2463441

LAURITZEN, S., RINALDO, A. and SADEGHI, K. (2018). Random networks, graphical models and exchangeability. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 481–508. MR3798875 https://doi.org/10.1111/rssb.12266

LAZEGA, E. and PATTISON, P. E. (1999). Multiplexity, generalized exchange and cooperation in organizations: A case study. *Soc. Netw.* **21** 67–90.

LAZEGA, E. and SNIJDERS, T. A. B., eds. (2016). *Multilevel Network Analysis for the Social Sciences*. Springer, Cham.

LEHMANN, E. L. (1999). *Elements of Large-Sample Theory. Springer Texts in Statistics*. Springer, New York. MR1663158 https://doi.org/10.1007/b98855

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

LEIFELD, P., CRANMER, S. J. and DESMARAIS, B. A. (2018). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *J. Stat. Softw.* **83** 1–36.

LIANG, F. and JIN, I.-H. (2013). A Monte Carlo Metropolis–Hastings algorithm for sampling from distributions with intractable normalizing constants. *Neural Comput.* **25** 2199–2234. MR3100001 https://doi.org/10.1162/NECO_a_00466

LIANG, F., JIN, I. H., SONG, Q. and LIU, J. S. (2016). An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *J. Amer. Statist. Assoc.* **111** 377–393. MR3494666 https://doi.org/10.1080/01621459.2015.1009072

LOMI, A., ROBINS, G. and TRANMER, M. (2016). Introduction to multilevel social networks. *Soc. Netw.* **44** 266–268. https://doi.org/10.1016/j.socnet.2015.10.006

LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035 https://doi.org/10.1090/coll/060

LUBBERS, M. J. (2003). Group composition and network structure in school classes: A multilevel application of the p* model. *Soc. Netw.* **25** 309–332.

LUBBERS, M. J. and SNIJDERS, T. A. B. (2007). A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Soc. Netw.* **29** 489–507.

LUNAGOMEZ, S. and AIROLDI, E. (2014). Bayesian inference from non-ignorable network sampling designs. Available at arXiv:1401.4718.

LUSHER, D., KOSKINEN, J. and ROBINS, G. (2013). *Exponential Random Graph Models for Social Networks*. Cambridge Univ. Press, Cambridge, UK.

LYNE, A.-M., GIROLAMI, M., ATCHADÉ, Y., STRATHMANN, H. and SIMPSON, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statist. Sci.* **30** 443–467. MR3432836 https://doi.org/10.1214/15-STS523

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0727836 https://doi.org/10.1007/978-1-4899-3244-0

MCPHERSON, J. M. (1983). An ecology of affiliation. *Am. Sociol. Rev.* **48** 519–532.

MELE, A. (2017). A structural model of dense network formation. *Econometrica* **85** 825–850. MR3664180 https://doi.org/10.3982/ECTA10400

MEREDITH, C., VAN DEN NOORTGATE, W., STRUYVE, C., GIELEN, S. and KYNDT, E. (2017). Information seeking in secondary schools: A multilevel network approach. *Soc. Netw.* **50** 35–45.

MØLLER, J., PETTITT, A. N., REEVES, R. and BERTHELSEN, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93** 451–458. MR2278096 https://doi.org/10.1093/biomet/93.2.451

MORRIS, M., HANDCOCK, M. S. and HUNTER, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *J. Stat. Softw.* **24** 1–24.

MUKHERJEE, S. (2013). Phase transition in the two star exponential random graph model. Available at arXiv:1310.4164.

MUKHERJEE, S. (2020). Degeneracy in sparse ERGMs with functions of degrees as sufficient statistics. *Bernoulli*. To appear.

MUKHERJEE, R., MUKHERJEE, S. and SEN, S. (2018). Detection thresholds for the β-model on sparse graphs. *Ann. Statist.* **46** 1288–1317. MR3798004 https://doi.org/10.1214/17-AOS1585

MURRAY, I., GHAHRAMANI, Z. and MACKAY, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence* 359–366. AUAI Press, Corvallis, OR.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255 https://doi.org/10.1198/016214501753208735

OBANDO, C. and DE VICO FALLANI, F. (2017). A statistical model for brain networks inferred from large-scale electrophysiological signals. *J. R. Soc. Interface* 1–10.

OKABAYASHI, S. and GEYER, C. J. (2012). Long range search for maximum likelihood in exponential families. *Electron. J. Stat.* **6** 123–147. MR2879674 https://doi.org/10.1214/11-EJS664

ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.

OUZIENKO, V., GUO, Y. and OBRADOVIC, Z. (2011). A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks. *Stat. Anal. Data Min.* **4** 470–486. MR2842404 https://doi.org/10.1002/sam.10130

PARK, J. and HARAN, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *J. Amer. Statist. Assoc.* **113** 1372–1390. MR3862364 https://doi.org/10.1080/01621459.2018.1448824

PARK, J. and NEWMAN, M. E. J. (2004). Solution of the two-star model of a network. *Phys. Rev. E* (3) **70** 066146, 5. MR2133810 https://doi.org/10.1103/PhysRevE.70.066146

PARK, J. and NEWMAN, M. E. J. (2005). Solution for the properties of a clustered network. *Phys. Rev. E* **72** 026136.

PATTISON, P. and ROBINS, G. (2002). Neighborhood-based models for social networks. In *Sociological Methodology* (R. M. Stolzenberg, ed.) **32** 301–337. Blackwell Publishing, Boston, MA.

PATTISON, P. and WASSERMAN, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *Br. J. Math. Stat. Psychol.* **52** 169–193.

PATTISON, P. E., ROBINS, G. L., SNIJDERS, T. A. B. and WANG, P. (2013). Conditional estimation of exponential random graph models from snowball sampling designs. *J. Math. Psych.* **57** 284–296. MR3137882 https://doi.org/10.1016/j.jmp.2013.05.004

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876 https://doi.org/10.1214/aos/1176350710

RADIN, C. and YIN, M. (2013). Phase transitions in exponential random graphs. *Ann. Appl. Probab.* **23** 2458–2471. MR3127941 https://doi.org/10.1214/12-AAP907

RAFTERY, A. E., NIU, X., HOFF, P. D. and YEUNG, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *J. Comput. Graph. Statist.* **21** 901–919. MR3005803 https://doi.org/10.1080/10618600.2012.679240

RAPOPORT, A. (1979/80). A probabilistic approach to networks. *Soc. Netw.* **2** 1–18. MR0551137 https://doi.org/10.1016/0378-8733(79)90008-X

RASTELLI, R., FRIEL, N. and RAFTERY, A. E. (2016). Properties of latent variable network models. *Netw. Sci.* **4** 407–432.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

RICHARDSON, M. and DOMINGOS, P. (2006). Markov logic networks. *Mach. Learn.* **62** 107–136.

RINALDO, A., FIENBERG, S. E. and ZHOU, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **3** 446–484. MR2507456 https://doi.org/10.1214/08-EJS350

RINALDO, A., PETROVIĆ, S. and FIENBERG, S. E. (2013). Maximum likelihood estimation in the β-model. *Ann. Statist.* **41** 1085–1110. MR3113804 https://doi.org/10.1214/12-AOS1078

ROBINS, G. and PATTISON, P. (2001). Random graph models for temporal processes in social networks. *J. Math. Sociol.* **25** 5–41.

ROBINS, G. L., PATTISON, P. E. and WANG, P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Soc. Netw.* **31** 105–117.

ROBINS, G., PATTISON, P. and WASSERMAN, S. (1999). Logit models and logistic regressions for social networks. III. Valued relations. *Psychometrika* **64** 371–394. MR1720089 https://doi.org/10.1007/BF02294302

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

ROLLS, D. A., WANG, P., JENKINSON, R., PATTISON, P. E., ROBINS, G. L., SACKS-DAVIS, R., DARAGANOVA, G., HELLARD, M. and MCBRYDE, E. (2013). Modelling a disease-relevant contact network of people who inject drugs. *Soc. Netw.* **35** 699–710.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

SALGANIK, M. J. and HECKATHORN, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–239.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Comput. Statist. Data Anal.* **57** 661–671. MR2981116 https://doi.org/10.1016/j.csda.2012.08.004

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2015). Role analysis in networks using mixtures of exponential random graph models. *J. Comput. Graph. Statist.* **24** 520–538. MR3357393 https://doi.org/10.1080/10618600.2014.923777

SALTER-TOWNSHEND, M., WHITE, A., GOLLINI, I. and MURPHY, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Stat. Anal. Data Min.* **5** 260–264. MR2958152 https://doi.org/10.1002/sam.11146

SCHALK, G., MCFARLAND, D. J., HINTERBERGER, T., BIRBAUMER, N. and WOLPAW, J. R. (2004). BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* **51** 1034–1043.

SCHWEINBERGER, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *J. Amer. Statist. Assoc.* **106** 1361–1370. MR2896841 https://doi.org/10.1198/jasa.2011.tm10747

SCHWEINBERGER, M. (2020). Consistent structure estimation of exponential-family random graph models with block structure. *Bernoulli*. To appear.

SCHWEINBERGER, M. and HANDCOCK, M. S. (2015). Local dependence in random graph models: Characterization, properties and statistical inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 647–676. MR3351449 https://doi.org/10.1111/rssb.12081

SCHWEINBERGER, M. and LUNA, P. (2018). HERGM: Hierarchical exponential-family random graph models. *J. Stat. Softw.* **85** 1–39.

SCHWEINBERGER, M. and SNIJDERS, T. A. B. (2003). Settings in social networks: A measurement model. In *Sociological Methodology* (R. M. Stolzenberg, ed.) **33** 307–341 10. Basil Blackwell, Boston & Oxford.

SCHWEINBERGER, M. and STEWART, J. (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *Ann. Statist.* To appear.

SENGUPTA, S. and CHEN, Y. (2018). A block model for node popularity in networks with community structure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 365–386. MR3763696 https://doi.org/10.1111/rssb.12245

SEWELL, D. K. and CHEN, Y. (2015). Latent space models for dynamic networks. *J. Amer. Statist. Assoc.* **110** 1646–1657. MR3449061 https://doi.org/10.1080/01621459.2014.988214

SHALIZI, C. R. and RINALDO, A. (2013). Consistency under sampling of exponential random graph models. *Ann. Statist.* **41** 508–535. MR3099112 https://doi.org/10.1214/12-AOS1044

SIMPSON, S. L., BOWMAN, F. D. and LAURIENTI, P. J. (2013). Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. *Stat. Surv.* **7** 1–36. MR3161730 https://doi.org/10.1214/13-SS103

SIMPSON, S. L., HAYASAKA, S. and LAURIENTI, P. J. (2011). Exponential random graph modeling for complex brain networks. *PLoS ONE* **6** e20039. https://doi.org/10.1371/journal.pone.0020039

SIMPSON, S. L., MOUSSA, M. N. and LAURIENTI, P. J. (2012). An exponential random graph modeling approach to creating group-based representative whole-brain connectivity networks. *NeuroImage* **60** 1117–1126.

SINKE, M. R. T., DIJKHUIZEN, R. M., CAIMO, A., STAM, C. J. and OTTE, W. M. (2016). Bayesian exponential random graph modeling of whole-brain structural networks across lifespan. *NeuroImage* **135** 79–91.

SLAUGHTER, A. J. and KOEHLY, L. M. (2016). Multilevel models for social networks: Hierarchical Bayesian approaches to exponential random graph modeling. *Soc. Netw.* **44** 334–345.

SMITH, T. W., MARSDEN, P., HOUT, M. and KIM, J. (1972–2016). General Social Surveys Technical Report NORC at the Univ. Chicago.

SNIJDERS, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *J. Soc. Struct.* **3** 1–40.

SNIJDERS, T. A. B. (2010). Conditional marginalization for exponential random graph models. *J. Math. Sociol.* **34** 239–252.

SNIJDERS, T. A. B. and BOSKER, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed. Sage Publications, Los Angeles, CA. MR3137621

SNIJDERS, T. A. B. and VAN DUIJN, M. A. J. (2002). Conditional maximum likelihood estimation under various specifications of exponential random graph models. In *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in Honour of Ove Frank* (J. Hagberg, ed.) 117–134. Dept. Statistics, Univ. Stockholm.

SNIJDERS, T. A. B., PATTISON, P. E., ROBINS, G. L. and HANDCOCK, M. S. (2006). New specifications for exponential random graph models. *Sociol. Method.* **36** 99–153.

STEIN, M. L. (1999). *Interpolation of Spatial Data. Springer Series in Statistics*. Springer, New York. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

STEWART, J., SCHWEINBERGER, M., BOJANOWSKI, M. and MORRIS, M. (2019). Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. *Soc. Netw.* **59** 98–119.

STRAUSS, D. (1986). On a general class of models for interaction. *SIAM Rev.* **28** 513–527. MR0867682 https://doi.org/10.1137/1028156

STRAUSS, D. and IKEDA, M. (1990). Pseudolikelihood estimation for social networks. *J. Amer. Statist. Assoc.* **85** 204–212. MR1137368

SUESSE, T. (2012). Marginalized exponential random graph models. *J. Comput. Graph. Statist.* **21** 883–900. MR3005802 https://doi.org/10.1080/10618600.2012.694750

TALAGRAND, M. (1996). A new look at independence. *Ann. Probab.* **24** 1–34. MR1387624 https://doi.org/10.1214/aop/1042644705

TANG, M., SUSSMAN, D. L. and PRIEBE, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *Ann. Statist.* **41** 1406–1430. MR3113816 https://doi.org/10.1214/13-AOS1112

THIEMICHEN, S. and KAUERMANN, G. (2017). Stable exponential random graph models with non-parametric components for large dense networks. *Soc. Netw.* **49** 67–80.

THIEMICHEN, S., FRIEL, N., CAIMO, A. and KAUERMANN, G. (2016). Bayesian exponential random graph models with nodal random effects. *Soc. Netw.* **46** 11–28.

THOMPSON, S. K. (2012). *Sampling*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2894042 https://doi.org/10.1002/9781118162934

THOMPSON, S. and FRANK, O. (2000). Model-based estimation with link-tracing sampling designs. *Surv. Methodol.* **26** 87–98.

VAN DUIJN, M. A. J. (1995). Estimation of a random effects model for directed graphs. In *Toeval Zit Overal: Programmatuur voor Random-Coëffciënt Modellen* (T. A. B. Snijders, B. Engel, J. C. Van Houwelingen, A. Keen, G. J. Stemerdink and M. Verbeek, eds.) 113–131. IEC ProGAMMA, Groningen.

VAN DUIJN, M. A. J., GILE, K. and HANDCOCK, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Soc. Netw.* **31** 52–62.

VAN DUIJN, M. A. J., SNIJDERS, T. A. B. and ZIJLSTRA, B. J. H. (2004). p_2: A random effects model with covariates for directed graphs. *Stat. Neerl.* **58** 234–254. MR2064846 https://doi.org/10.1046/j.0039-0402.2003.00258.x

VEITCH, V. and ROY, D. M. (2015). The class of random graphs arising from exchangeable random measures. Preprint. Available at arXiv:1512.03099.

VEITCH, V. and ROY, D. M. (2019). Sampling and estimation for (sparse) exchangeable graphs. *Ann. Statist.* **47** 3274–3299. MR4025742 https://doi.org/10.1214/18-AOS1778

WANG, J. and ATCHADÉ, Y. F. (2014). Approximate Bayesian computation for exponential random graph models for large social networks. *Comm. Statist. Simulation Comput.* **43** 359–377. MR3200975 https://doi.org/10.1080/03610918.2012.703359

WANG, P., ROBINS, G. and PATTISON, P. (2006). PNet. program for the simulation and estimation of exponential random graph (p*) models. Melbourne School of Psychological Sciences, University of Melbourne.

WANG, P., ROBINS, G., PATTISON, P. and LAZEGA, E. (2013). Exponential random graph models for multilevel networks. *Soc. Netw.* **35** 96–115.

WANG, P., ROBINS, G., PATTISON, P. and LAZEGA, E. (2016a). Social selection models for multilevel networks. *Soc. Netw.* **44** 346–362.

WANG, C., BUTTS, C. T., HIPP, J. R., JOSE, R. and LAKON, C. M. (2016b). Multiple imputation for missing edge data: A predictive evaluation method with application to add health. *Soc. Netw.* **45** 89–98. https://doi.org/10.1016/j.socnet.2015.12.003

WANG, Y., FANG, H., YANG, D., ZHAO, H. and DENG, M. (2018). Network clustering analysis using mixture exponential-family random graph models and its application in genetic interaction data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2017.2743711.

WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, Cambridge.

WASSERMAN, S. and PATTISON, P. (1996). Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and p. *Psychometrika* **61** 401–425. MR1424909 https://doi.org/10.1007/BF02294547

WILLINGER, W., ALDERSON, D. and DOYLE, J. C. (2009). Mathematics and the Internet: A source of enormous confusion and great potential. *Notices Amer. Math. Soc.* **56** 586–599. MR2509062

WYATT, D., CHOUDHURY, T. and BILMES, J. (2008). Learning hidden curved exponential random graph models to infer face-to-face interaction networks from situated speech data. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* 732–738.

XIANG, R. and NEVILLE, J. (2011). Relational learning with one network: An asymptotic analysis. In *Proceedings of the* 14*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 1–10.

YAN, T., LENG, C. and ZHU, J. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *Ann. Statist.* **44** 31–57. MR3449761 https://doi.org/10.1214/15-AOS1343

YAN, T., QIN, H. and WANG, H. (2016). Asymptotics in undirected random graph models parameterized by the strengths of vertices. *Statist. Sinica* **26** 273–293. MR3468353

YAN, T., ZHAO, Y. and QIN, H. (2015). Asymptotic normality in the maximum entropy models on graphs with an increasing number of parameters. *J. Multivariate Anal.* **133** 61–76. MR3282018 https://doi.org/10.1016/j.jmva.2014.08.013

YAN, T., JIANG, B., FIENBERG, S. E. and LENG, C. (2019). Statistical inference in a directed network model with covariates. *J. Amer. Statist. Assoc.* **114** 857–868. MR3963186 https://doi.org/10.1080/01621459.2018.1448829

YANG, X., RINALDO, A. and FIENBERG, S. E. (2014). Estimation for dyadic-dependent exponential random graph models. *J. Algebr. Stat.* **5** 39–63. MR3279953 https://doi.org/10.18409/jas.v5i1.24

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YIN, M., RINALDO, A. and FADNAVIS, S. (2016). Asymptotic quantization of exponential random graphs. *Ann. Appl. Probab.* **26** 3251–3285. MR3582803 https://doi.org/10.1214/16-AAP1175

ZAPPA, P. and LOMI, A. (2015). The analysis of multilevel networks in organizations: Models and empirical tests. *Organ. Res. Methods* **18** 542–569.

ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083 https://doi.org/10.1214/12-AOS1036