Local Dependence in Random Graph Models: Characterization, Properties and Statistical Inference

Michael Schweinberger

Department of Statistics, Rice University, Houston, TX, USA

Mark S. Handcock

Department of Statistics, University of California, Los Angeles, CA, USA

Summary. Dependent phenomena, such as relational, spatial, and temporal phenomena, tend to be characterised by local dependence in the sense that units which are close in a well-defined sense are dependent. In contrast to spatial and temporal phenomena, though, relational phenomena tend to lack a natural neighborhood structure in the sense that it is unknown which units are close and thus dependent. Owing to the challenge of characterising local dependence and constructing random graph models with local dependence, many conventional exponential-family random graph models induce strong dependence and are not amenable to statistical inference. We take first steps to characterise local dependence in random graph models, inspired by the notion of finite neighborhoods in spatial statistics and M-dependence in time series, and show that local dependence endows random graph models with desirable properties which make them amenable to statistical inference. We show that random graph models with local dependence satisfy a natural domain consistency condition which every model should satisfy, but conventional exponential-family random graph models do not satisfy. In addition, we establish a Central Limit Theorem for random graph models with local dependence, which suggests that random graph models with local dependence are amenable to statistical inference. We discuss how random graph models with local dependence can be constructed by exploiting either observed or unobserved neighborhood structure. In the absence of observed neighborhood structure, we take a Bayesian view and express the uncertainty about the neighborhood structure by specifying a prior on a set of suitable neighborhood structures. We present simulation results and applications to two real-world networks with ground truth.

Keywords: social networks, weak dependence, local dependence, M-dependence, exponential families, model degeneracy

1. Introduction

Network data arise in many fields, including biology, the health sciences, economics, political science, sociology, machine learning, and engineering. In these fields there are many applications with important societal implications, such as protein-protein interactions, the spread of infectious diseases, contagion in financial markets, insurgencies, terrorist networks, criminal networks, social networks, the internet, and power grids (e.g., Kolaczyk, 2009).

We consider a single observation of a network (e.g., a social network) with n nodes and $N = n(n-1)$ directed or $N = n(n-1)/2$ undirected edge variables. The statistical analysis of a single observation of a network is more challenging than the statistical analysis of multiple, independent networks, because such data are both dependent and high-dimensional and give rise to unique conceptual, computational, and statistical challenges.

We are concerned with problems of *specification* in the sense of Fisher (1922, p. 313), i.e., with the problem of identifying families of distributions which are capable of modeling a wide range of

E-mail: michael.schweinberger@rice.edu

dependencies of substantive interest and which are also amenable to statistical inference. In the past decade, Exponential-Family Random Graph Models (ERGMs) have attracted much attention (e.g., Frank and Strauss, 1986; Wasserman and Pattison, 1996; Lusher et al., 2013). Despite attractive finite-sample properties (e.g., Barndorff-Nielsen, 1978), ERGMs have turned out to be problematic models of real-world networks (e.g., Snijders, 2002; Handcock, 2003; Hunter et al., 2008).

One of the most striking observations is that some of the most interesting ERGMs do not place much probability mass on graphs which resemble real-world networks (e.g., Handcock, 2003; Hunter et al., 2008). Let $\boldsymbol{Y} \in \mathbb{Y}$ be a random graph on a finite set of nodes, corresponding to edge variables $Y_{i,j}$ between pairs of nodes i, j. To simplify the discussion, suppose that the edge variables $Y_{i,j}$ are binary, i.e., $Y_{i,j} \in \{0, 1\}$. A convenient representation of a distribution $\mathbb{P}$ with support $\mathbb{Y}$ is as an exponential family of the form

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) \;\;=\;\; \exp\left[\langle \boldsymbol{\theta}, s(\boldsymbol{y}) \rangle - \psi(\boldsymbol{\theta})\right], \; \boldsymbol{y} \in \mathbb{Y}, \tag{1}$$

where $\langle \boldsymbol{\theta}, s(\boldsymbol{y}) \rangle$ denotes the inner product of a d-vector of natural parameters $\boldsymbol{\theta}$ and a d-vector of sufficient statistics $s(\boldsymbol{y})$, and $\psi(\boldsymbol{\theta})$ is a log normalising constant. The d-vector of sufficient statistics may include statistics of interest, such as the number of transitive triples of nodes (e.g., in friendship networks, "a friend of my friend is my friend"). Such statistics induce dependence between edge variables and are of great interest in the fast-growing field of network science (e.g., Wasserman and Faust, 1994; Kolaczyk, 2009; Lusher et al., 2013). If $\boldsymbol{S} = s(\boldsymbol{Y})$ denotes the vector of sufficient statistics and $\mathbb{S}$ denotes the convex hull of $\{s(\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{Y}\}$, then the induced distribution of $\boldsymbol{S}$ is given by

$$\mathbb{Q}_{\boldsymbol{\theta}}(S \in \mathcal{S}) \;\;=\;\; \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} \in S^{-1}(\mathcal{S})) \;\;=\;\; \sum_{\boldsymbol{y} \,\in\, S^{-1}(\mathcal{S})} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}), \tag{2}$$

where $S^{-1}(\mathcal{S})$ denotes the subset of $\mathbb{Y}$ mapping into $\mathcal{S} \subset \mathbb{S}$. If the sufficient statistics include counts of the number of transitive triples and other subgraph configurations, then the induced distribution of $\boldsymbol{S}$ tends to place much probability mass on extreme graphs which do not resemble real-world networks. There is both theoretical and empirical evidence which suggests that many families of distributions with such count statistics place much mass on the relative boundary rather than the relative interior of $\mathbb{S}$ (e.g., Snijders, 2002; Handcock, 2003; Hunter et al., 2008). Worse, theoretical results indicate that the behavior of conventional ERGMs does not improve as the number of nodes n increases, but deteriorates (Strauss, 1986; Jonasson, 1999; Schweinberger, 2011; Butts, 2011). The best-known example are Markov random graph models (Frank and Strauss, 1986), though other interesting models are problematic as well. The flawed nature of conventional ERGMs is demonstrated by Figure 1. It relates to an ERGM of the form (1) with the number of edges and triangles as sufficient statistics with $n = 100$ nodes and $N = 4{,}950$ undirected edge variables. The figure shows the prior predictive distributions of the sufficient statistics from the model, which is described in detail in Section 5.1. It demonstrates that the prior predictive distribution places most of its mass on graphs which are extreme in terms of the number of edges and triangles.

In this paper, we address the root of the problem by characterising and constructing well-behaved random graph models which are amenable to statistical inference. The point of departure is the observation that many statistics of interest, $\boldsymbol{S}$, are sums of random variables, e.g., counts of the number of edges and transitive triples. The distribution of (normed) sums of independent or weakly dependent random variables tends to be Gaussian by virtue of some version of the Central Limit Theorem (e.g., Billingsley, 1995). If the expected value of $\boldsymbol{S}$ is in the relative interior of $\mathbb{S}$, then the model should place significant mass around the expected value of $\boldsymbol{S}$ by virtue of the approximate Gaussian distribution of $\boldsymbol{S}$. Therefore, for all expected values of $\boldsymbol{S}$ in the relative interior of $\mathbb{S}$, the model should place much mass on the relative interior of $\mathbb{S}$. The difficulty is that in the absence of spatial, temporal, and other structure, it is not evident which edge variables should be dependent. Therefore, the specification of random graph models with weak dependence is challenging and conventional

Fig. 1. Prior predictions of the number of edges (left) and triangles (right) under the global triangle model with $N = 4,950$ edge variables. Note the extreme polarisation.
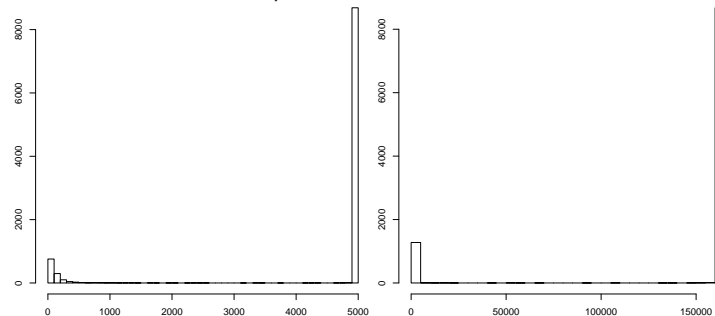
ERGMs induce either no dependence and are simplistic (e.g., Bernoulli random graph models) or strong dependence and are near-degenerate (e.g., Markov random graph models, Frank and Strauss, 1986).

We take first steps to characterise local dependence in random graph models in Section 2. We demonstrate that local dependence endows random graph models with desirable properties which makes them amenable to statistical inference. One property is a natural domain consistency condition that any probability model should satisfy, but many parametrisations of ERGMs do not satisfy (Shalizi and Rinaldo, 2013). A second, and more important property is asymptotic Gaussian behavior of statistics, which suggests that random graph models with local dependence place much probability mass around the expected value of statistics of interest. We discuss the construction of random graph models with local dependence in Section 3 and Bayesian inference given complete as well as incomplete data in Section 4. If suitable neighborhood structure is observed, at least two approaches to statistical inference are possible, depending on whether the observed neighborhood structure is regarded as fixed or random. If no suitable neighborhood structure is observed, we take a Bayesian view and express the uncertainty about the neighborhood structure by specifying a prior on a set of suitable neighborhood structures, using hierarchical parametric and non-parametric priors and auxiliary-variable Markov chain Monte Carlo methods. We present simulation results and applications to two real-world networks with ground truth in Section 5.

Other, related work. Snijders et al. (2006) and Hunter and Handcock (2006) considered non-linear constraints on the parameter space of ERGMs. Such curved ERGMs have been applied with some success (Hunter et al., 2008), but do not admit simple representations of dependencies and the interpretation of parameters is challenging, as noted by Snijders et al. (2006, p. 149). An alternative are latent variable models, which we discuss in Section 2.1. Selected, special cases and other, related work are discussed in Section 3.4.

2. Dependence

We discuss in Section 2.1 two broad approaches to modeling dependence, one based on latent variable models and the other one based on ERGMs, and we argue that ERGMs are attractive when dependence is of substantive interest. We discuss in Section 2.2 the challenges encountered in modeling dependence of substantive interest by ERGMs. In Section 2.3, we introduce a notion of local dependence and in Section 2.4 we show that that local dependence endows models with desirable properties which makes them amenable to statistical inference.

2.1. Dependence of substantive interest

Most relational phenomena are dependent phenomena, and dependence is often of substantive interest. Examples can be found in the social sciences (e.g., Wasserman and Faust, 1994; Lusher et al., 2013), economics (e.g., Jackson, 2008), the health sciences (e.g., Welch et al., 2011), and physics (Newman et al., 2002). As an example, if i, j, k is a triple of nodes and the edge variables $Y_{i,j}, Y_{j,k}, Y_{i,k} \in \{0, 1\}$ are binary and undirected, then the triple is called *transitive* if $Y_{i,j} Y_{j,k} Y_{i,k} = 1$, i.e., there exist edges between i and j as well as j and k and i and k. Other examples are discussed by Wasserman and Faust (1994) and Lusher et al. (2013).

When modeling transitivity and other dependencies, it is not attractive to assume conditional independence of edge variables, e.g.,

$$Y_{i,j} \mid p_{i,j} \overset{\text{ind}}{\sim} \quad \text{Bernoulli}(p_{i,j}), \quad i < j, \tag{3}$$

where $p_{i,j}$ denotes the probability of a binary, undirected edge between nodes i, j, which may depend on observed and unobserved, latent variables. Examples of models of the form (3) are stochastic block models (e.g., Nowicki and Snijders, 2001) and mixed membership models (e.g., Airoldi et al., 2008); random effects and mixed effects models (e.g., van Duijn et al., 2004; Hoff, 2005); and latent space models (Hoff et al., 2002; Schweinberger and Snijders, 2003; Handcock et al., 2007; Krivitsky et al., 2009). While models of the form (3) can capture transitive closure by introducing latent structure, such models induce dependence indirectly through latent variables rather than directly. In situations where dependence is of substantive interest, scientists tend to prefer models which allow to specify dependencies directly. Examples are the spatial covariance and variogram functions in spatial random field models (Cressie, 1993), interaction functions in spatial point processes (Møller and Waagepetersen, 2004), and covariance terms in time series (Granger and Morris, 1976). An additional, well-known example is the Ising model in physics (e.g., Georgii, 2011): The Ising model allows the explicit specification of the nature of interactions between particles. Physicists would hesitate to exchange the Ising model for a latent variable model which assumed that particles are independent conditional on observed structure (e.g., observed locations on a lattice) or unobserved, latent structure.

In the realm of networks, Frank and Strauss (1986) and Wasserman and Pattison (1996) introduced exponential-family models which resemble the Ising model in physics and lattice models in spatial statistics and which allow the modeling of a wide range of dependencies of substantive interest, including transitive closure. Such models have attracted much attention in the social sciences and health sciences and elsewhere, as the recent book by Lusher et al. (2013) testifies.

2.2. Modeling dependence: challenges

Despite the fact that ERGMs are the natural relatives of well-established models in physics, spatial statistics, machine learning and artificial intelligence, ERGMs lack something most other areas have: neighborhood structure. The lack of neighborhood structure makes modeling dependence challenging.

In spatial statistics (e.g., Cressie, 1993; Stein, 1999) and time series (e.g., Granger and Morris, 1976) and the related work on mixing conditions in probability theory (Billingsley, 1995, pp. 363–370, Dedecker et al., 2007), it is assumed that dependence decreases as the distance between random variables in the spatial or temporal domain increases. Thus, events may be dependent as long as the distance between the locations of the events is small, while distant events are almost dependent. In the realm of networks, though, it is not evident what distance between subgraphs means and how the dependence between subgraphs should decrease. A possible approach is to assume uniform weak dependence in the sense that all edges and subgraphs in large graphs are almost independent. Such dependence assumptions are not appealing, however, because network science would expect strong, local dependence between some of the edges and subgraphs.

In fact, even when spatial or temporal structure is available, there is often more local structure than would be expected based on the location of nodes in space and time: e.g., some subsets of nodes may be close in geographical space, but the members of the subset may be distant in "network space," while other subsets of nodes may be distant in geographical space, but close in "network space," where "network space" is understood as other structure not captured by geographical space; e.g., researchers in the same building, on the same floor, and in the same department may not collaborate, but may engage in transitive collaborations with other researchers distant in geographical space.

A well-known example of the challenge of modeling dependence are Markov random graph models (Frank and Strauss, 1986). Suppose that the random graph $\boldsymbol{Y}$ is undirected and binary. Motivated by the nearest neighbor definition in physics (Georgii, 2011) and spatial statistics (Besag, 1974), Frank and Strauss (1986) called two dyads $\{i, j\}$ and $\{k, l\}$ neighbors if $\{i, j\}$ and $\{k, l\}$ share a node and assumed that, if $\{i, j\}$ and $\{k, l\}$ are not neighbors, then $Y_{i,j}$ and $Y_{k,l}$ are independent conditional on the rest of random graph $\boldsymbol{Y}$. Markov random graph models can be represented in exponential-family form (1) with the number of edges $s_1(\boldsymbol{y}) = \sum_{i<j} y_{i,j}$, the number of k-stars $s_k(\boldsymbol{y}) = \sum_i \sum_{j_1 < \cdots < j_k} y_{i,j_1} \cdots y_{i,j_k}$, and the number of triangles $s_n(\boldsymbol{y}) = \sum_{i<j<k} y_{i,j}\, y_{j,k}\, y_{i,k}$ as sufficient statistics. Markov random graph models and generalisations to ERGMs (Wasserman and Pattison, 1996) allow scientists to model transitive closure and other dependencies along with covariate-related similarity, which scientists have long considered to be of great interest (e.g., Wasserman and Faust, 1994; Lusher et al., 2013). Despite the underlying nearest-neighbor assumption and its scientific appeal, however, Markov random graph models are problematic models of real-world networks. A simple observation by Strauss (1986) that demonstrates the fundamental flaws of Markov random graph models is that, for any given pair of nodes $\{i, j\}$, the number of neighbors is $2(n-2)$ and thus increases with the number of nodes n. The large and growing neighborhoods indicate that Markov random graph models induce stronger and stronger dependence as n increases and are problematic when n is large. This has been confirmed by a growing body of theoretical and empirical results (Strauss, 1986; Jonasson, 1999; Snijders, 2002; Handcock, 2003; Hunter et al., 2008; Rinaldo et al., 2009; Schweinberger, 2011; Butts, 2011; Shalizi and Rinaldo, 2013).

2.3. Characterising local dependence

We take first steps to characterise local dependence, drawing inspiration from two sources: network science and probability theory. On the one hand, network science (e.g., Homans, 1950; Wasserman and Faust, 1994; Pattison and Robins, 2002) suggests that interactions in networks are local. On the other hand, weak dependence conditions in probability theory, such as mixing conditions (e.g. Billingsley, 1995; Dedecker et al., 2007), suggest that dependence should be local to ensure weak dependence and thus desirable behavior, such as Central Limit Theorems. The study of probability measures on infinite domains in physics (e.g., Georgii, 2011), spatial statistics (e.g., Cressie, 1993, Section 7.3.1, Stein, 1999, Chapter 3), machine learning and artificial intelligence (e.g., Singla and Domingos, 2007; Xiang and Neville, 2011) as well as the notion of M-dependence in time series (e.g., Billingsley, 1995, pp. 363–370) suggest each random variable should depend on a finite subset of other random variables. In other words, a natural starting point is to assume that each edge variable depends on a finite subset of other edge variables. We adapt the idea of finite neighborhoods and M-dependence to random graphs as follows.

Definition: **local dependence.** Let $\boldsymbol{Y}$ be a random graph with domain $\mathcal{D} = \mathcal{N} \times \mathcal{N}$ and sample space $\mathbb{Y}$, where $\mathcal{N}$ is a finite set of nodes. The dependence induced by a probability measure $\mathbb{P}$ on $\mathbb{Y}$ is called *local* if there exists a partition of the set of nodes $\mathcal{A}$ into $K \geq 2$ non-empty, finite subsets $\mathcal{A}_1, \ldots, \mathcal{A}_K$, called neighborhoods, such that the within- and between-neighborhood subgraphs $\boldsymbol{Y}_{k,l}$

with domains $\mathcal{A}_k \times \mathcal{A}_l$ and sample spaces $\mathbb{Y}_{k,l}$ satisfy, for all $\mathcal{Y} \subseteq \mathbb{Y}$ and $\mathcal{Y}_{k,l} \subseteq \mathbb{Y}_{k,l}$,

$$\mathbb{P}_K(\boldsymbol{Y} \in \mathcal{Y}) = \prod_{k=1}^{K} \mathbb{P}_{k,k}(\boldsymbol{Y}_{k,k} \in \mathcal{Y}_{k,k}) \prod_{l=1}^{k-1} \mathbb{P}_{k,l}(\boldsymbol{Y}_{k,l} \in \mathcal{Y}_{k,l}, \boldsymbol{Y}_{l,k} \in \mathcal{Y}_{l,k}), \tag{4}$$

where within-neighborhood probability measures $\mathbb{P}_{k,k}$ induce dependence within subgraphs $\boldsymbol{Y}_{k,k}$, while between-neighborhood probability measures $\mathbb{P}_{k,l}$ induce independence between subgraphs $\boldsymbol{Y}_{k,l}$, i.e., for all $\mathcal{Y}_{k,l} \subseteq \mathbb{Y}_{k,l}$, $\mathcal{Y}_{B,i,j} \subseteq \mathbb{Y}_{B,i,j}$ and $\mathcal{Y}_{B,j,i} \subseteq \mathbb{Y}_{B,j,i}$,

$$\mathbb{P}_{k,l}(\boldsymbol{Y}_{k,l} \in \mathcal{Y}_{k,l}, \boldsymbol{Y}_{l,k} \in \mathcal{Y}_{l,k}) = \prod_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} \mathbb{P}_{B,i,j}(Y_{B,i,j} \in \mathcal{Y}_{B,i,j}, Y_{B,j,i} \in \mathcal{Y}_{B,j,i}), \quad l < k, \tag{5}$$

where $Y_{B,i,j}$ denote edge variables corresponding to nodes i and j with samples spaces $\mathbb{Y}_{B,i,j}$ and $\mathbb{Y}_{B,j,i}$.□

Thus, local dependence breaks down the dependence of the random graph $\boldsymbol{Y}$ into dependence within subgraphs $\boldsymbol{Y}_{k,k}$. The construction of random graph models with local dependence is discussed in Section 3.

The first and foremost advantage of local dependence is that it makes no assumptions about the form and strength of dependence within subgraphs. Scientists are free to incorporate dependencies of interest, such as transitive closure within subgraphs. In contrast, conventional ERGMs (e.g., Frank and Strauss, 1986) induce unbounded neighborhoods and global dependence, as discussed in Section 2.2.

A second advantage is that local dependence endows models with desirable properties, which we discuss in Section 2.4.

2.4. Properties of local dependence

We show that random graphs with local dependence possess two natural properties. The first property is a domain consistency property that any probability model should possess. The second property is asymptotic Gaussian behavior of statistics of interest. The two properties help to address both problems of *specification* and *distribution* in the sense of Fisher (1922, p. 313) by allowing to model a wide range of dependencies within subgraphs and facilitating the derivation the distributions of estimators and goodness-of-fit statistics.

Since we are considering a single observation of a graph, we follow a domain-increasing approach to asymptotics that resembles the domain-increasing approach in spatial statistics (e.g., Cressie, 1993, Section 7.3.1, Stein, 1999, Chapter 3). Suppose that the domain of the random graph increases as follows. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of non-empty, finite sets of nodes and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ be a sequence of random graphs with increasing domain $\mathcal{N}_1 \times \mathcal{N}_1, \mathcal{N}_2 \times \mathcal{N}_2, \ldots$, where the set of nodes $\mathcal{N}_K = \bigcup_{k=1}^{K} \mathcal{A}_k$ is the union of the sets of nodes $\mathcal{A}_1, \ldots, \mathcal{A}_K$.

The first property is a domain consistency property that should be satisfied by any probability model (e.g., Billingsley, 1995, Section 36), but which many parametrisations of ERGMs do not satisfy (Shalizi and Rinaldo, 2013).

Theorem 1. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of non-empty, finite sets of nodes and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ be a sequence of random graphs with increasing domain $\mathcal{N}_1 \times \mathcal{N}_1, \mathcal{N}_2 \times \mathcal{N}_2, \ldots$, where $\mathcal{N}_K = \bigcup_{k=1}^{K} \mathcal{A}_k$. Let $\boldsymbol{Y}_{K+1\setminus K}$ be the random graph $\boldsymbol{Y}_{K+1}$ excluding $\boldsymbol{Y}_K$, i.e., $\boldsymbol{Y}_{K+1\setminus K}$ corresponds to the within-neighborhood subgraph $\boldsymbol{Y}_{K+1,K+1}$ and the between-neighborhood subgraphs $\boldsymbol{Y}_{k,K+1}$ and $\boldsymbol{Y}_{K+1,k}$, $k = 1, \ldots, K$; and let $\mathbb{Y}_{K+1\setminus K}$ be the sample space of $\boldsymbol{Y}_{K+1\setminus K}$. If a sequence of random graphs $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ satisfies local dependence, then it is *domain consistent* in the sense that, for all $K > 0$ and $\mathcal{Y}_K \subseteq \mathbb{Y}_K$,

$$\mathbb{P}_K(\boldsymbol{Y}_K \in \mathcal{Y}_K) = \mathbb{P}_{K+1}(\boldsymbol{Y}_K \in \mathcal{Y}_K, \boldsymbol{Y}_{K+1\setminus K} \in \mathbb{Y}_{K+1\setminus K}).□ \tag{6}$$

In other words, the probability measure $\mathbb{P}_K$ of random graph $\boldsymbol{Y}_K$ with domain $\mathcal{N}_K \times \mathcal{N}_K$ can be recovered from the probability measure $\mathbb{P}_{K+1}$ of random graph $\boldsymbol{Y}_{K+1}$ with domain $\mathcal{N}_{K+1} \times \mathcal{N}_{K+1}$ by marginalising with respect to $\boldsymbol{Y}_{K+1\setminus K}$. It is worth noting that the domain consistency condition considered here is weaker than the domain consistency condition considered by Shalizi and Rinaldo (2013) and is motivated by the way the domain of local random graphs increases.

In addition to domain consistency, it is desirable that random graphs with increasing domain satisfy sparsity. A random graph can be called *sparse* if the expected degrees $\mathbb{E}(\sum_j Y_{i,j})$ of nodes i are bounded, suggesting $\mathbb{E}(Y_{i,j}) \to 0$ as the number of nodes increases. The importance of sparsity has been recognised by social scientists, computer scientists, mathematicians (e.g., Jonasson, 1999, Lovász, 2012, p. viii and p. 4), and statisticians (e.g., Krivitsky et al., 2011; Vu et al., 2013). Sparsity embodies the notion that in the real world, resources are bounded—animals and humans face real-world constraints such as limited time and therefore cannot maintain arbitrarily many relationships. If random graphs with local dependence do not satisfy sparsity, then the expected degrees of nodes would be dominated by edges to nodes in other neighborhoods, which would be in conflict with the notion of local interaction in network science. We therefore focus on graphs where within-neighborhood subgraphs may be dense, but between-neighborhood subgraphs are sparse.

Definition: **sparsity.** Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of non-empty, finite sets of nodes and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ be a sequence of random graphs with increasing domain $\mathcal{N}_1 \times \mathcal{N}_1, \mathcal{N}_2 \times \mathcal{N}_2, \ldots$, where $\mathcal{N}_K = \bigcup_{k=1}^K \mathcal{A}_k$. A sequence of random graphs $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ satisfying local dependence is called δ-*sparse* if there exist constants $A > 0$ and $\delta > 0$ such that

$$\mathbb{E}(|Y_{B,i,j}|^p) \quad \leq \quad A\,n^{-\delta}, \quad p = 1, 2. \square \tag{7}$$

The second, and more important property is the fact that the asymptotic distribution of statistics of interest is Gaussian, which helps to address problems of estimation and goodness-of-fit. In practice, most statistics of interest are sums of subgraph configurations. Let $\mathcal{S}_K \subseteq \bigtimes_{i=1}^d \mathcal{N}_K$ be a subset of the d-dimensional Cartesian product of $\mathcal{N}_K$ with itself and $S_K : \mathcal{S}_K \mapsto \mathbb{R}$ be a real-valued function with domain $\mathcal{S}_K$, e.g., the number of edges or transitive triples. Such sums of subgraph configurations can be written as

$$S_K \quad = \quad \sum_{i \in \mathcal{S}_K} S_{K,i}, \tag{8}$$

where $S_{K,i} = \prod_{k=1}^q Y_{i,a_k,b_k}$ are interactions of q distinct edge variables Y_{i,a_k,b_k}, which resemble the interactions in undirected graphical models and related models in physics (e.g., Ising model) and spatial statistics (e.g., random fields). If, e.g., edge variables are binary and $\mathcal{S}_K \subseteq \mathcal{A}^3$, then $S_{K,i} = Y_{a,b}\,Y_{b,c}\,Y_{a,c}$ is an indicator of whether the triple of nodes (a, b, c) is transitive. The sum S_K can be decomposed into within- and between neighborhood-sums W_K and B_K:

$$S_K \quad = \quad W_K + B_K, \tag{9}$$

where $W_K = \sum_{k=1}^K W_{K,k}$ is the total within-neighborhood sum, composed of the within-neighborhood sums $W_{K,k} = \sum_{i \in \mathcal{S}_K} 1_{W,k,i} S_{K,i}$ in neighborhoods k, and $B_K = \sum_{i \in \mathcal{S}_K} 1_{B,i} S_{K,i}$ is the total between-neighborhood sum. The indicator function $1_{W,k,i}$ is 1 if subgraph configuration i involves nodes in neighborhood k and neighborhood k only and is 0 otherwise, whereas $1_{B,i}$ is 1 if $S_{K,i}$ involves nodes of more than one neighborhood and is 0 otherwise.

It turns out that sequences of local and sparse random graphs with increasing domain are well-behaved in the sense of satisfying a Central Limit Theorem for weakly dependent random variables.

Theorem 2. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of non-empty, finite sets of nodes and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ be a sequence of random graphs with increasing domain $\mathcal{N}_1 \times \mathcal{N}_1, \mathcal{N}_2 \times \mathcal{N}_2, \ldots$, where $\mathcal{N}_K = \bigcup_{k=1}^K \mathcal{A}_k$.

Consider sums of the form $S_K = \sum_{i \in \mathcal{S}_K} S_{K,i}$, where $\mathcal{S}_K \subseteq \times_{i=1}^d \mathcal{N}_K$ and $S_{K,i} = \prod_{k=1}^q Y_{i,a_k,b_k}$. Suppose that the edge variables $Y_{a,b}$ satisfy uniform boundedness in the sense that there exists a constant $C > 0$ such that, for all $K > 0$, $a \in \mathcal{N}_K$ and $b \in \mathcal{N}_K$, $\mathbb{P}(|Y_{a,b}| \leq C) = 1$. Without loss of generality, assume that, for all $K > 0$ and $i \in \mathcal{S}_K$, $\mathbb{E}(S_{K,i}) = 0$. If the sequence of random graphs $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ is local and $\delta > d$—sparse and $\mathbb{V}(W_K) \to \infty$ as $K \to \infty$, then

$$\lim_{K \longrightarrow \infty} \max_{1 \leq k \leq K} \mathbb{P}\left(|W_{K,k}| > \epsilon \sqrt{\mathbb{V}(W_K)} \right) = 0 \tag{10}$$

and

$$\frac{S_K}{\sqrt{\mathbb{V}(S_K)}} \xrightarrow{\mathrm{d}} N(0,1) \quad \text{as } K \longrightarrow \infty, \tag{11}$$

where $\mathbb{V}(W_K)$ and $\mathbb{V}(S_K)$ denote the variance of W_K and S_K, respectively.□

We discuss implications, starting with the most important one: The theorem respects the desiderata that random graphs be local and sparse and imposes no constraints on the form and shape of within-neighborhood probability measures, granting scientists complete freedom to specify arbitrary dependencies of interest within neighborhoods, such as transitive closure. At the same time, local and sparse random graphs tend to be well-behaved in the sense that neighborhoods cannot dominate the whole graph by (10) and the distribution of statistics, e.g., the number of transitive triples, tends to be Gaussian by (11) provided the number of neighborhoods K is large. As a result, random graph models with local dependence can be expected to place much probability mass around the expected values of statistics. If a graph is observed and the method of estimation (e.g., the method of maximum likelihood in exponential families) matches the expected and observed values of selected statistics, then the goodness-of-fit of the model with respect to the selected statistics can be expected to be acceptable.

Some additional remarks are in place. The uniform boundedness condition covers the most common cases, including the case of binary edge variables $Y_{i,j} \in \{0, 1\}$. Multivariate extensions of the Central Limit Theorem may be obtained by the Cramer-Wold theorem (e.g., Billingsley, 1995, p. 383). If suitable parametrisations of random graph models with local dependence are chosen, then the δ-method can be used to establish asymptotic normality of maximum likelihood estimators and test statistics along the lines of, e.g., DasGupta (2008, Section 16.3).

3. Model construction and parametrisations

To construct random graph models with local dependence, a suitable neighborhood structure is needed. In practice, suitable neighborhood structure may or may not be observed.

Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be membership indicators, where $\boldsymbol{Z}_i$ is the vector of membership indicators Z_{ik} of node i, where $Z_{ik} = 1$ if node i is member of neighborhood $\mathcal{A}_k$ and $Z_{ik} = 0$ otherwise. Since most network data are discrete, we consider throughout discrete network data and densities with respect to counting measure. We assume that the conditional probability mass function (PMF) of a random graph $\boldsymbol{Y}$ given a neighborhood structure $\boldsymbol{Z} = \boldsymbol{z}$ can be written as

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) &= \prod_{k=1}^K \mathbb{P}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k} \mid \boldsymbol{Z} = \boldsymbol{z}) \\
&\times \prod_{l=1}^{k-1} \mathbb{P}(\boldsymbol{Y}_{k,l} = \boldsymbol{y}_{k,l}, \boldsymbol{Y}_{l,k} = \boldsymbol{y}_{l,k} \mid \boldsymbol{Z} = \boldsymbol{z}),
\end{aligned} \tag{12}
$$

where between-neighborhood PMFs can be factorised into dyad-bound PMFs:

$$\mathbb{P}(\boldsymbol{Y}_{k,l} = \boldsymbol{y}_{k,l}, \boldsymbol{Y}_{l,k} = \boldsymbol{y}_{l,k} \mid \boldsymbol{Z} = \boldsymbol{z}) \quad = \quad \prod_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} \mathbb{P}(Y_{B,i,j} = y_{B,i,j}, Y_{B,j,i} = y_{B,j,i} \mid \boldsymbol{Z} = \boldsymbol{z}), \quad (13)$$

while the within-neighborhood PMFs are not assumed to be factorisable.

In practice, the question is the source of the neighborhood structure $\boldsymbol{Z}$. If suitable neighborhood structure were observed, then it should be used. We discuss model construction and two approaches to statistical inference given observed neighborhood structure in Section 3.1. An important practical problem is that in most applications no suitable neighborhood structure is observed. We discuss model construction and statistical inference in the absence of observed neighborhood structure in Section 3.2. Parametrisations are discussed in Section 3.3 and selected special cases in Section 3.4.

3.1. Model construction with observed neighborhoods

Consider the situation where a suitable neighborhood structure is observed and is used. It is worth noting that the theoretical results of Section 2.4 suggest that to be suitable, none of the neighborhoods should dominate the whole graph.

We consider two approaches to statistical inference given observed neighborhood structure $\boldsymbol{Z} = \boldsymbol{z}_{obs}$. Suppose that the conditional PMF (12) of random graph $\boldsymbol{Y}$ given observed neighborhood structure $\boldsymbol{Z} = \boldsymbol{z}_{obs}$ is parametrised by $\boldsymbol{\theta}$.

The first approach regards the observed neighborhood structure $\boldsymbol{z}_{obs}$ as fixed and bases statistical inference on the likelihood function

$$L(\boldsymbol{\theta}) \quad = \quad \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}_{obs}). \tag{14}$$

The second approach considers the observed neighborhood $\boldsymbol{z}_{obs}$ as the outcome of a random variable $\boldsymbol{Z}$ with a distribution $\mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}_{obs})$ parametrised by $\boldsymbol{\pi}$ and bases statistical inference on the likelihood function

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) \quad = \quad \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}_{obs}) \, \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}_{obs}). \tag{15}$$

If the two parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are variation-independent in the sense that the parameter space $\Omega_{\boldsymbol{\theta}, \boldsymbol{\pi}}$ is a product space of the form $\Omega_{\boldsymbol{\theta}, \boldsymbol{\pi}} = \Omega_{\boldsymbol{\theta}} \times \Omega_{\boldsymbol{\pi}}$, where $\Omega_{\boldsymbol{\theta}}$ is the parameter space of $\boldsymbol{\theta}$ and $\Omega_{\boldsymbol{\pi}}$ is the parameter space of $\boldsymbol{\pi}$, then the likelihood function is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) \quad = \quad \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}_{obs}) \, \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}_{obs}) \quad = \quad L(\boldsymbol{\theta}) \, L(\boldsymbol{\pi}), \tag{16}$$

where $L(\boldsymbol{\theta})$ is given by (14) and $L(\boldsymbol{\pi})$ is given by $L(\boldsymbol{\pi}) = \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}_{obs})$. Thus, if the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are variation-independent, then the likelihood function factorises and statistical inference for $\boldsymbol{\theta}$ can be based on $L(\boldsymbol{\theta})$. In other words, the two approaches are equivalent as long as the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are variation-independent. In general, the random-neighborhood approach may be more suitable than the fixed-neighborhood approach when it is believed that the neighborhoods are generated by a stochastic mechanism and that mechanism is itself of interest. The likelihood function $L(\boldsymbol{\theta})$ can be maximised by the maximum likelihood methods of Hunter and Handcock (2006) by using $\boldsymbol{z}_{obs}$ as a covariate and Bayesian inference can be conducted by using the Bayesian methods of Koskinen et al. (2010) and Caimo and Friel (2011).

There are multiple data structures which could be exploited to construct random graph models with local dependence. Strauss and Ikeda (1990) suggested to construct neighborhoods by exploiting the categories of categorical covariates, e.g., the two categories of the categorical covariate gender could be used to form two neighborhoods, corresponding to females and males. However, many categorical covariates are collected by surveys and have a small number of categories. The theoretical

results in Section 2.4 suggest that the number of neighborhoods should be large and none of the neighborhoods should dominate the graph, making categorical covariates with a small number of categories problematic when the number of nodes is large. A second approach exploits multilevel structure. In the health sciences and social sciences, many network data have a multilevel structure in the sense that subgraphs are nested in graphs, e.g., researchers are located in departments, departments are nested in buildings, and buildings belong to a campus. Such multilevel structure could be exploited to form the neighborhood structure. A third approach exploits spatial structure provided it is available, though spatial structure may not capture the whole dependence, as discussed in Section 2.2.

3.2. Model construction without observed neighborhoods

It is common that no suitable neighborhood structure is observed. In such cases, we follow a Bayesian approach and express the uncertainty about the neighborhood structure by a prior on a set of suitable neighborhood structures. We consider both hierarchical parametric and non-parametric priors.

In principle, statistical inference could be based on the likelihood function

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) \quad = \quad \sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) \, \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}), \tag{17}$$

where $\mathcal{Z}$ is the sample space of $\boldsymbol{Z}$. The difficulty is that $\mathcal{Z}$ is either a finite, but large set with $\exp(n \log K)$ elements—provided the number of neighborhoods K is fixed and known—or a countably infinite set, and in general the sum cannot be computed by complete enumeration. To facilitate statistical inference, we augment the observed data $\boldsymbol{Y}$ by the unobserved data $\boldsymbol{Z}$ and exploit hierarchical parametric and non-parametric models.

To be specific, assume

$$\boldsymbol{Z}_i \mid \pi_1, \ldots, \pi_K \quad \overset{\text{iid}}{\sim} \quad \text{Multinomial}(1; \pi_1, \ldots, \pi_K), \; i = 1, \ldots, n \tag{18}$$

is the distribution of membership vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. We note that one could incorporate predictors of memberships by using multinomial logit or probit link functions along the lines of Tallberg (2005).

A parametric approach could be based on Dirichlet priors:

$$\pi_1, \ldots, \pi_K \quad \sim \quad \text{Dirichlet}(\omega_1, \ldots, \omega_K). \tag{19}$$

A potential problem with Dirichlet priors is that the number of neighborhoods K must either be known or selected by model selection methods, which is not straightforward. An alternative is to express the uncertainty about K by specifying a prior for K (e.g., Richardson and Green, 1997), which leads to complicated Markov chain Monte Carlo algorithms. We follow a non-parametric approach based on stick-breaking priors (Ishwaran and James, 2001), which sidesteps both model selection and complicated Markov chain Monte Carlo algorithms. It allows the number of non-empty neighborhoods a posteriori to be large, while encouraging it a priori to be small. Suppose that there is an infinite number of neighborhoods and that nodes belong to neighborhood $k = 1, 2, \ldots$ with probability π_k, $k = 1, 2, \ldots$, where

$$\pi_1 \quad = \quad V_1 \tag{20}$$

$$\pi_k \quad = \quad V_k \prod_{j=1}^{k-1}(1 - V_j), \; k = 2, 3, \ldots, \tag{21}$$

where

$$V_k \mid \alpha \quad \overset{\text{iid}}{\sim} \quad \text{Beta}(1, \alpha), \; k = 1, 2, \ldots, \tag{22}$$

where $\alpha > 0$ is a parameter and $\sum_{k=1}^{\infty} \pi_k = 1$ with probability 1 (Ishwaran and James, 2001).

3.3. Parametrisations

Exponential parametrisations of the conditional PMF (12) are convenient, though other parametrisations may be used as well.

The dyad-bound between-neighborhood PMFs can be written as

$$\mathbb{P}_{\boldsymbol{\theta}}(Y_{B,i,j} = y_{B,i,j}, Y_{B,j,i} = y_{B,j,i} \mid \boldsymbol{Z} = \boldsymbol{z}) \quad = \quad \exp[\langle \boldsymbol{\theta}_B, s_{B,i,j}(y_{B,i,j}, y_{B,j,i}) \rangle - \psi_{B,i,j}(\boldsymbol{\theta}_B)], \quad (23)$$

where $s_{B,i,j}(y_{B,i,j}, y_{B,j,i})$ is a vector of between-neighborhood sufficient statistics, $\boldsymbol{\theta}_B$ is a vector of between-neighborhood natural parameters, and $\psi_{B,i,j}(\boldsymbol{\theta}_B)$ is the between-neighborhood log normalising constant,

$$\psi_{B,i,j}(\boldsymbol{\theta}_B) \quad = \quad \log \sum_{y'_{B,i,j} \in \mathbb{Y}_{B,i,j}} \sum_{y'_{B,j,i} \in \mathbb{Y}_{B,j,i}} \exp[\langle \boldsymbol{\theta}_B, s_{B,i,j}(y'_{B,i,j}, y'_{B,j,i}) \rangle]. \tag{24}$$

The between-neighborhood sufficient statistics $s_{B,i,j}(y_{B,i,j}, s_{B,j,i})$ may be functions of edges $y_{B,i,j}$ and $y_{B,j,i}$ and covariates. In the interest of model parsimony, we assume that the between-neighborhood parameter vector $\boldsymbol{\theta}_B$ is constant across dyads.

The within-neighborhood PMFs can be written as

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k} \mid \boldsymbol{Z} = \boldsymbol{z}) \quad = \quad \exp[\langle \boldsymbol{\theta}_{W,k}, s_{W,k}(\boldsymbol{y}_{k,k}) \rangle - \psi_{W,k}(\boldsymbol{\theta}_{W,k})], \tag{25}$$

where $s_{W,k}(\boldsymbol{y}_{k,k})$ is a vector of within-neighborhood sufficient statistics, $\boldsymbol{\theta}_{W,k}$ is a vector of within-neighborhood natural parameters, and $\psi_{W,k}(\boldsymbol{\theta}_{W,k})$ is the within-neighborhood log normalising constant,

$$\psi_{W,k}(\boldsymbol{\theta}_{W,k}) \quad = \quad \log \sum_{\boldsymbol{y}'_{k,k} \in \mathbb{Y}_{k,k}} \exp[\langle \boldsymbol{\theta}_{W,k}, s_{W,k}(\boldsymbol{y}'_{k,k}) \rangle]. \tag{26}$$

The within-neighborhood sufficient statistics $s_{W,k}(\boldsymbol{y}_{k,k})$ may include interactions, such as the number of triangles within neighborhood k, which induce dependence within neighborhoods. In addition, covariates can be used.

The exponential parametrisation of the between- and within-neighborhood PMFs implies that the conditional PMF of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ can be written as

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) \quad = \quad \exp\left[\langle \boldsymbol{\eta}(\boldsymbol{\theta}), s(\boldsymbol{y}) \rangle - \psi(\boldsymbol{\theta})\right], \tag{27}$$

where the vector of parameters $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a linear function of the vectors of between- and within-neighborhood parameters, the vector of sufficient statistics $s(\boldsymbol{y})$ is a linear function of between- and within-neighborhood vectors of sufficient statistics, and the log normalising constant $\psi(\boldsymbol{\theta})$ is given by

$$\psi(\boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{K} \sum_{l=1}^{k-1} \sum_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} \psi_{B,i,j}(\boldsymbol{\theta}_B) + \sum_{k=1}^{K} \psi_{W,k}(\boldsymbol{\theta}_{W,k}). \tag{28}$$

The between- and within-neighborhood parameter vectors $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_{W,k}$ index exponential families and therefore conjugate priors exist, though direct sampling from the resulting full conditional distributions is infeasible. In the absence of computational advantages, multivariate Gaussian priors are convenient alternatives:

$$\begin{aligned} \boldsymbol{\theta}_B \mid \boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B^{-1} \quad &\sim \quad \mathrm{MVN}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B^{-1}) \\ \boldsymbol{\theta}_{W,k} \mid \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1} \quad &\overset{\text{iid}}{\sim} \quad \mathrm{MVN}(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}), \ k = 1, 2, \ldots, \end{aligned} \tag{29}$$

where $\boldsymbol{\mu}_B$ and $\boldsymbol{\mu}_W$ are mean parameter vectors and $\boldsymbol{\Sigma}_B^{-1}$ and $\boldsymbol{\Sigma}_W^{-1}$ are precision matrices of suitable order.

To acknowledge the uncertainty about the hyper-parameters α, $\boldsymbol{\mu}_W$, and $\boldsymbol{\Sigma}_W^{-1}$, we assign conjugate Gamma, multivariate Gaussian, and Wishart hyper-priors to α, $\boldsymbol{\mu}_W$, and $\boldsymbol{\Sigma}_W^{-1}$, respectively.

3.4. Special cases and related models

Special cases of interest are the block models of Wang and Wong (1987), the stochastic block models of Nowicki and Snijders (2001), and the mixed membership models of Airoldi et al. (2008). These models assume that edge variables are independent conditional on an observed or unobserved partition of the set of nodes into subsets, which are called blocks and correspond to neighborhoods. These models satisfy local dependence, but do not allow scientists to directly specify the nature of interactions of interest, as discussed in Section 2.1.

The models of Strauss and Ikeda (1990) were discussed in Section 3.1. The usefulness of the models is limited, because neighborhood structure may either not be observed, in which case the models cannot be used, or neighborhood structure is observed but is unsuitable in the sense that the observed number of neighborhoods is small and the number of nodes is large.

Last, the model of Koskinen (2009) does not restrict dependence to blocks and therefore does not satisfy local dependence.

4. Bayesian inference

We focus on Bayesian inference without observed neighborhood structure, which is the most common and most challenging case. A Bayesian approach must overcome multiple obstacles. The most serious obstacle is the fact that with positive probability one or more neighborhoods k contains $n_k \gg 5$ nodes and thus one or more within-neighborhood log normalising constants, which are log sums of $\exp[\binom{n_k}{2} \log 2]$ terms (see (26)), is intractable. To facilitate posterior computations, we approximate the prior and augment the posterior.

We describe the approximation of the prior in Section 4.1; discuss the augmentation of the posterior and sampling from the augmented posterior in Section 4.2, with additional details in Supplements A and B; and address the non-identifiability of within-neighborhood parameter vectors and membership indicators in Supplement C.

4.1. Prior truncation

The stick-breaking prior of Section 3.2 can be approximated by a truncated stick-breaking prior along the lines of Ishwaran and James (2001), which facilitates posterior computations.

We choose a maximum number of neighborhoods, denoted by K_{max}. Some general advice concerning the choice of K_{max} is given by Ishwaran and James (2001). We are here more concerned with the goodness of fit of the model than the approximation of the stick-breaking prior and choose K_{max} in accordance. In practice, we choose K_{max} by either I. trying out multiple values of K_{max} and comparing the goodness of fit of the model; II. exploiting on-the-ground knowledge; or III. setting $K_{max} = n$. Strategy I is motivated by the fact that model estimation is time-consuming and the computing time increases with K_{max}, thus there is an incentive to choose $K_{\max}$ as small as possible. We demonstrate strategies I and II in Section 5.3.

Given K_{max}, the membership probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K_{max}})$ are constructed by truncated stick-breaking (Ishwaran and James, 2001):

$$\pi_1 = V_1 \tag{30}$$

$$\pi_k = V_k \prod_{j=1}^{k-1} (1 - V_j), \ k = 2, \ldots, K_{max}, \tag{31}$$

where

$$V_k \mid \alpha \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \ k = 1, \ldots, K_{max} - 1$$
$$V_{K_{max}} = 1, \tag{32}$$

where $\alpha > 0$ is a parameter and $V_{K_{max}} = 1$ ensures $\sum_{k=1}^{K_{max}} \pi_k = 1$. The truncated stick-breaking construction of $\boldsymbol{\pi}$ implies that $\boldsymbol{\pi}$ is generalised Dirichlet distributed, which is conjugate to multinomial sampling (Ishwaran and James, 2001).

4.2. Posterior augmentation

Under the truncated prior described in Section 4.1, the posterior is of the form

$$
\begin{aligned}
p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z} \mid \boldsymbol{y}) \quad &\propto \quad p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W) \\
&\times \quad \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}) \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}),
\end{aligned}
\tag{33}
$$

where the truncated prior is of the form

$$
\begin{aligned}
p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W) \quad &= \quad p(\alpha) \, p(\boldsymbol{\mu}_W) \, p(\boldsymbol{\Sigma}_W^{-1}) \, p(\boldsymbol{\pi} \mid \alpha) \, p(\boldsymbol{\theta}_B) \\
&\times \quad \prod_{k=1}^{K_{max}} p(\boldsymbol{\theta}_{W,k} \mid \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}),
\end{aligned}
\tag{34}
$$

where $\boldsymbol{\theta}_W = (\boldsymbol{\theta}_{W,1}, \ldots, \boldsymbol{\theta}_{W,K_{max}})$ denotes the within-neighborhood parameter vectors.

Owing to the fact that the conditional PMF of $\boldsymbol{Y}$ is not, in general, tractable, the posterior is doubly intractable, implying that standard Markov chain Monte Carlo methods (e.g., Metropolis-Hastings) cannot be used to sample from the posterior. Auxiliary-variable Markov chain Monte Carlo methods for sampling from doubly intractable posteriors arising in complete-data problems were introduced by Møller et al. (2006) and extended by Murray et al. (2006) and Liang (2010); and adapted to networks by Koskinen et al. (2010); Caimo and Friel (2011); and Wang and Atchade (2014). We extend them from the complete-data problems considered there to the incomplete-data problem considered here.

To facilitate posterior computations, we augment α, $\boldsymbol{\mu}_W$, $\boldsymbol{\Sigma}_W^{-1}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}_B$, $\boldsymbol{\theta}_W$, $\boldsymbol{Z}$, and $\boldsymbol{Y}$ by auxiliary variables $\boldsymbol{\theta}_W^\star$, $\boldsymbol{Z}^\star$, and $\boldsymbol{Y}^\star$. The auxiliary variable $\boldsymbol{Y}^\star$ can be interpreted as an auxiliary random graph, $\boldsymbol{Z}^\star$ can be interpreted as an auxiliary neighborhood structure, and $\boldsymbol{\theta}_W^\star$ can be interpreted as auxiliary within-neighborhood parameter vectors. We assume that the joint distribution of α, $\boldsymbol{\mu}_W$, $\boldsymbol{\Sigma}_W^{-1}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}_B$, $\boldsymbol{\theta}_W$, $\boldsymbol{Z}$, $\boldsymbol{Y}$, $\boldsymbol{\theta}_W^\star$, $\boldsymbol{Z}^\star$, and $\boldsymbol{Y}^\star$ is of the form

$$
\begin{aligned}
p(\alpha, &\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star, \boldsymbol{y}^\star) \\
&= p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W) \, \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}) \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) \\
&\times q(\boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star \mid \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{y}) \, \mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z}^\star),
\end{aligned}
\tag{35}
$$

where $q(\boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star \mid \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{y})$ is a suitable, auxiliary distribution, the conditional distributions $\boldsymbol{Y}$ and $\boldsymbol{Y}^\star$ belong to the same exponential family of distributions, and $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_B, \boldsymbol{\theta}_W^\star)$. The augmented posterior is of the form

$$
\begin{aligned}
p(\alpha, &\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star, \boldsymbol{y}^\star \mid \boldsymbol{y}) \\
&\propto p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star, \boldsymbol{y}^\star).
\end{aligned}
\tag{36}
$$

Integrating out the auxiliary variables $\boldsymbol{\theta}_W^\star$, $\boldsymbol{Z}^\star$, and $\boldsymbol{Y}^\star$ results in the posterior of α, $\boldsymbol{\mu}_W$, $\boldsymbol{\Sigma}_W^{-1}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}_B$, $\boldsymbol{\theta}_W$, and $\boldsymbol{Z}$. While sampling from the posterior (33) is infeasible, sampling from the augmented posterior (36) and integrating out the auxiliary variables $\boldsymbol{\theta}_W^\star$, $\boldsymbol{Z}^\star$, and $\boldsymbol{Y}^\star$ turns out to be feasible. We discuss Markov chain Monte Carlo steps and improved Markov chain Monte Carlo through variational methods in the supplement.

Fig. 2. Prior predictions of the number of edges (left) and triangles (right) under the local triangle model with parametric prior and $K = 150$ neighborhoods and $N = 499{,}500$ edge variables. Most prior predictive mass is concentrated around the means of the prior predictions, which are indicated by red, vertical lines.
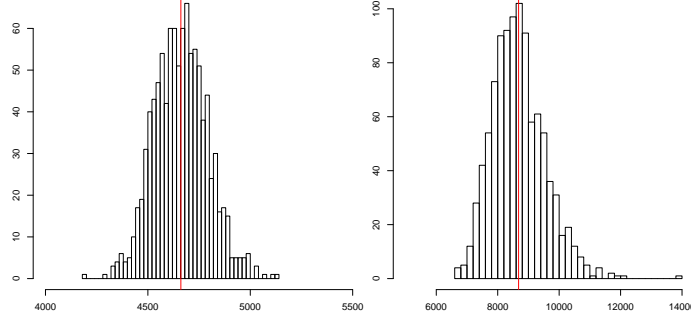
5. Assessment of local and global models of transitive closure

We compare random graph models with local and global dependence by comparing

- Prior predictions of graphs in order to assess whether models are able to produce data that resemble real-world networks (Section 5.1).

- Sampling distributions of Bayesian point and interval estimators (Section 5.2).

- Posterior predictions of graphs in order to assess whether models make sense in the light of observed data (Section 5.3).

We consider undirected, binary edge variables, i.e., $Y_{i,j} \in \{0, 1\}$ and $Y_{i,j} = Y_{j,i}$ with probability 1, and random graph models capturing transitive closure, because it is one of the most fundamental and problematic forms of dependence. A well-known random graph capturing transitive closure is the triangle model, which is an ERGM of the form (1) with the number of edges $y_{i,j}$ and triangles $y_{i,j}\, y_{j,h}\, y_{i,h}$ as sufficient statistics (Jonasson, 1999; Handcock et al., 2008). Its natural relative is the random graph model of the form (23) and (25), where the between-neighborhood sufficient statistics are the edges $y_{i,j}$ between nodes i, j in neighborhoods k, l, and the within-neighborhood sufficient statistics are the number of edges $y_{i,j}$ and triangles $y_{i,j}\, y_{j,h}\, y_{i,h}$ within neighborhood k. We refer to the two models as the global and local triangle model, respectively. The name global triangle model is motivated by the fact that the model is a special case of a Markov random graph model with unbounded neighborhoods—as explained in Section 2.2—and therefore does not satisfy local dependence.

5.1. Comparison of local and global models of transitive closure

In this subsection, we assess whether models are realistic by considering the prior predictive distributions of the statistics. As discussed in Section 1, the global triangle model places much probability mass on the relative boundary of the convex hull of $\{s(\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{Y}\}$. As networks with extreme values on network statistics may be considered odd, one approach to assessing the realism of a model is to consider the distribution of the statistics it produces.

The prior predictive distribution under the global triangle model can be written as

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) \;\; = \;\; \int p(\boldsymbol{\theta})\, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})\, \mathrm{d}\,\boldsymbol{\theta}, \tag{37}$$

Fig. 3. Sampling distributions of posterior medians of within-neighborhood means $\mu_{W,1}$ (left) and $\mu_{W,2}$ (right) under the local triangle model with parametric prior and $K = 7$ neighborhoods and $N = 1,225$ edge variables, where data-generating values are indicated by red, vertical lines. Despite the small number of $K = 7$ neighborhoods, the data-generating values of the Gaussian mean μ_W of the $K = 7$ within-neighborhood parameters $\theta_{W,1}, \ldots, \theta_{W,7}$ can be recovered.
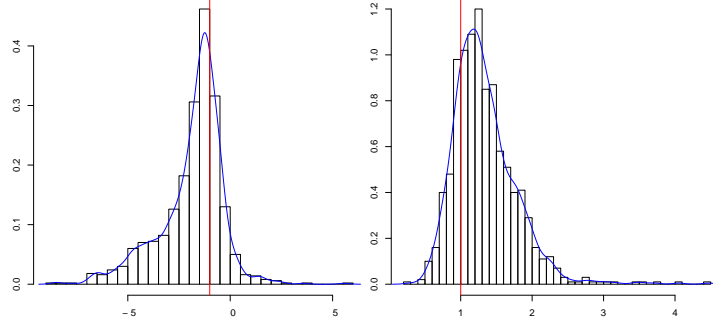
where $p(\boldsymbol{\theta})$ denotes the prior. Based on experience, values of θ_1 outside of $(-5,0)$ and values of θ_2 outside of $(0,5)$ index near-degenerate distributions. Therefore, we choose independent, uniform priors given by $\theta_1 \sim \text{Uniform}(-5,0)$ and $\theta_2 \sim \text{Uniform}(0,5)$.

The prior predictive distribution under the local triangle model can be written as

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \int \cdots \int \sum_{\boldsymbol{z} \in \mathcal{Z}} p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W) \tag{38}$$
$$\times \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}) \, \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}) \, \mathrm{d}\,\alpha \, \mathrm{d}\,\boldsymbol{\mu}_W \, \mathrm{d}\,\boldsymbol{\Sigma}_W^{-1} \, \mathrm{d}\,\boldsymbol{\pi} \, \mathrm{d}\,\boldsymbol{\theta}_B \, \mathrm{d}\,\boldsymbol{\theta}_W,$$

where $p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W)$ denotes the prior. We assign independent Gaussian priors with means -1 and $+1$ and standard deviations .25 and .25 to within-neighborhood parameters $\theta_{W,k,1}$ and $\theta_{W,k,2}$, respectively. The marginal priors of $\theta_{W,k,1}$ and $\theta_{W,k,2}$ ensure that most of the prior probability mass of $\theta_{W,k,1}$ is concentrated on $(-2,0)$ and most of the prior probability mass of $\theta_{W,k,2}$ is concentrated on $(0,2)$, which covers the most reasonable value of $\theta_{W,k,1}$ and $\theta_{W,k,2}$, respectively. To respect the sparse nature of graphs, we assume that the between-neighborhood parameter θ_B is governed by a Gaussian prior with mean μ_B and standard deviation 1, where the Gaussian is centered at $\mu_B = 3/(n-1)$, the value of θ_B under which the expected number of edges of nodes between neighborhoods is at most 3. The prior of $\boldsymbol{\pi}$ is given by the $\text{Dirichlet}(10, \ldots, 10)$ prior.

We generated 1,000 model predictions from the local triangle model with $K = 150$ neighborhoods, $n = 1,000$ nodes, and $N = 499,500$ edge variables and 1,000 realisations from the global triangle model with $n = 100$ nodes and $N = 4,950$ edge variables; the difference in the size of the graphs is due to the fact that the flawed nature of the global triangle model makes it infeasible to sample much larger graphs than $N = 4950$. Monte Carlo samples of size 1,000 were generated from the prior and, for every one of the draws from the prior, a prediction was generated by a Markov chain of length 10,000,000 for the local triangle model and 100,000 for the global triangle model, accepting the final draw of the Markov chain as a draw from the prior predictive distribution; the sample size is proportional to the size of the graphs.

Figure 1 shows prior predictions of the number of edges and triangles under the global triangle model. The bulk of the prior predictive mass is placed on extreme graphs with few edges and triangles and graphs with almost all possible edges and triangles. We note that the behavior of the global triangle model tends to deteriorate as the size of the graph increases, as discussed in Section 1. In contrast, Figure 2 demonstrates that random graph models with local dependence

Fig. 4. Sampling distributions of posterior medians of within-neighborhood means $\mu_{W,1}$ (left) and $\mu_{W,2}$ (right) under the local triangle model with non-parametric prior and $K = K_{\max} = 7$ neighborhoods and $N = 1,225$ edge variables, where data-generating values are indicated by red, vertical lines. Despite the small number of neighborhoods $K = 7$, the data-generating values of the Gaussian mean μ_W of the $K = 7$ within-neighborhood parameters $\theta_{W,1}, \ldots, \theta_{W,7}$ can be recovered.
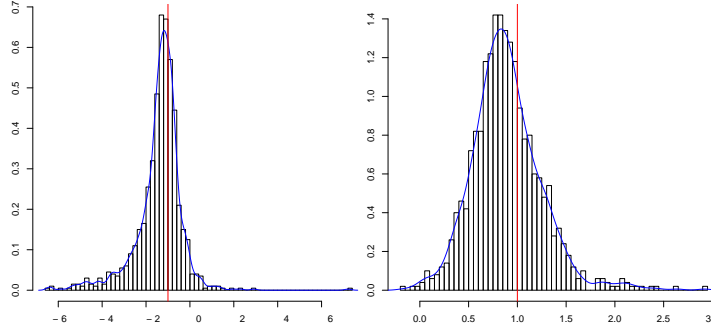
Table 1. Frequentist properties of posterior medians and 95%-posterior credibility intervals of within-neighborhood means $\mu_{W,1}$ and $\mu_{W,2}$. Despite the small number of $K = 7$ neighborhoods, the data-generating values of the Gaussian mean μ_W of the $K = 7$ within-neighborhood parameters $\theta_{W,1}, \ldots, \theta_{W,7}$ can be recovered.

prior	parameter	.025 quantile	.50 quantile	.975 quantile	coverage
parametric	$\mu_{W,1} = -1$	-5.75	-1.48	.48	86%
parametric	$\mu_{W,2} = 1$	.69	1.26	2.27	98%
non-parametric	$\mu_{W,1} = -1$	-4.19	-1.25	.26	95%
non-parametric	$\mu_{W,2} = 1$	.26	.86	1.68	$> 99\%$

place much prior predictive mass around the mean and all of its mass on graphs which resemble real-world networks: i.e., graphs where the average number of edges of nodes ranges from 4 to 6 and where the number of triangles is a small multiple of the number of edges. It is worth repeating that the number of edge variables is 499,500, which demonstrates that random graph models with local dependence are well-behaved when the number of edge variables is large, in sharp contrast to conventional ERGMs.

In short, the model predictions confirm what the theoretical results of Section 2.4 suggested: In contrast to conventional ERGMs, random graph models with local dependence are capable of generating graphs which resemble real-world networks and can thus be recommended a priori as models of real-world networks.

5.2. Sampling distributions of Bayesian point and interval estimators

We shed light on the frequentist properties of estimators by simulation. We focus on random graph models with local dependence, because flawed models of the form (1) generate many graphs which fall onto the relative boundary of the convex hull of $\{s(y) : y \in \mathbb{Y}\}$ and make statistical inference problematic (e.g., Barndorff-Nielsen, 1978, p. 151, Handcock, 2003, Rinaldo et al., 2009, Koskinen et al., 2010, Bhamidi et al., 2011).

Here, we focus on the frequentist properties of posterior point estimators and interval estimators of the local triangle model with $K = 7$ neighborhoods, $n = 50$ nodes, and $N = 1,225$ edge variables. We generated 1,000 graphs from the local triangle model using the same prior as used in Section

5.1. To infer from the simulated graphs to the data-generating values of the parameters, we used two priors: a parametric Dirichlet$(\alpha, \ldots, \alpha)$ prior for $\boldsymbol{\pi}$ with a Gamma$(1, .1)$ hyper-prior for α; and a non-parametric a truncated stick-breaking prior with $K_{\max} = 7$ with a Gamma$(1, .1)$ hyper-prior for α. In both cases, the marginal priors of $\theta_{W,1}$, $\theta_{W,2}$, and θ_B are independently $N(0, 100)$. We construct 1,000 Markov chains with 100,000 iterations, discarding the first 20,000 iterations as burn-in and recording every 10-th post-burn-in iterations.

In practice, within-neighborhood parameters are of primary interest, because it is the within-neighborhood models which capture the dependencies of interest. Here, we focus on the within-neighborhood means. The sampling distributions of posterior point estimators for $\boldsymbol{\theta}_{W,1}$ and $\boldsymbol{\theta}_{W,2}$ under parametric and non-parametric priors are shown in Figures 3 and 4. Despite the small number of neighborhoods $K = 7$, the data-generating parameter of the Gaussian mean $\boldsymbol{\mu}_W$ of the $K = 7$ within-neighborhood parameters $\boldsymbol{\theta}_{W,1}, \ldots, \boldsymbol{\theta}_{W,7}$ can be recovered in the sense that the posterior median clusters around the data-generating parameter value. The distributions are somewhat asymmetric, which is not surprising considering the small number of $K = 7$ neighborhoods. Table 1 shows that interval estimators, e.g., 95% posterior credibility intervals have acceptable coverage properties, considering the small number of $K = 7$ neighborhoods.

The results also suggest that the non-parametric approach seems to outperform the parametric approach, at least in terms of coverage for $\mu_{W,1}$. In addition, we found in the applications in Section 5.3 that the hierarchical non-parametric prior is not overly sensitive to the choice of the hyper-parameters of the priors for α, $\boldsymbol{\mu}_W$, $\boldsymbol{\mu}_B$, $\boldsymbol{\Sigma}_W$, and $\boldsymbol{\Sigma}_B$, whereas the hierarchical parametric prior sometimes is sensitive to the choice of hyper-parameters, and more so when the specified number of neighborhoods exceeds the true number of neighborhoods. Thus, the non-parametric approach seems to have advantages over the parametric approach.

5.3. Application to a terrorist network and a social network

A natural approach to comparing ERGMs and random graph models with local dependence is based on their predictions about observable quantities. Hunter et al. (2008) argued that, in practice, it is imperative to generate model predictions to assess the goodness-of-fit of network models.

In this section, we compare ERGMs and random graph models with local dependence in terms of posterior predictions in two real-world networks with known ground truth: the terrorist network behind the Bali bombing in 2002 as well as social relations amongst novices within a novitiate.

The posterior predictive distribution under the global triangle model given data $\boldsymbol{y}$ can be written as

$$\mathbb{P}(\dot{\boldsymbol{Y}} = \dot{\boldsymbol{y}} \mid \boldsymbol{Y} = \boldsymbol{y}) \quad = \quad \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathbb{P}_{\boldsymbol{\theta}}(\dot{\boldsymbol{Y}} = \dot{\boldsymbol{y}}) \, \mathrm{d}\, \boldsymbol{\theta}, \tag{39}$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ denotes the posterior. The posterior predictive distribution under the local triangle model can be written as

$$\mathbb{P}(\dot{\boldsymbol{Y}} = \dot{\boldsymbol{y}} \mid \boldsymbol{Y} = \boldsymbol{y}) \quad = \quad \int \cdots \int \sum_{\boldsymbol{z} \in \mathcal{Z}} p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z} \mid \boldsymbol{y})$$
$$\times \quad \mathbb{P}_{\boldsymbol{\theta}}(\dot{\boldsymbol{Y}} = \dot{\boldsymbol{y}} \mid \boldsymbol{Z} = \boldsymbol{z}) \, \mathrm{d}\, \alpha \, \mathrm{d}\, \boldsymbol{\mu}_W \, \mathrm{d}\, \boldsymbol{\Sigma}_W^{-1} \, \mathrm{d}\, \boldsymbol{\pi} \, \mathrm{d}\, \boldsymbol{\theta}_B \, \mathrm{d}\, \boldsymbol{\theta}_W, \tag{40}$$

where $p(\alpha, \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z} \mid \boldsymbol{y})$ denotes the posterior. Independent priors $\theta_i \sim N(0, 25)$ are used in the case of the ERGM and independent priors $\alpha \sim$ Gamma$(1, 1)$, $\mu_{W,i} \sim N(0, 1)$, and $\sigma_{W,i}^{-2} \sim$ Gamma$(10, 10)$ in the case of the random graph model with local dependence. 120,000 draws from the posterior predictive distribution of the ERGM were generated by the Markov chain Monte Carlo algorithm of Caimo and Friel (2011), with a burn-in of 20,000 and saving every 10-th post-burn-in draw, and 1,200,000 draws from the posterior predictive distribution of the random

Fig. 5. Terrorist network behind Bali bombing in 2002 with $N = 136$ edge variables. The colored pie charts represent posterior membership probabilities. The clustering is a by-product of model estimation and is of secondary interest, but it is comforting that it is consistent with ground truth.
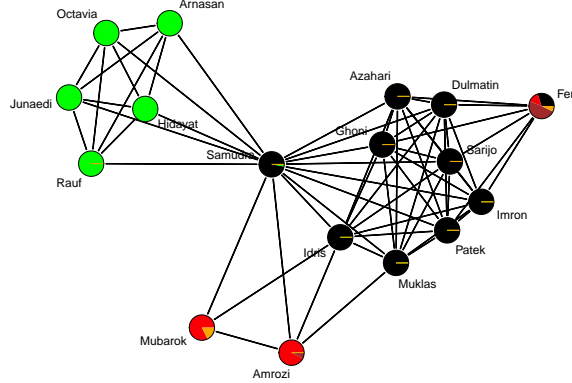
graph model with local dependence were generated by the Markov chain Monte Carlo algorithm of Section 4, with a burn-in of 200,000 and saving every 100-th post-burn-in draw.

5.3.1. Terrorist network behind Bali bombing in 2002

The structure of terrorist networks is of interest with a view to understanding how terrorists communicate, to identify cells (i.e., subsets of terrorists), to isolate cells, and to dismantle them. We consider here the network of terrorists behind the Bali, Indonesia, bombing in 2002, killing 202 (Koschade, 2006). The 17 terrorists who carried out the bombing were members of the Southeast Asian al-Qaeda affiliate Jemaah Islamiyah. The terrorist network can be represented by a graph with $n = 17$ nodes and $N = 136$ edge variables, where $Y_{i,j} = 1$ if terrorists i and j were in contact prior to the bombing and $Y_{i,j} = 0$ otherwise. The terrorist network is shown in Figure 5.

We start by determining the maximum number of neighborhoods K_{max} to truncate the prior. Using strategy I described in Section 4.1, we compare the local triangle model with $K_{max} = 2, 3, 4, 5$ neighborhoods in terms of predictive power. Predictive power is taken to be the root mean square deviation of the predicted number of triangles. According to the left panel of Figure 6, the local triangle model with $K_{max} = 2$ neighborhoods is far superior to the global triangle model, which corresponds to $K_{max} = 1$ neighborhood. The local triangle model with $K_{max} = 3$ in turn is superior to the local triangle model with $K_{max} = 2$, but increasing K_{max} from 3 to 5 does not increase the predictive power much. The right panel of Figure 6 compares the local triangle model to stochastic neighborhoods. The stochastic block model used here is a special case of the local triangle model where the within-neighborhood sufficient statistics are reduced to the number of edges, which induces conditional independence of edges within neighborhoods. Stochastic block models are special cases of random graph models with local dependence and not appealing when dependence is of substantive interest, because they assume conditional independence within neighborhoods, as discussed in Sections 2.1 and 3.4. The right panel of Figure 6 demonstrates that the stochastic block model has much lower predictive power than the local triangle model.

We compare the global and local triangle model with up to $K_{max} = 5$ neighborhoods in terms of the posterior predictive distribution of the number of edges and triangles, shown in Figures 7 and 8. Under the global triangle model, the posterior predictive distribution is bimodal. In contrast, the posterior predictive distribution under the local triangle model is unimodal and places

Fig. 6. Terrorist network: root mean square deviation of predicted number of triangles plotted against $K_{max} = 2, 3, 4, 5$ neighborhoods. Left: The global triangle model with $K_{\max} = 1$ is far inferior to the local triangle model with $K_{\max} = 2, 3, 4, 5$ neighborhoods. Right: The stochastic block model (line with 1's) is likewise far inferior to the local triangle model (line with 2's).
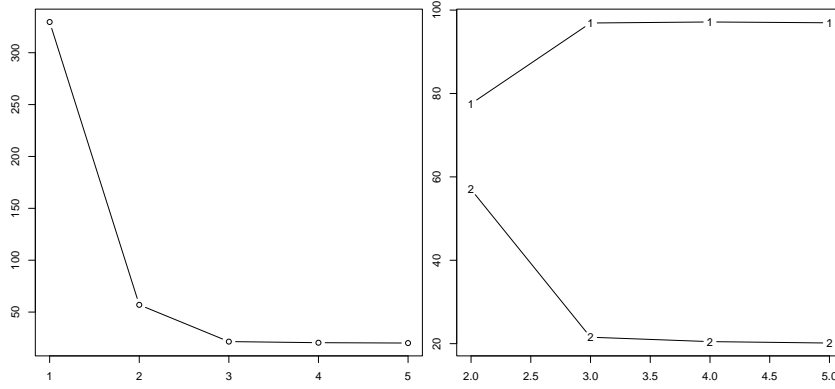
**Table 2.** Terrorist network: posterior of parameters α, $\mu_{W,1}$, and $\mu_{W,2}$

parameter	.05 quantile	.50 quantile	.95 quantile	odds of parameter being positive
α	.36	1.32	3.43	∞
$\mu_{W,1}$	-1.03	.45	2.00	2.22
$\mu_{W,2}$	-.27	.91	2.22	8.74

most mass on graphs which are close to the observed graph in terms of the number of edges and triangles. The fact that the global triangle model places so much mass on dense graphs with almost all edges and triangles indicates that the global triangle model fits much worse than the random graph model with local dependence, no matter which goodness of fit statistics are chosen, because the topology of graphs which are local in nature—such as the observed graph—stands in sharp contrast to the topology of dense graphs in terms of connectivity, centrality, transitivity, and other interesting features of graphs (e.g., Kolaczyk, 2009). We note that, while other statistics may be used to compare the two models in terms of goodness of fit, the choice of goodness of fit statistics here present compelling evidence.

The posterior of α, $\mu_{W,1}$, and $\mu_{W,2}$ is shown in Table 2. The mean parameters $\mu_{W,1}$ and $\mu_{W,2}$ governing the within-neighborhood parameters tend to be both positive—and more so the mean parameter $\mu_{W,2}$ governing the within-neighborhood triangle parameters—which is not surprising in the light of the large number of edges and triangles within neighborhoods.

Last, while the primary purpose of introducing neighborhoods is the desire to address the model degeneracy and striking lack of fit of ERGMs, predictions of the memberships to neighborhoods may be of interest as well, for example, to identify cells. The pie charts in Figure 5 represent the posterior membership probabilities reported by the stochastic relabeling algorithm described in Supplement C. The 5 green-colored terrorists turn out to be the 5 members of the so-called support group, which was to supposed to support the so-called main group consisting of all other terrorists. The members of the main group tend to be black-colored, with the exception of Amrozi and Mubarok who are more red-colored than black-colored. Indeed, while Amrozi and Mubarok belonged to the main group, both resided elsewhere and were almost isolated from the rest of the main group (Koschade, 2006). Most interesting is the membership of Feri. He was a member of the main group and was the suicide bomber who initiated the attack. Feri arrived two days before the attack, whereas all other members

Fig. 7. Terrorist network: posterior predictions of the number of edges (left) and triangles (right) under the global triangle model; vertical lines represent observed numbers. Although the number of edge variables $N = 136$ is not large, the polarisation is evident.
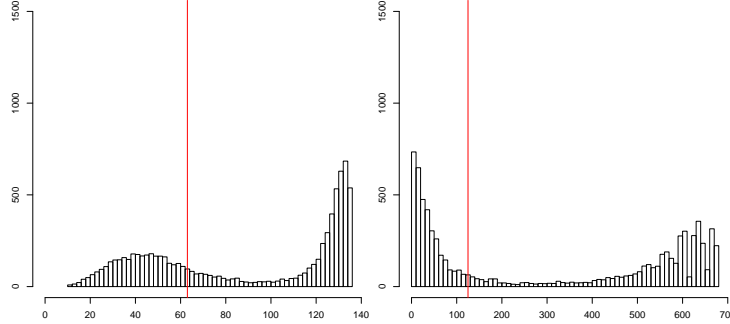
Fig. 8. Terrorist network: posterior predictions of the number of edges (left) and triangles (right) under the local triangle model; vertical lines represent observed numbers. The posterior predictive distributions are unimodal and short-tailed, in contrast to Figure 7.
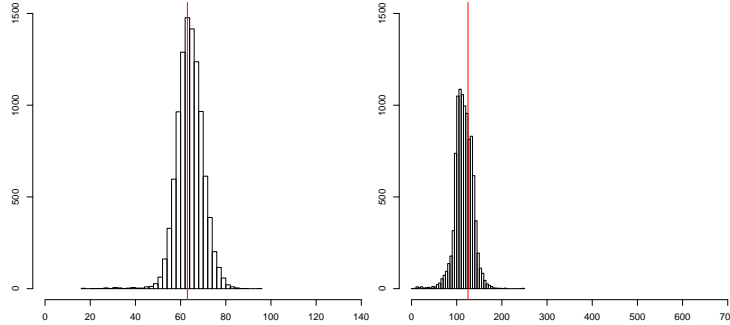
of the main group had arrived days or weeks earlier and in fact started leaving the night Feri arrived (Koschade, 2006). As a result, Feri had limited opportunities to communicate with others. In particular, Feri was the one and only member of the main group who did not communicate with the three commanders Muklas (the Jemaah Islamiyah head of operations in Singapore and Malaysia), Samudra (the field commander), and Idris (the logistics commander) (Koschade, 2006). Therefore, the network position of Feri is unique and the uncertainty about his membership is reflected in the posterior membership probability distribution.

In conclusion, random graph models with local dependence capture simple and interesting features of the terrorist network and, under the parametrisation considered here, posterior membership predictions are consistent with on-the-ground knowledge of the terrorist network.

5.3.2. Social relations within a novitiate

Sampson (1968) studied social relations among a group of novices who were preparing to enter a monastic order. The network is a classic data set in social network analysis (White et al., 1976; Handcock et al., 2007) and corresponds to $N = 306$ relationships among the $n = 18$ novices measured at three time points spreads our over a twelve month period. We consider here the following directed edge variables $Y_{i,j}$: If novice i liked novice j at any of the three time points, then $Y_{i,j} = 1$, otherwise $Y_{i,j} = 0$. The network is plotted in Figure 9.

A natural extension of the triangle model to directed graphs is given by a model of the form

Fig. 9. Sampson network with $N = 306$ edge variables. The colored pie charts represent posterior membership probabilities. The clustering is a by-product of model estimation and is of secondary interest, but it is comforting that it is consistent with ground truth.
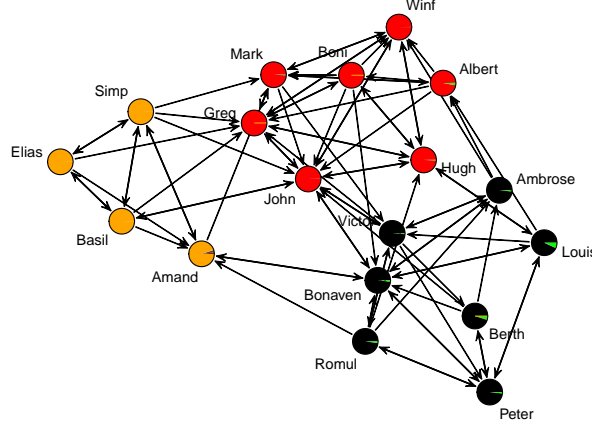
(1) with the number of edges $y_{i,j}$, mutual edges $y_{i,j}\,y_{j,i}$, and transitive triples $y_{i,j}\,y_{j,h}\,y_{i,h}$ as sufficient statistics. Its local relative is given by the random graph model (23) and (25) with the number of edges $y_{i,j}$ and mutual edges $y_{i,j}\,y_{j,i}$ between nodes i, j in neighborhoods k, l as between-neighborhood sufficient statistics, and the number of edges $y_{i,j}$, mutual edges $y_{i,j}\,y_{j,i}$, and transitive triples $y_{i,j}\,y_{j,h}\,y_{i,h}$ within neighborhood k as with-neighborhood sufficient statistics. Since experts argue that the novices are divided into 3 or 4 groups (White et al., 1976; Handcock et al., 2007), we follow strategy II described in Section 4.1 and set $K_{max} = 5$, which can be considered to be an upper bound on the number of neighborhoods.

Figures 10 and 11 show posterior predictions of the number of edges, mutual edges, and transitive triples. The contrast between the global and local triangle model in terms of goodness of fit is at least as striking as in the case of the terrorist network in Section 5.3.1.

The problematic nature of the global triangle model is underlined by the posterior of the number of non-empty neighborhoods of the random graph model with local dependence. Figure 12 shows that the posterior places negligible mass on partitions of the set of nodes where all nodes are assigned to one neighborhood, which corresponds to the global triangle model. In addition, the posterior mode is 3, which is in line with expert knowledge (White et al., 1976; Handcock et al., 2007).

The neighborhoods correspond, once again, to physical groups: The posterior membership probabilities shown in Figure 9 agree with the three-group division of novices into "Loyals," "Turks," and "Outcasts" advocated by most experts (White et al., 1976; Handcock et al., 2007).

6. Discussion

We have demonstrated that the notion of local dependence, as introduced here, endows models with desirable properties and also makes them amenable to statistical inference. Models with local dependence can be considered to be models of the *"next generation of social network models"* (Snijders, 2007, p. 324): i.e., models which combine latent structure models (e.g., Nowicki and Snijders, 2001; Hoff et al., 2002) and exponential-family random graph models (e.g., Frank and Strauss, 1986) in a way that takes advantage of the strengths of ERGMs—i.e., the power of ERGMs to model dependencies—while reducing the weaknesses of ERGMs—i.e., the fact that Markov dependence

Fig. 10. Sampson network: posterior predictions of the number of edges (left), mutual edges (middle), and transitive triples (right) under the global triangle model; vertical lines represent observed numbers. With $N = 306$ edge variables, the polarisation is more pronounced than in Figure 7 with $N = 136$ edge variables.
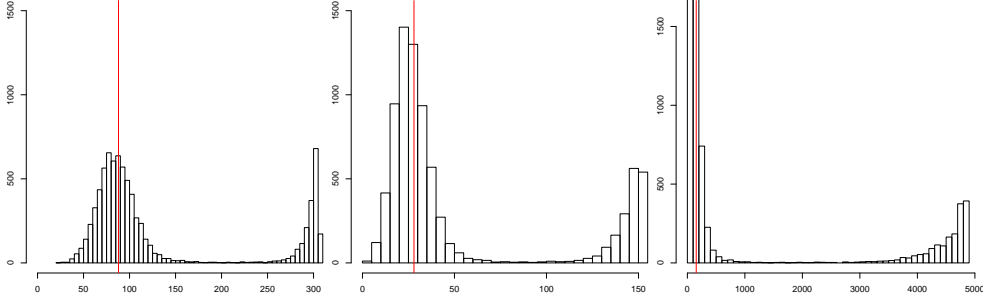
Fig. 11. Sampson network: posterior predictions of the number of edges (left), mutual edges (middle), and transitive triples (right) under the local triangle model; vertical lines represent observed numbers. The posterior predictive distributions are unimodal and short-tailed, in contrast to Figure 10.
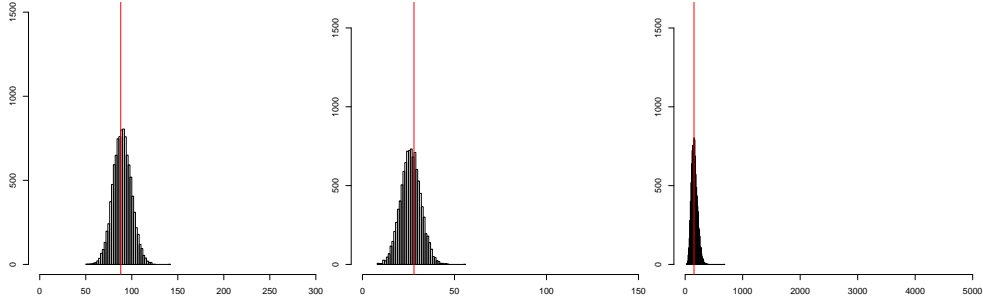
along the lines of Frank and Strauss (1986) is more global than local in nature and are not amenable to statistical inference; note that a partition of the set of nodes can be considered to constitute a latent, discrete space.

We believe that random graph models with local dependence constitute a promising and versatile approach to modeling real-world networks. Models with finite neighborhoods have been used in physics, machine learning, artificial intelligence, and spatial statistics with success, and so have models with M-dependence in time series. We believe that the notion of local dependence introduced here is a natural relative of the notions of local dependence in spatial statistics and time series, and as a result can be expected to be useful in applications.

The desirable properties of local dependence suggests that researchers should make every effort to identify and collect information on suitable neighborhood structures. If suitable neighborhood structures are not observed, then the auxiliary-variable Bayesian methods developed here can be used.

We have implemented statistical inference for random graph models with local dependence in the open-source R package `hergm`. It will be made publicly available on CRAN.

Acknowledgements

Fig. 12. Sampson network: posterior of number of non-empty neighborhoods under the local triangle model. The posterior is consistent with ground truth and confirms that the global triangle model (assuming that all nodes are in one neighborhood) makes no sense in the light of the observed data.
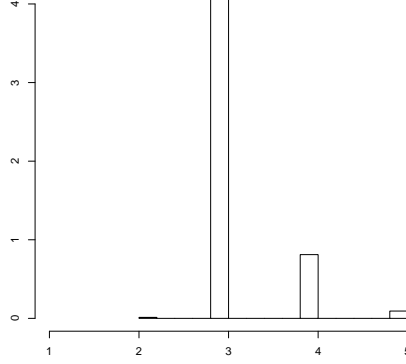
A. Proofs

Proof of Theorem 1. By local dependence, for all $K > 0$ and $\mathcal{Y}_K \subseteq \mathbb{Y}_K$,

$$\mathbb{P}_{K+1}(\boldsymbol{Y}_K \in \mathcal{Y}_K,\ \boldsymbol{Y}_{K+1 \setminus K} \in \mathbb{Y}_{K+1 \setminus K})$$

$$= \left[\prod_{k=1}^{K} \mathbb{P}_{k,k}(\boldsymbol{Y}_{k,k} \in \mathcal{Y}_{k,k}) \prod_{l=1}^{k-1} \mathbb{P}_{k,l}(\boldsymbol{Y}_{k,l} \in \mathcal{Y}_{k,l}, \boldsymbol{Y}_{l,k} \in \mathcal{Y}_{l,k}) \right]$$

$$\times \left[\mathbb{P}_{K+1,K+1}(\boldsymbol{Y}_{K+1,K+1} \in \mathbb{Y}_{K+1,K+1}) \prod_{l=1}^{K} \mathbb{P}_{K+1,l}(\boldsymbol{Y}_{k,l} \in \mathbb{Y}_{k,l}, \boldsymbol{Y}_{l,k} \in \mathbb{Y}_{l,k}) \right] \qquad (41)$$

$$= \prod_{k=1}^{K} \mathbb{P}_{k,k}(\boldsymbol{Y}_{k,k} \in \mathcal{Y}_{k,k}) \prod_{l=1}^{k-1} \mathbb{P}_{k,l}(\boldsymbol{Y}_{k,l} \in \mathcal{Y}_{k,l}, \boldsymbol{Y}_{l,k} \in \mathcal{Y}_{l,k})$$

$$= \mathbb{P}_K(\boldsymbol{Y}_K \in \mathcal{Y}_K). \square$$

Proof of Theorem 2. By uniform boundedness, $\mathbb{E}(|S_{K,i}|^p) \leq C^{pq}$, $p = 1, 2$. Therefore, we can assume $\mathbb{E}(S_{K,i}) = 0$ without loss of generality, which implies $\mathbb{E}(W_{K,k}) = 0$, $\mathbb{E}(W_K) = 0$, $\mathbb{E}(B_K) = 0$, and $\mathbb{E}(S_K) = 0$. The variance of S_K can be written as

$$\mathbb{V}(S_K) = \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} \mathbb{C}(S_{K,i}, S_{K,j}) \leq \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} |\mathbb{C}(S_{K,i}, S_{K,j})|$$

$$= \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} I_{W,k,i,j} |\mathbb{C}(S_{K,i}, S_{K,j})| + \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} I_{B,i,j} |\mathbb{C}(S_{K,i}, S_{K,j})|, \qquad (42)$$

where $\mathbb{C}(S_{K,i}, S_{K,j})$ denotes the covariance of $S_{K,i}$ and $S_{K,j}$ and $1_{W,k,i,j}$ indicates that both $S_{K,i}$ and $S_{K,j}$ are functions of edge variables in neighborhood k and neighborhood k only, whereas $1_{B,i,j}$

indicates that either $S_{K,i}$ or $S_{K,j}$ or both are functions of between-neighborhood edge variables. The within-neighborhood covariances are non-zero, but are bounded by uniform boundedness:

$$1_{W,k,i,j} \, |\mathbb{C}(S_{K,i}, S_{K,j})| \;\; \leq \;\; 1_{W,k,i,j} \, \sqrt{\mathbb{V}(S_{K,i})} \, \sqrt{\mathbb{V}(S_{K,j})} \;\; \leq \;\; 1_{W,k,i,j} \, C^{2q}. \tag{43}$$

Some of the between-neighborhood covariances may be non-zero as well, because some of the statistics $S_{K,i}$ and $S_{K,j}$ may share edge variables and may therefore be dependent. But $1_{B,i,j} = 1$ implies that either $S_{K,i}$ or $S_{K,j}$ or both are functions of at least one between-neighborhood edge variable. Without loss of generality, assume that Y_{i,a_1,b_1} is one of the between-neighborhood edge variables, and note that both $S_{K,i}$ and $S_{K,j}$ may be functions of Y_{i,a_1,b_1}. The between-neighborhood covariances can be written as

$$1_{B,i,j} \, |\mathbb{C}(S_{K,i}, S_{K,j})| \;\; = \;\; 1_{B,i,j} \, \left| \mathbb{E}\left(\prod_{k=1}^{q} Y_{i,a_k,b_k} \, Y_{j,a_k,b_k} \right) \right|. \tag{44}$$

The product $\prod_{k=1}^{q} Y_{i,a_k,b_k} \, Y_{j,a_k,b_k}$ can be written as $Y_{i,a_1,b_1}^{p} \, Y_{-i,a_1,b_1}$, where $p = 1, 2$ because $S_{K,i}$ and $S_{K,j}$ are functions of q distinct edge variables, and Y_{-i,a_1,b_1} is the product of $2q - p$ edge variables distinct from Y_{i,a_1,b_1}. By the independence of the between-neighborhood edge variable Y_{i,a_1,b_1} and uniform boundedness,

$$1_{B,i,j} \, \left| \mathbb{E}\left(\prod_{k=1}^{q} Y_{i,a_k,b_k} \, Y_{j,a_k,b_k} \right) \right| \;\; = \;\; 1_{B,i,j} \, \left| \mathbb{E}(Y_{i,a_1,b_1}^{p}) \, \mathbb{E}\left(Y_{-i,a_1,b_1} \right) \right|$$
$$\leq \;\; 1_{B,i,j} \, \mathbb{E}(|Y_{i,a_1,b_1}^{p}|) \, \mathbb{E}\left(|Y_{-i,a_1,b_1}| \right) \;\; \leq \;\; 1_{B,i,j} \, \mathbb{E}(|Y_{i,a_1,b_1}^{p}|) \, C^{2q-p}. \tag{45}$$

By sparsity, $\mathbb{E}(|Y_{i,a_1,b_1}|^{p}) \leq A \, n^{-\delta}$, $\delta > d$, $p = 1, 2$. Thus, by (44), (45), and sparsity, all covariances vanish in the limit, with the exception of within-neighborhood covariances. As a result,

$$\lim_{K \longrightarrow \infty} \mathbb{V}(S_K) \;\; = \;\; \lim_{K \longrightarrow \infty} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} 1_{W,k,i,j} \, \mathbb{C}(S_{K,i}, S_{K,j}) \;\; = \;\; \lim_{K \longrightarrow \infty} \sum_{k=1}^{K} \mathbb{V}(W_{K,k}) \tag{46}$$

$$\lim_{K \longrightarrow \infty} \mathbb{V}(W_K) \;\; = \;\; \lim_{K \longrightarrow \infty} \sum_{k=1}^{K} \mathbb{V}(W_{K,k}) \tag{47}$$

$$\lim_{K \longrightarrow \infty} \mathbb{V}(B_K) \;\; = \;\; 0. \tag{48}$$

Since the subsets of nodes $\mathcal{A}_k$ contain at most $M < \infty$ nodes and thus subsets $\mathcal{S}_{K,k} \subseteq \bigtimes_{k=1}^{d} \mathcal{A}_k$ contain at most $M^d < \infty$ elements, the within-neighborhood variances $\mathbb{V}(W_{K,k})$ are bounded:

$$\mathbb{V}(W_{K,k}) \;\; = \;\; \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} 1_{W,k,i,j} \, \mathbb{C}(S_{K,i}, S_{K,j}) \;\; \leq \;\; \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} 1_{W,k,i,j} \, |\mathbb{C}(S_{K,i}, S_{K,j})|$$
$$\leq \;\; \sum_{i \in \mathcal{S}_K} \sum_{j \in \mathcal{S}_K} 1_{W,k,i,j} \, C^{2q} \;\; \leq \;\; M^{2d} \, C^{2q}. \tag{49}$$

By uniform boundedness, (47), and (49), the within-neighborhood sums $W_{K,k}$ satisfy Lyaponouv's and thus Lindeberg's condition:

$$\lim_{K \longrightarrow \infty} \sum_{k=1}^{K} \frac{\mathbb{E}(|W_{K,k}|^4)}{[\mathbb{V}(W_K)]^2} \quad \leq \quad \lim_{K \longrightarrow \infty} \sum_{k=1}^{K} \frac{M^{2d} \, C^{2q} \, \mathbb{E}(|W_{K,k}|^2)}{[\mathbb{V}(W_K)]^2}$$

$$= \quad \frac{M^{2d} \, C^{2q} \lim\limits_{K \longrightarrow \infty} \sum\limits_{k=1}^{K} \mathbb{V}(W_{K,k})}{\lim\limits_{K \longrightarrow \infty} \mathbb{V}(W_K) \lim\limits_{K \longrightarrow \infty} \mathbb{V}(W_K)} \quad = \quad \frac{M^{2d} \, C^{2q} \lim\limits_{K \longrightarrow \infty} \mathbb{V}(W_K)}{\lim\limits_{K \longrightarrow \infty} \mathbb{V}(W_K) \lim\limits_{K \longrightarrow \infty} \mathbb{V}(W_K)} \quad = \quad 0. \tag{50}$$

By (50), the within-neighborhood sums $W_{K,k}$ satisfy the uniform asymptotic negligibility condition (10) (e.g., Resnick, 1999, p. 315). By (49), (50), and the Lindeberg-Feller Central Limit Theorem (e.g., Billingsley, 1995, p. 359, Theorem 27.2) applied to the double sequence of random variables $W_K = W_{K,1} + \cdots + W_{K,K}$, $K = 1, 2, \ldots$,

$$\frac{W_K}{\sqrt{\mathbb{V}(W_K)}} \quad \xrightarrow{\text{d}} \quad N(0,1) \quad \text{as } K \longrightarrow \infty. \tag{51}$$

By Chebyshev's inequality and (48), for all $\epsilon > 0$,

$$\lim_{K \longrightarrow \infty} \mathbb{P}(|B_K| > \epsilon) \quad \leq \quad \lim_{K \longrightarrow \infty} \frac{1}{\epsilon^2} \mathbb{V}(B_K) \quad = \quad 0, \tag{52}$$

implying

$$B_K \quad \xrightarrow{\text{p}} \quad 0 \quad \text{as} \quad K \quad \longrightarrow \quad \infty. \tag{53}$$

By definition, $S_K = W_{K,1} + \cdots + W_{K,K} + B_K$, thus (11) follows from (51), (53), and Slutsky's theorem.$\square$

References

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

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. New York: Wiley.

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

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

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: Wiley.

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

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

Cressie, N. A. C. (1993). *Statistis for Spatial Data*. Wiley.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. New York: Springer.

Dedecker, J., P. Doukhan, G. Lang, J. R. Leon, S. Louhichi, and C. Prieur (Eds.) (2007). *Weak Dependence: With Examples and Applications*. Springer.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A 222*, 309–368.

Frank, O. and D. Strauss (1986). Markov graphs. *Journal of the American Statistical Association 81*(395), 832–842.

Georgii, H. (2011). *Gibbs measures and phase transitions* (2 ed.). De Gruyter.

Granger, C. W. J. and M. J. Morris (1976). Time series modelling and interpretation. *Journal of the Royal Statistical Society. Series A (General) 139*(2), pp. 246–257.

Handcock, M. (2003). Assessing degeneracy in statistical models of social networks. Technical report, Center for Statistics and the Social Sciences, University of Washington. http://www.csss.washington.edu/Papers.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris (2008). statnet: Software tools for the representation, visualization, analysis and simulation of social network data. *Journal of Statistical Software 24*(1).

Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A 170*, 301–354. With discussion.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association 100*(469), 286–295.

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

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

Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit of social network models. *Journal of the American Statistical Association 103*(481), 248–258.

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

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton: Princeton University Press.

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

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.

Koschade, S. (2006). A social network analysis of Jemaah Islamiyah: The applications to counter-terrorism and intelligence. *Studies in Conflict and Terrorism 29*, 559–575.

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

Koskinen, J. H., G. L. Robins, and P. E. Pattison (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology 7*(3), 366–384.

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

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

Liang, F. (2010). A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computing and Simulation 80*, 1007–1022.

Lovász, L. (2012). *Large Networks and Graph Limits*. American Mathematical Society.

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

Møller, J., A. N. Pettitt, R. Reeves, and K. K. Berthelsen (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika 93*, 451–458.

Møller, J. and R. P. Waagepetersen (2004). *Statistical inference and simulation for spatial point processes*. Boca Raton (Fl.), London, New York: Chapman & Hall/CRC, 2003.

Murray, I., Z. Ghahramani, and D. J. MacKay (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 359–366. AUAI Press.

Newman, M. E. J., D. J. Watts, and S. H. Strogatz (2002). Random graph models of social networks. In *Proceedings of the National Academy of Sciences USA*, Volume 99, pp. 2566–2572.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*(455), 1077–1087.

Pattison, P. and G. Robins (2002). Neighborhood-based models for social networks. In R. M. Stolzenberg (Ed.), *Sociological Methodology*, Volume 32, Chapter 9, pp. 301–337. Boston: Blackwell Publishing.

Resnick, S. I. (1999). *A Probability Path*. Boston: Birkhäuser.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B 59*, 731–792.

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

Sampson, S. F. (1968). A novitiate in a period of change: An experimental and case study of relationships. Unpublished ph.d. dissertation, Department of Sociology, Cornell University, Ithaca, New York.

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association 106*(496), 1361–1370.

Schweinberger, M. and T. A. B. Snijders (2003). Settings in social networks: A measurement model. In R. M. Stolzenberg (Ed.), *Sociological Methodology*, Volume 33, Chapter 10, pp. 307–341. Boston & Oxford: Basil Blackwell.

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

Singla, P. and P. Domingos (2007). Markov logic in infinite domains. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 368–375.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure 3*, 1–40.

Snijders, T. A. B. (2007). Contribution to the discussion of Handcock, M.S., Raftery, A.E., and J.M. Tantrum, Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A 170*, 322–324.

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

Stein, M. L. (1999). New York: Springer.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review 28*, 513–527.

Strauss, D. and M. Ikeda (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association 85*(409), 204–212.

Tallberg, C. (2005). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology 29*, 1–23.

van Duijn, M. A. J., T. A. B. Snijders, and B. J. H. Zijlstra (2004). P2: a random effects model with covariates for directed graphs. *Statistica Neerlandica 58*, 234–254.

Vu, D. Q., D. R. Hunter, and M. Schweinberger (2013). Model-based clustering of large networks. *Annals of Applied Statistics 7*, 1010–1039.

Wang, J. and Y. F. Atchade (2014). Bayesian inference of exponential random graph models for large social networks. *Communications in Statistics - Simulation and Computation 43*, 359–377.

Wang, Y. J. and G. Y. Wong (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association 82*(397), 8–19.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

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

Welch, D., S. Bansal, and D. R. Hunter (2011). Statistical inference to advance network models in epidemiology. *Epidemics 3*, 38–45.

White, H. C., S. Boorman, and R. L. Breiger (1976). Social structure from multiple networks. *American Journal of Sociology 81*, 730–770.

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