Computational Statistics and Data Analysis 152 (2020) 107029

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Large-scale estimation of random graph models with local dependence

Sergii Babkin [a], Jonathan R. Stewart [b], Xiaochen Long [c], Michael Schweinberger [c,*]

[a] Microsoft, United States of America
[b] Department of Statistics, Florida State University, United States of America
[c] Department of Statistics, Rice University, United States of America

ARTICLE INFO

Article history:
Received 6 May 2019
Received in revised form 11 March 2020
Accepted 4 June 2020
Available online 9 June 2020

Keywords:
Exponential random graph models
Latent structure models
Stochastic block models
Variational methods
EM algorithms
MM algorithms

ABSTRACT

A class of random graph models is considered, combining features of exponential-family models and latent structure models, with the goal of retaining the strengths of both of them while reducing the weaknesses of each of them. An open problem is how to estimate such models from large networks. A novel approach to large-scale estimation is proposed, taking advantage of the local structure of such models for the purpose of local computing. The main idea is that random graphs with local dependence can be decomposed into subgraphs, which enables parallel computing on subgraphs and suggests a two-step estimation approach. The first step estimates the local structure underlying random graphs. The second step estimates parameters given the estimated local structure of random graphs. Both steps can be implemented in parallel, which enables large-scale estimation. The advantages of the two-step estimation approach are demonstrated by simulation studies with up to 10,000 nodes and an application to a large Amazon product recommendation network with more than 10,000 products.
© 2020 Elsevier B.V. All rights reserved.

1. Introduction

The statistical analysis of network data is an emerging area in statistics (Kolaczyk, 2009). Network data arise in the study of insurgent and terrorist networks, contact networks facilitating the spread of infectious diseases (e.g., coronaviruses such as SARS, MERS, and COVID-19; Ebola; HIV), social networks, and the World Wide Web.

Many models of network data have been proposed, as described in recent review papers (Fienberg, 2012; Hunter et al., 2012; Salter-Townshend et al., 2012; Smith et al., 2019; Schweinberger et al., 2020; Hoff, 2020). Among the plethora of models, two broad classes of models can be distinguished: models with latent structure, including stochastic block models (e.g., Nowicki and Snijders, 2001; Bickel and Chen, 2009; Rohe et al., 2011) and latent space models (e.g., Hoff et al., 2002; Handcock et al., 2007; Sewell and Chen, 2015; Smith et al., 2019); and exponential-family models of random graphs (Lusher et al., 2013; Harris, 2013), including Erdős and Rényi random graphs, logistic regression models, and models resembling Markov random fields in spatial statistics (Besag, 1974). Both have advantages and disadvantages, as one of the pioneers of statistical network analysis observed:

* Correspondence to: Department of Statistics, Rice University, 6100 Main St, Houston, TX 77005, USA.
E-mail addresses: sergii.babkin@gmail.com (S. Babkin), jrstewart@fsu.edu (J.R. Stewart), xl81@rice.edu (X. Long), michael.schweinberger@rice.edu (M. Schweinberger).

https://doi.org/10.1016/j.csda.2020.107029
0167-9473/© 2020 Elsevier B.V. All rights reserved.

> *"I expect that, especially for modeling larger networks (with, say, a few hundred or more nodes), the latent space models will not be able to represent network structures as expressed by subgraph counts sufficiently well and the exponential random-graph models will not be able to represent the cohesive structure sufficiently well. Models that combine important features of these two approaches may be the next generation of social network models" (Snijders, 2007, p. 324).*

In other words, latent structure models capture who is close to whom, but are not flexible models of many network phenomena of interest (despite the fact that latent space models induce a weak tendency towards transitivity, see Hoff et al., 2002). Well-posed exponential-family models of random graphs can capture a vast range of network phenomena of interest (including, but not limited to, transitivity), but the underlying assumptions of many of those models make more sense in small networks than large networks. A case in point is the Markov random graphs of Frank and Strauss (1986). These models assume that possible edges $X_{i,j} \in \{0, 1\}$ and $X_{k,l} \in \{0, 1\}$ of pairs of nodes $\{i, j\}$ and $\{k, l\}$ are independent conditional on the rest of the graph provided $\{i, j\}$ and $\{k, l\}$ do not overlap, but are otherwise dependent. As a consequence, each possible edge $X_{i,j}$ depends on $2\,(n-2)$ other possible edges, where n is the number of nodes. If Markov random graphs were applied to online social networks such as Facebook, then each possible friendship would depend on billions of other possible friendships. Such dependence assumptions are implausible, and when coupled with homogeneity assumptions regarding parameters can give rise to the well-studied issue of model near-degeneracy (Handcock, 2003; Schweinberger, 2011; Chatterjee and Diaconis, 2013; Mele, 2017).

One class of next-generation random graph models was introduced in Schweinberger and Handcock (2015), which combines important features of stochastic block models and exponential-family models, with the goal of retaining the strengths of both of them while reducing the weaknesses of each of them. The basic idea is that a set of nodes is partitioned into blocks, and edges among nodes within and between blocks are governed by exponential-family models of random graphs with local dependence within blocks. A simple example is a model where edges between blocks are independent Bernoulli random variables, whereas edges within blocks are generated by an exponential-family model which encourages triangles within blocks but ensures that, for each pair of nodes within a block, the added value of additional edges and triangles decays. Such models induce local dependence within blocks and the overall dependence induced by the model is weak provided the blocks are not too large. We have shown elsewhere that such models are well-behaved—in contrast to the infamous triangle model first studied by Strauss (1986), Jonasson (1999), Häggström and Jonasson (1999), and others (e.g., Chatterjee and Diaconis, 2013)—and that statistical inference is possible and supported by statistical theory: e.g., we have established concentration and consistency results for canonical and curved exponential-family models of random graphs with local dependence under the assumption that the blocks are known and the sizes of the blocks are similar in a well-defined sense (Schweinberger and Stewart, 2020). In addition, when the blocks are unknown, the block memberships of most nodes can be recovered with high probability under weak dependence and smoothness conditions (Schweinberger, 2020).

While some progress has been made in terms of statistical theory, computing remains challenging. When the block structure is known (which is the case in multilevel networks, e.g., networks consisting of units of armed forces, Lazega and Snijders, 2016), statistical inference for parameters can rely on existing methods for estimating parameters of exponential-family models of random graphs (e.g., Strauss and Ikeda, 1990; Snijders, 2002; Hunter and Handcock, 2006; Caimo and Friel, 2011; Hummel et al., 2012; Okabayashi and Geyer, 2012; Atchade et al., 2013; Jin and Liang, 2013; Thiemichen and Kauermann, 2017; Krivitsky, 2017; Byshkin et al., 2018; Tan and Friel, 2020). If the block structure is unknown, however, it needs to be estimated based on the observed network. The recovery of unknown block structure resembles the recovery of unknown block structure in stochastic block models (e.g., Nowicki and Snijders, 2001; Bickel and Chen, 2009; Bickel et al., 2011; Rohe et al., 2011; Choi et al., 2012; Celisse et al., 2012; Priebe et al., 2012; Amini et al., 2013; Rohe et al., 2014; Zhang and Zhou, 2016; Binkiewicz et al., 2017; Gao et al., 2017). However, the recovery of unknown block structure is more challenging for exponential-family models with local dependence than stochastic block models. The main challenge is the intractability of the complete-data likelihood function, i.e., the likelihood function given an observation of the network as well as the block structure. The intractability of the complete-data likelihood function is rooted in the intractability of the normalizing constants of within-block probability mass functions, which stems from the local dependence within blocks.

We present here a tractable approximation of the likelihood function, leveraging concentration results for random graphs with local dependence. Based on the approximation of the likelihood function, we propose a two-step estimation approach that exploits the local structure of random graphs with local dependence for the purpose of local computing. The first step estimates the block structure and decomposes the random graph into subgraphs with local dependence. The decomposition of the random graph relies on approximations of the likelihood function supported by theoretical results. The second step estimates parameters given the estimated block structure by using Monte Carlo maximum likelihood methods (Hunter and Handcock, 2006) or maximum pseudolikelihood methods (Strauss and Ikeda, 1990). Both steps can be implemented in parallel, which enables large-scale estimation on multi-core computers or computing clusters. We demonstrate the advantages of the two-step estimation approach by simulation studies with up to 10,000 nodes and an application to a large Amazon product recommendation network with more than 10,000 products.

The remainder of our paper is organized as follows. Section 2 introduces models. Section 3 discusses likelihood-based inference based on approximations of the likelihood function motivated by theoretical results, and Section 4 takes advantage of such approximations to estimate models. Section 5 presents simulation results and Section 6 presents an application.

2. Models

We consider random graphs with a set of nodes $\mathcal{A} = \{1, \ldots, n\}$ and a set of edges $\mathcal{E} \subset \mathcal{A} \times \mathcal{A}$. Here, edges are regarded as random variables $X_{i,j} \in \{0, 1\}$, where $X_{i,j} = 1$ if nodes i and j are connected by an edge and $X_{i,j} = 0$ otherwise. We focus on undirected random graphs without self-edges, although the methods introduced here can be extended to directed random graphs. Throughout, we denote by $\boldsymbol{X} = (X_{i,j})_{i<j}^{n}$ the set of possible edges of a random graph, and by $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$ the sample space of $\boldsymbol{X}$.

Since the pioneering research of Holland and Leinhardt (1970, 1972, 1976) in the 1970s, it is known that network data are dependent data. In small networks, each possible edge might depend on many other possible edges – in fact, it might depend on all other possible edges. In large networks, however, it is not credible that each possible edge depends on all other possible edges: e.g., in networks with $n \gg 10{,}000$ nodes – such as the network used in Section 6 – it is implausible that each possible edge depends on all $\binom{n}{2} - 1 \gg 50{,}000{,}000$ other possible edges. Instead, it is tempting to believe that the dependence among edges in large networks is local, in the sense that each edge depends on a subset of other possible edges. We consider here a simple form of local dependence, following Schweinberger and Handcock (2015). Assume that $\mathcal{A}$ is partitioned into $K \geq 2$ subsets of nodes $\mathcal{A}_1, \ldots, \mathcal{A}_K$, called blocks, and let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be the block memberships of nodes, where $z_{i,k} = 1$ if node i belongs to block $\mathcal{A}_k$ and $z_{i,k} = 0$ otherwise. We henceforth denote by $\boldsymbol{X}_{k,k} = (X_{i,j})_{i<j:\, z_{i,k}=z_{j,k}=1}^{n}$ the within-block edges among nodes in block $\mathcal{A}_k$ $(k = 1, \ldots, K)$ and by $\boldsymbol{X}_{k,l} = (X_{i,j})_{i,j:\, z_{i,k}=z_{j,l}=1}^{n}$ the between-block edges among nodes in block $\mathcal{A}_k$ and nodes in block $\mathcal{A}_l$ $(l < k = 1, \ldots, K)$.

Definition (*Local Dependence*)**.** A model of a random graph $\boldsymbol{X}$ satisfies local dependence if the probability mass function of $\boldsymbol{X}$ can be factorized as follows:

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) = \prod_{k=1}^{K} p_{\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k}) \prod_{l=1}^{k-1} \prod_{i,j:\, z_{i,k}=z_{j,l}=1}^{n} p_{\eta_{B,k,l}(\boldsymbol{\theta},\boldsymbol{z})}(x_{i,j}), \quad \boldsymbol{x} \in \mathbb{X}, \tag{1}$$

where

- $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is the probability mass of graph $\boldsymbol{x}$, parameterized by $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^d$;
- $p_{\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$ is the probability mass of within-block subgraph $\boldsymbol{x}_{k,k}$, parameterized by $\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^{d_{W,k}}$ $(k = 1, \ldots, K)$;
- $p_{\eta_{B,k,l}(\boldsymbol{\theta},\boldsymbol{z})}(x_{i,j})$ is the probability mass of between-block edge $x_{i,j}$, parameterized by $\boldsymbol{\eta}_{B,k,l}(\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^{d_{B,k,l}}$ $(i, j : z_{i,k} = z_{j,l} = 1)$;
- the parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^d$ consists of the subvectors $\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^{d_{W,k}}$ $(k = 1, \ldots, K)$ and $\boldsymbol{\eta}_{B,k,l}(\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^{d_{B,k,l}}$ $(l < k = 1, \ldots, K)$.

The map $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \mathbb{R}^d$ depends on the model. In general, $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \mathbb{R}^d$ may be a linear or non-linear function of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^q$ of dimension $q \leq d$. An example is given by the curved exponential-family parameterizations used in Section 6, where $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \mathbb{R}^d$ is a non-linear function of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^q$ of dimension $q < d$. Well-chosen curved exponential-family parameterizations help ensure that the added value of additional edges, triangles, and other subgraph configurations decays (Stewart et al., 2019), which is plausible in many applications and can improve the in-sample performance of models (Hunter et al., 2008) as well as the out-of-sample performance of models (Stewart et al., 2019). The factorization of the probability mass function of models with local dependence has at least three implications. First, edges among nodes in block $\mathcal{A}_k$ can depend on other edges in block $\mathcal{A}_k$ $(k = 1, \ldots, K)$. Second, the factorization implies that edges among nodes in block $\mathcal{A}_k$ do not depend on edges that involve nodes in other blocks $(k = 1, \ldots, K)$. Third, edges between blocks are independent. In other words, the dependence is local in the sense that it is confined to within-block subgraphs.

We introduce here two examples of models with local dependence, to be used as motivating examples throughout Sections 3–5.

Example 1 (*Stochastic Block Model*)**.** An important special case of models with local dependence is given by stochastic block models (e.g., Nowicki and Snijders, 2001; Bickel and Chen, 2009; Rohe et al., 2011). Stochastic block models assume that

possible edges $X_{i,j}$ between nodes i and j in blocks k and l are independent Bernoulli($\mu_{k,l}$) random variables, where the probability $\mu_{k,l} \in (0, 1)$ of an edge depends on the blocks k and l of nodes i and j, respectively. Stochastic block models are special cases of models with local dependence, having within-block probability mass functions of the form

$$\begin{aligned} p_{\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k}) &= \prod_{i<j:\, z_{i,k}=z_{j,k}=1}^{n} \mu_{k,k}^{x_{i,j}} (1-\mu_{k,k})^{1-x_{i,j}} \\ &= \prod_{i<j:\, z_{i,k}=z_{j,k}=1}^{n} \exp\left(\theta_{W,k,k}\, x_{i,j} - \log(1+\exp(\theta_{W,k,k}))\right) \\ &\propto \exp\left(\theta_{W,k,k} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}\right), \end{aligned}$$

where

$$\theta_{W,k,k} \;=\; \text{logit}(\mu_{k,k}) \;\in\; \mathbb{R}, \quad k = 1, \ldots, K.$$

The between-block probability mass functions are of the form

$$\begin{aligned} p_{\boldsymbol{\eta}_{B,k,l}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,l}) &= \prod_{i<j:\, z_{i,k}=z_{j,l}=1}^{n} \mu_{k,l}^{x_{i,j}} (1-\mu_{k,l})^{1-x_{i,j}} \\ &= \prod_{i<j:\, z_{i,k}=z_{j,l}=1}^{n} \exp\left(\theta_{B,k,l}\, x_{i,j} - \log(1+\exp(\theta_{B,k,l}))\right) \\ &\propto \exp\left(\theta_{B,k,l} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,l}\right), \end{aligned}$$

where

$$\theta_{B,k,l} \;=\; \text{logit}(\mu_{k,l}) \;\in\; \mathbb{R}, \quad l < k = 1, \ldots, K.$$

As a consequence, the joint probability mass function of random graph $\boldsymbol{X}$ is proportional to

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \;\propto\; \prod_{k=1}^{K} \exp\left(\theta_{W,k,k} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}\right) \prod_{l=1}^{k-1} \exp\left(\theta_{B,k,l} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,l}\right).$$

However, while stochastic block models are popular, the assumption that edges within and between blocks are independent is restrictive, because edges among nodes that are close – in the sense of being members of the same block – may very well be dependent.

Example 2 (*Model with Local Dependence*)**.** To allow edges within blocks to be dependent, we build on the stochastic block model described above, but tilt the within-block probability mass function of the stochastic block model as follows:

$$p_{\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k}) \;\propto\; \exp\left(\theta_{W,k,k,1} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}\right) \exp\left(\theta_{W,k,k,2} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}\, \mathbb{1}_{i,j}(\boldsymbol{x}_{k,k})\right),$$

where $\theta_{W,k,k,1} \in \mathbb{R}$ and $\theta_{W,k,k,2} \in \mathbb{R}$ and $\mathbb{1}_{i,j}(\boldsymbol{x}_{k,k})$ is an indicator function, which is 1 if nodes i and j are both connected to one or more other nodes in block k, and is 0 otherwise. If $x_{i,j} = 1$ and $\mathbb{1}_{i,j}(\boldsymbol{x}_{k,k}) = 1$, the edge between nodes i and j is called transitive, because i and j form transitive triples with other nodes, which are known as triangles in the random graph literature. The first term of the within-block probability mass function shown above corresponds to the within-block probability mass function of the stochastic block model. The second term tilts the within-block probability mass function of the stochastic block model: If $\theta_{W,k,k,2} > 0$, the model rewards within-block subgraphs with transitive edges, whereas $\theta_{W,k,k,2} < 0$ penalizes them, and $\theta_{W,k,k,2} = 0$ neither rewards nor penalizes them—in which case the model reduces to the stochastic block model.

Using the same between-block probability mass functions as the stochastic block model, the joint probability mass function of random graph $\boldsymbol{X}$ is proportional to

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \propto \prod_{k=1}^{K} \exp\left(\theta_{W,k,k,1} \sum_{i<j} x_{i,j}\, z_{i,k}\, z_{j,k}\right) \exp\left(\theta_{W,k,k,2} \sum_{i<j} x_{i,j}\, z_{i,k}\, z_{j,k}\, \mathbb{1}_{i,j}(\boldsymbol{x}_{k,k})\right) \times \prod_{l=1}^{k-1} \exp\left(\theta_{B,k,l} \sum_{i<j} x_{i,j}\, z_{i,k}\, z_{j,l}\right).$$

The resulting model can capture an excess in the expected number of transitive edges within blocks, relative to the stochastic block model. To demonstrate, note that the resulting model is an exponential-family model and – by exponential-family theory (Brown, 1986, Corollary 2.5, p. 37) – the expected number of transitive edges in block $\mathcal{A}_k$ satisfies

$$\underbrace{\mathbb{E}_{\theta_{W,k,k,1},\,\theta_{W,k,k,2}>0}\, s_{k,k,2}(\boldsymbol{X})}_{\text{model with local dependence}} > \underbrace{\mathbb{E}_{\theta_{W,k,k,1},\,\theta_{W,k,k,2}=0}\, s_{k,k,2}(\boldsymbol{X})}_{\text{stochastic block model}}, \quad k = 1, \ldots, K,$$

where $s_{k,k,2}(\boldsymbol{X})$ is the number of transitive edges in block $\mathcal{A}_k$ and $\mathbb{E}_{\theta_{W,k,k,1},\,\theta_{W,k,k,2}}\, s_{k,k,2}(\boldsymbol{X})$ is the expectation of $s_{k,k,2}(\boldsymbol{X})$. In other words, the expected number of transitive edges in block $\mathcal{A}_k$ is greater under the model with $\theta_{W,k,k,2} > 0$ than under the stochastic block model with $\theta_{W,k,k,2} = 0$, assuming that both have the same edge parameters $\theta_{W,k,k,1}$ $(k = 1, \ldots, K)$. Therefore, the model with local dependence can capture an excess in the expected number of transitive edges within blocks, relative to the stochastic block model. In addition, models with local dependence can capture excesses in the expected number of other subgraph statistics within blocks by adding suitable model terms.

How models with within-block edge and transitive edge terms differ from the "triangle model". It is worth noting that the model with local dependence induced by within-block edge and transitive edge terms differs in two important ways from the infamous triangle model with edge and triangle terms, which has been known to be ill-behaved since the 1980s (Strauss, 1986; Jonasson, 1999; Schweinberger, 2011; Chatterjee and Diaconis, 2013):

- The model with local dependence restricts dependence among edges to subsets of nodes, i.e., blocks. As long as the blocks are not too large, the overall dependence induced by the model is weak. By contrast, the triangle model does not restrict dependence to subsets of nodes, leading to undesirable behavior in large graphs.
- Within blocks, the model with local dependence and positive within-block transitive edge parameters assumes that, for each pair of nodes, the value added by the first triangle to the log odds of the conditional probability of an edge is positive, but additional triangles add nothing. By contrast, the triangle model assumes that, for each pair of nodes, each additional triangle has the same added value, which is unreasonable and results in undesirable behavior in large graphs.

These sensible assumptions ensure that models with local dependence have more desirable properties than the triangle model (Schweinberger and Stewart, 2020).

Exponential-family representations of models. It is worth noting that Models 1 and 2 can be represented in exponential-family form:

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}),\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}))), \quad \boldsymbol{x} \in \mathbb{X},$$

where $\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}),\, s(\boldsymbol{x})\rangle$ denotes the inner product of a vector of natural parameters $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \mathbb{R}^d$ and a vector of sufficient statistics $s : \mathbb{X} \mapsto \mathbb{R}^d$, and $\psi(\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}))$ ensures that $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ sums to 1. While exponential-family representations are not needed to introduce Models 1 and 2, the properties of exponential families facilitate the theoretical results in Section 3.2, concerned with approximations of likelihood functions.

3. Likelihood-based inference

While the block structure $\boldsymbol{z}$ is observed in some applications (e.g., in multilevel networks), it is unobserved in other applications. We focus here on unobserved block structure $\boldsymbol{z}$.

It is tempting to base statistical inference for the unknown block structure $\boldsymbol{z}$ and the unknown parameter vector $\boldsymbol{\theta}$ on the likelihood function, but likelihood-based inference for models with local dependence is challenging. The main reason is that the probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable, because the within-block probability mass functions $p_{\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$ are intractable $(k = 1, \ldots, K)$. The intractability of $p_{\boldsymbol{\eta}_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$ is rooted in the fact that its normalizing constant is a sum over all $\exp(\binom{|\mathcal{A}_k|}{2} \log 2)$ possible within-block subgraphs of block $\mathcal{A}_k$, which cannot be computed unless $\mathcal{A}_k$ is small, i.e., unless $|\mathcal{A}_k| \ll 10$ $(k = 1, \ldots, K)$.

To facilitate likelihood-based inference, we introduce tractable approximations of the intractable probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ in Section 3.1 and support them by theoretical results in Section 3.2. A statistical algorithm that takes advantage of such approximations is introduced in Section 4.

3.1. Approximate likelihood functions: motivation

Suppose that we want to estimate both $\boldsymbol{z}$ and $\boldsymbol{\theta}$. It is natural to estimate them by using an iterative algorithm that cycles through updates of $\boldsymbol{z}$ and $\boldsymbol{\theta}$ as follows:

Step 1: Update $\boldsymbol{z}$ given $\boldsymbol{\theta}$.
Step 2: Update $\boldsymbol{\theta}$ given $\boldsymbol{z}$.

The algorithm sketched above is generic and cannot be used in practice, but regardless of which specific algorithm is used – whether EM, Monte Carlo EM, variational EM, Bayesian Markov chain Monte Carlo, or other algorithms – most of them have in common that Step 1 is either infeasible or time-consuming, whereas Step 2 is less problematic than Step 1.

Step 1. Step 1 is either infeasible or time-consuming, because the probability mass function $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable. To demonstrate, consider a Bayesian Markov chain Monte Carlo algorithm that updates $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ given $\boldsymbol{\theta}$ by Gibbs sampling. Gibbs sampling of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ turns out to be infeasible, because the full conditional distributions of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ depend on the intractable within-block probability mass functions $p_{\eta_{W,k}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$ $(k = 1, \ldots, K)$. One could approximate them by Monte Carlo samples of within-block subgraphs, but such approximations may not generate Markov chain Monte Carlo samples from the target distribution (Liang et al., 2016) and are problematic on computational grounds:

- Using Monte Carlo approximations of within-block probability mass functions is infeasible when the number of nodes n is large, because such approximations are needed for each update of each of the n block memberships $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.
- Worse, the n block memberships $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ cannot be updated in parallel, because the block membership of one node depends on the block memberships of other nodes.

As a consequence, Step 1 is infeasible when n is large.

Step 2. Step 2 is less problematic than Step 1. While the probability mass function $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable and may have to be approximated by Monte Carlo methods (Hunter and Handcock, 2006), such Monte Carlo approximations are needed once to update $\boldsymbol{\theta}$ given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, whereas Monte Carlo approximations are needed n times to update $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ given $\boldsymbol{\theta}$ one by one. In addition, the probability mass function $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ decomposes into between- and within-block probability mass functions and hence within-block probability mass functions can be approximated in parallel, i.e., on multi-core computers or computing clusters.

Approximations. To enable feasible updates of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ when n is large, we are interested in approximating the intractable probability mass function $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ by a tractable probability mass function. To do so, we focus on models with between-block edge terms $\theta_{B,k,l} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,l}$ and within-block edge terms $\theta_{W,k,k,1} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}$ along with additional within-block terms that induce local dependence within blocks. We denote the vector of between- and within-block edge parameters by $\boldsymbol{\theta}_1$ and the vector of all other within-block parameters by $\boldsymbol{\theta}_2$. We assume that $\boldsymbol{\theta}_2 = \boldsymbol{0}$ reduces the model to the stochastic block model. An example is given by the model with between-block edge terms and within-block edge and transitive edge terms described in Section 2: the parameter vector $\boldsymbol{\theta}_1$ consists of the between-block edge parameters $\theta_{B,k,l}$ $(l < k = 1, \ldots, K)$ and the within-block edge parameters $\theta_{W,k,k,1}$ $(k = 1, \ldots, K)$, whereas the parameter vector $\boldsymbol{\theta}_2$ consists of the within-block transitive edge parameters $\theta_{W,k,k,2}$ $(k = 1, \ldots, K)$. If the parameter vector $\boldsymbol{\theta}_2$ is set to $\boldsymbol{0}$, the model reduces to the stochastic block model, under which edges within and between blocks are independent.

Such models have two useful properties:

- The probability mass functions $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ impose the same probability law on between-block subgraphs.
- The probability mass function $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ is tractable, because edges between and within blocks are independent under all $\boldsymbol{z}$.

We henceforth approximate $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ by $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$, which corresponds to the probability mass function of the stochastic block model.

The idea underlying the approximation is that when the blocks are not too large, most pairs of nodes are not members of the same block. Since $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ impose the same probability law on possible edges between pairs of nodes that are not members of the same block, $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ agree on most of the random graph. Therefore, $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be approximated by $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ for the purpose of updating $\boldsymbol{z}$ given $\boldsymbol{\theta}$. Suppose, e.g., that we consider to update $\boldsymbol{z}$ given $\boldsymbol{\theta}$ by replacing $\boldsymbol{z}$ by $\boldsymbol{z}' \neq \boldsymbol{z}$. We may decide to do so if the loglikelihood ratio

$$\log \frac{p_{\eta(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x})}{p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})} = \log p_{\eta(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$$

is large: e.g., the acceptance probability of Metropolis–Hastings algorithms depends on the loglikelihood ratio above. If $p_{\eta(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be approximated by $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$, we can base the decision on $\log p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ rather

than $\log p_{\eta(\theta,\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x})$, because

$$\begin{aligned}\log p_{\eta(\theta,\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x}) \quad &= \quad [\log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})] \\ + [\log p_{\eta(\theta,\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z}')}(\boldsymbol{x})] \;-\; &[\log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})].\end{aligned}$$

Therefore, as long as

$$\max_{\boldsymbol{z}} \; |\log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})|$$

is small, we have

$$\log p_{\eta(\theta,\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x}) \;\approx\; \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x}).$$

The advantage of approximating $p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x})$ by $p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ is that there exist methods for stochastic block models to estimate the block structure from large random graphs (e.g., Daudin et al., 2008; Rohe et al., 2011; Amini et al., 2013; Vu et al., 2013). We take advantage of such methods in Section 4, but we first shed light on the conditions under which $\max_{\boldsymbol{z}} \; |\log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})|$ is small.

3.2. Approximate likelihood functions: theoretical results

We show that updates of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ can be based on $p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ rather than $p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x})$ by showing that

$$\max_{\boldsymbol{z}} \; |\log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})|$$

is small with high probability when the blocks are not too large.

We start with a special case in Theorem 1 and then present more general results in Theorem 2. To prepare the ground for Theorems 1 and 2, let $\mathbb{Z} = \{1, \ldots, K\}^n$ be the set of all block structures, and denote by $m(\boldsymbol{z})$ the size of the largest block under $\boldsymbol{z} \in \mathbb{Z}$. Let $\mathbb{S} \subseteq \mathbb{Z}$ be any subset of block structures that includes the data-generating block structure $\boldsymbol{z}^\star \in \mathbb{Z}$, and denote by $m(\mathbb{S})$ the size of the largest block among all block structures $\boldsymbol{z} \in \mathbb{S}$, so that $m(\boldsymbol{z}) \leq m(\mathbb{S})$ for all $\boldsymbol{z} \in \mathbb{S}$. We allow the size of the largest block $m(\mathbb{S})$ to increase as a function of the number of nodes n: e.g., $m(\mathbb{S})$ may be a constant multiple of $\log n$ or n^α ($\alpha < 1$). The size of the largest data-generating block is denoted by $\|\mathcal{A}\|_\infty = \max_{1 \leq k \leq K} |\mathcal{A}_k|$.

Theorem 1. *Consider the model with between-block edge terms and within-block edge and transitive edge terms described in Section 2. Let $\boldsymbol{\Theta} = \prod_{l<k}^{K} \boldsymbol{\Theta}_{k,l}$ be the parameter space, $\boldsymbol{\Theta}_{k,l}$ be a compact subset of $\mathbb{R}$ ($l < k = 1, \ldots, K$), and $\boldsymbol{\Theta}_{k,k}$ be a compact subset of $\mathbb{R}^2$ ($k = 1, \ldots, K$). Choose $\epsilon \in (0, 1)$ as small as desired. Then there exists a universal constant $c > 0$ such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, with at least probability $1 - \epsilon$,*

$$\max_{\boldsymbol{z} \in \mathbb{S}} |\log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})| \;<\; 2\,c \sqrt{-\log \frac{\epsilon}{2} + n \log K}\, \sqrt{K}\, m(\mathbb{S})^2\, \|\mathcal{A}\|_\infty^2\, \log n.$$

The proof of Theorem 1 can be found in the supplement. The basic idea underlying Theorem 1 is that the deviation $\max_{\boldsymbol{z}} \; |\log p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})|$ cannot be too large when the blocks are not too large, because most pairs of nodes are not members of the same block and $p_{\eta(\theta,\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\eta(\theta_1,\theta_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ impose the same probability law on possible edges between pairs of nodes that are not members of the same block.

Theorem 1 implies that the largest deviation of the loglikelihood function under the unrestricted model and the restricted model is smaller than n^2 with high probability, provided the blocks are not too large. To see that, observe that $\|\mathcal{A}\|_\infty \leq m(\mathbb{S})$ and $K \leq n$ imply

$$2\,c \sqrt{-\log \frac{\epsilon}{2} + n \log K}\, \sqrt{K}\, m(\mathbb{S})^2\, \|\mathcal{A}\|_\infty^2\, \log n \leq \delta(\epsilon)\; n\; m(\mathbb{S})^4\; (\log n)^{3/2},$$

where $\delta(\epsilon) > 0$ is a function of ϵ but is not a function of the number of nodes n. In other words, as long as the size $m(\mathbb{S})$ of the largest block in $\mathbb{S}$ satisfies $m(\mathbb{S}) \ll n^{1/4} \,/\, (\log n)^{3/8}$, the maximum deviation is much smaller than n^2 with high probability. As a consequence, the largest deviation of the loglikelihood function under the unrestricted model and the restricted model is small with high probability – note that in dense random graphs many quantities are of order n^2, so quantities of order less than n^2 may be considered small. For example, consider Bernoulli(μ) random graphs, under which possible edges $X_{i,j}$ are independent Bernoulli(μ) random variables (Erdős and Rényi, 1959, 1960). If a Bernoulli(μ) random graph is dense in the sense that $\mu = \mathbb{E} X_{i,j} \in (0, 1)$ does not decrease as a function of the number of nodes n, then the expected number of edges is of order n^2,

$$\mathbb{E} \sum_{i<j}^{n} X_{i,j} \;=\; \binom{n}{2} \mu,$$

and so is the expected loglikelihood function of the natural parameter $\theta = \text{logit}(\mu) \in \mathbb{R}$,

$$\mathbb{E} \log p_\theta(\boldsymbol{X}) \;=\; \binom{n}{2} \left(\theta\; \mu - \log(1 + \exp(\theta))\right).$$

Other quantities are likewise of order n^2 in dense random graphs, so quantities of order less than n^2 may be considered small. Thus, the largest deviation of the loglikelihood function under the unrestricted model and the restricted model is small with high probability.

Last, but not least, note that $m(\mathbb{S}) \ll n^{1/4}/(\log n)^{3/8}$ implies that the number of blocks K must satisfy $K \gg n^{3/4}(\log n)^{3/8}$. In other words, the number of blocks K needs to grow as function of the number of nodes n. It is worth noting that in the special case of stochastic block models, it is known that K is allowed to grow as fast as $n/\log n$ in dense-graph settings (Zhang and Zhou, 2016).

It turns out that the result in Theorem 1 is not limited to the model with between-block edge terms and within-block edge and transitive edge terms, but is a special case of more general results. To introduce these more general results, we make the following assumptions. We assume that the probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be represented in exponential-family form,

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \;=\; \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}),\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}))), \quad \boldsymbol{x} \in \mathbb{X},$$

where $\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}),\, s(\boldsymbol{x})\rangle$ denotes the inner product of a vector of natural parameters $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \mathbb{R}^d$ and a vector of sufficient statistics $s : \mathbb{X} \mapsto \mathbb{R}^d$, and $\psi(\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}))$ ensures that $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ sums to 1; note that all models used in our paper can be represented in exponential-family form. We assume that $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \boldsymbol{\Xi}$ and that $\boldsymbol{\Xi} \subseteq \text{int}(\mathbb{N})$ is a subset of the interior $\text{int}(\mathbb{N})$ of the natural parameter space $\mathbb{N}$ of the exponential family (Brown, 1986). Let $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\eta}^\star}$ be the expectation under the data-generating parameter vector $\boldsymbol{\eta}^\star \equiv \boldsymbol{\eta}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star)$, where $(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) \in \boldsymbol{\Theta} \times \mathbb{Z}$ denotes the data-generating value of $(\boldsymbol{\theta},\boldsymbol{z}) \in \boldsymbol{\Theta} \times \mathbb{Z}$. We denote by $d : \mathbb{X} \times \mathbb{X} \mapsto \{0, 1, 2, \dots\}$ the Hamming metric, which is defined by

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) \;=\; \sum_{i<j}^{n} \mathbb{1}_{x_{1,i,j} \neq x_{2,i,j}}, (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X},$$

where $\mathbb{1}_{x_{1,i,j} \neq x_{2,i,j}}$ is 1 if $x_{1,i,j} \neq x_{2,i,j}$ and is 0 otherwise. The main assumptions can then be stated as follows.

[C.1] There exists a constant $c > 0$ such that, for all $(\boldsymbol{\theta},\boldsymbol{z}) \in \boldsymbol{\Theta} \times \mathbb{Z}$ and all $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X}$,

$$|\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}),\, s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2)\rangle| \;\leq\; c\; d(\boldsymbol{x}_1, \boldsymbol{x}_2)\; m(\boldsymbol{z})\; \log n.$$

[C.2] There exists a constant $c > 0$ such that, for all $(\boldsymbol{\theta}_{k,l,1},\, \boldsymbol{\theta}_{k,l,2}) \in \boldsymbol{\Theta}_{k,l} \times \boldsymbol{\Theta}_{k,l}$ and all $\boldsymbol{z} \in \mathbb{Z}$,

$$\left|\left\langle \boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}_{k,l,1}, \boldsymbol{z}) - \boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}_{k,l,2}, \boldsymbol{z}),\, \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}\, s_{k,l}(\boldsymbol{X})\right\rangle\right| \;\leq\; c\; \|\boldsymbol{\theta}_{k,l,1} - \boldsymbol{\theta}_{k,l,2}\|\; m(\boldsymbol{z})^2\; \log n,$$

where $\boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}_{k,l}, \boldsymbol{z})$, $\boldsymbol{\theta}_{k,l}$, and $s_{k,l}(\boldsymbol{x})$ denote the subvectors of $\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})$, $\boldsymbol{\theta}$, and $s(\boldsymbol{x})$ corresponding to the subgraph between blocks k and l ($l < k$) or the subgraph of block k ($k = l$) and $\boldsymbol{\Theta}_{k,l}$ is a compact subset of $\mathbb{R}^{q_{k,l}}$ ($k \leq l = 1, \dots, K$).

Conditions [C.1] and [C.2] are smoothness conditions: [C.1] states that the inner product of natural parameters and sufficient statistics must not be too sensitive to changes of edges, whereas [C.2] states that the inner product of between- and within-block natural parameters and expected sufficient statistics must not be too sensitive to changes of parameters.

An example of a model violating condition [C.1] is a model containing a sufficient statistic $s(\boldsymbol{x})$ of the form

$$s(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \sum_{i<j}^{n} x_{i,j} = 0 \\ \binom{n}{2} & \text{if } \sum_{i<j}^{n} x_{i,j} \in \{1, \dots, \binom{n}{2}\}. \end{cases}$$

Under such models, adding a single edge can change the inner product of natural parameters and sufficient statistics by a multiple of n^2. That would violate [C.1], because [C.1] assumes that changing a single edge changes the inner product by at most $c\; m(\boldsymbol{z})\; \log n \leq c\; n \log n$.

An example of a model violating condition [C.2] is a model with within-block sufficient statistics that count the number of triangles within blocks,

$$s_{k,k}(\boldsymbol{x}) \;=\; \sum_{h<i<j:\; h,\, i,\, j \,\in\, \mathcal{A}_k}^{n} x_{i,h}\, x_{j,h}\, x_{i,j}, \quad k = 1, \dots, K,$$

with within-block parameters of the form $\eta_{k,k}(\boldsymbol{\theta},\boldsymbol{z}) = \theta_{k,k} \in \mathbb{R}$ ($k = 1, \dots, K$). If all nodes belong to block $\mathcal{A}_1$, the inner product of the block's within-block natural parameter vector and expected sufficient statistic vector is a multiple of $|\theta_{1,1} - \theta'_{1,1}|\, n^3$, where $\theta_{1,1}$ and $\theta'_{1,1}$ are two possible values of the within-block triangle parameter of block $\mathcal{A}_1$. As a result, [C.2] would be violated, because [C.2] requires the inner product to be at most $c\; |\theta_{1,1} - \theta'_{1,1}|\; m(\boldsymbol{z})^2\; \log n \;\leq\; c\; |\theta_{1,1} - \theta'_{1,1}|\; n^2\; \log n$.

However, the fact that some model specifications violate conditions [C.1] and [C.2] is not necessarily a major concern, for two reasons. First, the specifications that violate conditions [C.1] and [C.2] are, more often than not, of limited interest in applications: e.g., we are not aware of any good reason for using the sufficient statistic in the first example, and the triangle term in the second example is known to induce model near-degeneracy in large networks and may therefore not be useful in practice (Strauss, 1986; Jonasson, 1999; Schweinberger, 2011; Chatterjee and Diaconis, 2013). Second, while some ill-posed specifications do not satisfy conditions [C.1] and [C.2], there are many well-posed specifications that do

satisfy them: e.g., conditions [C.1] and [C.2] are satisfied by the model with between-block edge and within-block edge and transitive edge terms, which we verify in the proof of Theorem 1. In addition, conditions [C.1] and [C.2] cover the models with size-dependent parameterizations used in Sections 5 and 6.

The following result, Theorem 2, is a generalization of Theorem 1.

Theorem 2. *Consider a model with local dependence satisfying conditions [C.1] and [C.2]. Choose* $\epsilon \in (0, 1)$ *as small as desired. Then there exists a universal constant* $c > 0$ *such that, for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, with at least probability* $1 - \epsilon$*,*

$$\max_{\boldsymbol{z} \in \mathbb{S}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})| \;<\; 2\,c \sqrt{-\log \frac{\epsilon}{2} + n \log K}\, \sqrt{K}\, m(\mathbb{S})^2\, \|\mathcal{A}\|_\infty^2\, \log n.$$

The proof of Theorem 2 can be found in the supplement. An application of Theorem 2 to the model with between-block edge terms and within-block edge and transitive edge terms can be found in Theorem 1.

Last, but not least, it is worth noting that the upper bound $m(\mathbb{S})$ on the sizes of blocks cannot be too small, because it is not possible to recover the block structure with high probability when the blocks are too small. However, Zhang and Zhou (2016) showed that under stochastic block models the number of blocks K is allowed to grow as fast as $n / \log n$ in dense-graph settings, which implies that the sizes of the blocks can be as small as $\log n$. These important issues are studied in more depth in Zhang and Zhou (2016) and Gao et al. (2017).

4. Two-step estimation approach

We propose a two-step estimation approach that takes advantage of the theoretical results of Section 3 and enables large-scale estimation of models with local dependence.

To describe the two-step estimation approach, assume that $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ is the observed value of a random variable $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ with distribution

$$\boldsymbol{Z}_i \overset{\text{iid}}{\sim} \text{Multinomial}(1,\ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)), \quad i = 1, \ldots, n.$$

It is natural to base statistical inference on the observed-data likelihood function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi}) \;=\; \sum_{\boldsymbol{z} \in \mathbb{Z}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}).$$

The problem is that $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi})$ is intractable, because $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable and the set $\mathbb{Z}$ contains $\exp(n \log K)$ elements.

The first problem can be solved by taking advantage of the theoretical results of Section 3, which suggest that $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be approximated by $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ provided that the blocks are not too large. A complication is that $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ may not be close when the block structure $\boldsymbol{z} \in \mathbb{Z} \setminus \mathbb{S}$ is not contained in $\mathbb{S}$, in which case some of the blocks can be large and the within-block models of large blocks can induce strong dependence. In the worst case, all nodes belong to a single block, in which case $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ may induce strong dependence throughout the random graph while $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ induces no dependence at all, so $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ may be very different. However, the basic inequality

$$\sum_{\boldsymbol{z} \in \mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \le \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi}) \le \sum_{\boldsymbol{z} \in \mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) + \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S})$$

suggests that as long as the event $\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S}$ is a rare event in the sense that $\mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S})$ is close to 0, $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi})$ can be approximated by $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2{=}\boldsymbol{0}, \boldsymbol{\pi})$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi}) \;&=\; \sum_{\boldsymbol{z} \in \mathbb{Z}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \qquad \approx \sum_{\boldsymbol{z} \in \mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \\ &\approx \sum_{\boldsymbol{z} \in \mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \;\approx \sum_{\boldsymbol{z} \in \mathbb{Z}} p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \;=\; \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2{=}\boldsymbol{0}, \boldsymbol{\pi}). \end{aligned}$$

The assumption that $\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S}$ is a rare event – i.e., the probabilities $\pi_1, \ldots, \pi_K$ are small – makes sense in a wide range of applications, because communities in real-world networks tend to be small (see, e.g., the discussion of Rohe et al., 2011). Therefore, as long as $\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S}$ is a rare event, we can base statistical inference concerning the block structure on $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2{=}\boldsymbol{0}, \boldsymbol{\pi})$ rather than $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi})$. To simplify the notation, we write henceforth $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ instead of $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2{=}\boldsymbol{0}, \boldsymbol{\pi})$.

The second problem can be solved by methods developed for stochastic block models, because $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ is the observed-data likelihood function of a stochastic block model. There are many stochastic block model methods that could be used, such as profile likelihood (Bickel and Chen, 2009), pseudo-likelihood (Amini et al., 2013), spectral clustering (Rohe et al., 2011), and variational methods (Daudin et al., 2008; Vu et al., 2013). Among these methods, we found that the variational methods of Vu et al. (2013) work best in practice: the simulation results in Section 5 demonstrate this. In addition, the variational methods of Vu et al. (2013) have the advantage of being able to estimate stochastic block models from networks

with hundreds of thousands of nodes due to a running time of $O(n)$ for sparse random graphs and $O(n^2)$ for dense random graphs (Vu et al., 2013). Some consistency and asymptotic normality results for variational methods for stochastic block models were established by Celisse et al. (2012) and Bickel et al. (2013).

Variational methods approximate $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi}) = \log \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ by introducing an auxiliary distribution $a(\boldsymbol{z})$ with support $\mathbb{Z}$ and bounding $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ from below by using Jensen's inequality:

$$\begin{aligned}\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi}) &= \log \sum_{\boldsymbol{z}\in\mathbb{Z}} a(\boldsymbol{z}) \frac{p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z})}{a(\boldsymbol{z})} \\ &\geq \sum_{\boldsymbol{z}\in\mathbb{Z}} a(\boldsymbol{z}) \log \frac{p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z})}{a(\boldsymbol{z})} \stackrel{\text{def}}{=} \hat{\ell}(\boldsymbol{\theta}_1, \boldsymbol{\pi}).\end{aligned}$$

Each auxiliary distribution with support $\mathbb{Z}$ gives rise to a lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$. To choose the best auxiliary distribution – i.e., the auxiliary distribution that gives rise to the tightest lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ – we choose a family of auxiliary distributions and select the best member of the family. In practice, an important consideration is that the resulting lower bound is tractable. Therefore, we confine attention to a family of auxiliary distributions under which the resulting lower bounds are tractable. A natural choice is given by a family of auxiliary distributions under which the block memberships are independent:

$$\boldsymbol{Z}_i \overset{\text{ind}}{\sim} \text{Multinomial}(1,\ \boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,K})), \quad i = 1, \ldots, n.$$

By the independence of block memberships under the auxiliary distribution, we obtain the following tractable lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$:

$$\begin{aligned}\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}) &\stackrel{\text{def}}{=} \sum_{\boldsymbol{z}\in\mathbb{Z}} a_{\boldsymbol{\alpha}}(\boldsymbol{z}) \log \frac{p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z})}{a_{\boldsymbol{\alpha}}(\boldsymbol{z})} \\ &= \sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_{i,k}\, \alpha_{j,l} \log p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},z_{i,k}=1,\, z_{j,l}=1,\boldsymbol{z}_{-i,j})}(x_{i,j}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{i,k} (\log \pi_k - \log \alpha_{i,k}),\end{aligned}$$

where $p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},z_{i,k}=1,\, z_{j,l}=1,\boldsymbol{z}_{-i,j})}(x_{i,j})$ denotes the marginal probability mass function of $X_{i,j}$ and $\boldsymbol{z}_{-i,j}$ denotes the block memberships of all nodes excluding nodes i and j.

In practice, we obtain the best lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ by maximizing $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\alpha}$. Direct maximization of $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\alpha}$ is possible but inconvenient, because $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ contains products of $\alpha_{i,k}$ and $\alpha_{j,l}$. As a consequence, a fixed-point update of $\alpha_{i,k}$ depends on $(n-1)K$ other terms $\alpha_{j,l}$ and hence fixed-point updates tend to be time-consuming and get stuck in local maxima, as demonstrated by Vu et al. (2013). Vu et al. (2013) proposed an elegant approach to alleviating the problem by using minorization–maximization (MM) methods (Hunter and Lange, 2004). Such methods construct a minorizing function that approximates $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ but is easier to maximize than $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$. A function $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ of $\boldsymbol{\alpha}$ minorizes $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ at point $\boldsymbol{\alpha}^{(t)}$ at iteration t of an iterative algorithm for maximizing $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ if

$$\begin{aligned}M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}) &\leq \hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}) \quad \text{for all} \quad \boldsymbol{\alpha}, \\ M(\boldsymbol{\alpha}^{(t)}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}) &= \hat{\ell}(\boldsymbol{\alpha}^{(t)}; \boldsymbol{\theta}_1, \boldsymbol{\pi}),\end{aligned}$$

where $\boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}$ are fixed. In other words, $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ is bounded above by $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ for all $\boldsymbol{\alpha}$ and touches $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}$. As a result, increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ with respect to $\boldsymbol{\alpha}$ increases $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$. Vu et al. (2013) showed that the following function minorizes $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ at point $\boldsymbol{\alpha}^{(t)}$:

$$\begin{aligned}M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}) = &\sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(\alpha_{i,k}^2 \frac{\alpha_{j,l}^{(t)}}{2\,\alpha_{i,k}^{(t)}} + \alpha_{j,l}^2 \frac{\alpha_{i,k}^{(t)}}{2\,\alpha_{j,l}^{(t)}} \right) \log p_{\eta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},z_{i,k}=1,\, z_{j,l}=1,\boldsymbol{z}_{-i,j})}(x_{i,j}) \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{i,k} \left[\log \pi_k^{(t)} - \log \alpha_{i,k}^{(t)} + \left(1 - \frac{\alpha_{i,k}}{\alpha_{i,k}^{(t)}} \right) \right].\end{aligned}$$

The minorizing function $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ is easier to maximize than $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$, because it replaces the products of $\alpha_{i,k}$ and $\alpha_{j,l}$ by sums of $\alpha_{i,k}^2$ and $\alpha_{j,l}^2$. An additional advantage is that the maximization of $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ amounts to n quadratic programming problems, which can be solved in parallel.

We therefore propose a two-step estimation approach as described in Table 1. We discuss the two steps below and conclude with some comments on parallel computing.

Step 1. The first step estimates $\boldsymbol{z}$ based on $\boldsymbol{\alpha}$. We do so by increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ with respect to $\boldsymbol{\alpha}_i$ subject to the constraints $0 < \alpha_{i,k} < 1$ $(k = 1, \ldots, K)$ and $\sum_{k=1}^{K} \alpha_{i,k} = 1$ $(i = 1, \ldots, n)$. We increase rather than

Table 1
Two-step estimation approach.

1. Estimate $\boldsymbol{z}$ along with $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1$ by iterating:
   - 1.1 Update $\boldsymbol{\alpha}$ by increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\alpha}^{(t)})$ with respect to $\boldsymbol{\alpha}_i$ subject to $0 < \alpha_{i,k} < 1$ $(k = 1, \ldots, K)$ and $\sum_{k=1}^{K} \alpha_{i,k} = 1$ and denote the update by $\boldsymbol{\alpha}_i^{(t+1)}$ $(i = 1, \ldots, n)$.
   - 1.2 Update $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1$ by maximizing $\hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1$:
     - Update $\pi_k^{(t+1)} = (1/n) \sum_{i=1}^{n} \alpha_{i,k}^{(t+1)}$, $k = 1, \ldots, K$.
     - Update $\boldsymbol{\theta}_1^{(t+1)} = \arg\max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1, \boldsymbol{\pi}^{(t+1)})$.

   Upon convergence, we estimate the block membership indicators $\widehat{\boldsymbol{z}}$ by computing $k = \arg\max_{1 \le l \le K} \widehat{\alpha}_{i,l}$ and setting $\widehat{z}_{i,k} = 1$ and $\widehat{z}_{i,l} = 0$ for all $l \neq k$ $(i = 1, \ldots, n)$, where $\widehat{\boldsymbol{\alpha}}$ denotes the final value of $\boldsymbol{\alpha}$.
2. Estimate $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ by $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta})$ or $\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \tilde{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta})$.

maximize $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$, because maximizing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ is more time-consuming and algorithms maximizing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ are more prone to end up in local maxima than algorithms increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$. Since $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ and $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ depend on $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$ and both are unknown, we iterate between updates of $\boldsymbol{\alpha}$ and updates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$. The updates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$ are based on maximizing $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$ and are identical to the updates of Vu et al. (2013), because $\boldsymbol{\theta}_2 = \mathbf{0}$ reduces the model to a stochastic block model. As a convergence criterion, we use

$$\frac{|\hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\pi}^{(t+1)}) - \hat{\ell}(\boldsymbol{\alpha}^{(t)}; \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\pi}^{(t)})|}{\hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\pi}^{(t+1)})} < \gamma,$$

where $\gamma > 0$ is a small constant. Upon convergence, we estimate the block membership indicators $\widehat{\boldsymbol{z}}$ by computing $k = \arg\max_{1 \le l \le K} \widehat{\alpha}_{i,l}$ and setting $\widehat{z}_{i,k} = 1$ and $\widehat{z}_{i,l} = 0$ for all $l \neq k$ $(i = 1, \ldots, n)$, where $\widehat{\boldsymbol{\alpha}}$ denotes the final value of $\boldsymbol{\alpha}$.

Step 2. We estimate $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ by Monte Carlo maximum likelihood methods (Hunter and Handcock, 2006) or maximum pseudolikelihood methods (Strauss and Ikeda, 1990). Monte Carlo maximum likelihood methods exploit the fact that the loglikelihood function of $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$, defined by

$$\ell_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}) \;=\; \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})}(\boldsymbol{x}),$$

can be written as

$$\ell_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}) \;=\; \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}}),\, s(\boldsymbol{x}) \rangle - \log \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})} \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}}),\, s(\boldsymbol{X}) \rangle),$$

where $\boldsymbol{\theta}_0$ is a fixed parameter vector (e.g., $\boldsymbol{\theta}_0$ may be an educated guess of the data-generating parameter vector). In general, the expectation $\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})}$ is intractable, but it can be estimated by a Monte Carlo sample average based on a Monte Carlo sample of graphs generated under $\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})$. Therefore, we can approximate $\ell_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta})$ by

$$\hat{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}) \;=\; \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}}), s(\boldsymbol{x}) \rangle - \log \widehat{\mathbb{E}}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})} \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}}), s(\boldsymbol{X}) \rangle),$$

where $\widehat{\mathbb{E}}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})}$ is a Monte Carlo approximation of $\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})}$ based on a Monte Carlo sample of graphs generated by using $\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})$. Hence $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ can be estimated by

$$\widehat{\boldsymbol{\theta}} \;=\; \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}).$$

Additional details on Monte Carlo maximum likelihood methods can be found in Hunter and Handcock (2006). We note that the local dependence of the model facilitates parallel computing, which is discussed in the following paragraph. Standard errors of $\widehat{\boldsymbol{\theta}}$ can be based on the estimated Fisher information matrix, although such standard errors are conditional on the estimated block structure $\widehat{\boldsymbol{z}}$ and therefore do not reflect the uncertainty about $\widehat{\boldsymbol{z}}$. A parametric bootstrap approach would be an interesting alternative to capturing the additional uncertainty due to $\widehat{\boldsymbol{z}}$, but would be time-consuming.

An alternative is to estimate $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ by maximum pseudolikelihood estimators (Strauss and Ikeda, 1990). Maximum pseudolikelihood estimators were invented by Besag (1974) to sidestep intractable likelihood functions of Markov random fields in spatial statistics, and are known to be consistent estimators of Markov random fields with exponential-family parameterizations (e.g., Comets, 1992). Strauss and Ikeda (1990) adapted them to exponential-family random graph models. While believed to be inferior to maximum likelihood estimators when the dependence among edges is strong and propagates throughout the random graph (e.g., van Duijn et al., 2009), maximum pseudolikelihood estimators seem to perform well when the dependence is weak, e.g., when the dependence among edges is confined to non-overlapping or overlapping blocks (Stewart and Schweinberger, 2020). The maximum pseudolikelihood estimator is defined as

$$\widetilde{\boldsymbol{\theta}} \;=\; \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \tilde{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}),$$

where

$$\tilde{\ell}_{\hat{\boldsymbol{z}}}(\boldsymbol{\theta}) \;=\; \sum_{i<j}^{n} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\hat{\boldsymbol{z}})}(x_{i,j} \mid \boldsymbol{x}_{-i,j}).$$

Here, $\boldsymbol{x}_{-i,j}$ denotes $\boldsymbol{x}$ excluding $x_{i,j}$, and $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\hat{\boldsymbol{z}})}(x_{i,j} \mid \boldsymbol{x}_{-i,j})$ denotes the conditional probability of $X_{i,j} = x_{i,j}$ given $\boldsymbol{X}_{-i,j} = \boldsymbol{x}_{-i,j}$. Since the conditional distributions of $X_{i,j}$ given the rest of the random graph are Bernoulli distributions, computing maximum pseudolikelihood estimators consumes less time than computing Monte Carlo maximum likelihood estimators.

Parallel computing. In Step 1, the maximization of the minorizing function amounts to n quadratic programming problems, which can be solved in parallel. In Step 2, the local dependence induced by the model implies that the contributions of the between- and within-block subgraphs to the loglikelihood function and its gradient and Hessian can be computed in parallel. Hence both steps can be implemented in parallel, which suggests that the two-step estimation approach can be applied to large networks as long as the blocks are not too large and multi-core computers or computing clusters are available.

5. Simulation results

We assess the performance of the two-step estimation approach by conducting multiple simulation studies, with the number of nodes n ranging from 30 to 10,000, the number of blocks K ranging from 3 to 100, and the sizes of blocks ranging from 5 to 100:

I. A small-scale simulation study to compare the block recovery of the gold standard, the Bayesian approach of Schweinberger and Handcock (2015), to the two-step estimation approach. Since the Bayesian approach is too time-consuming to be applied to large networks, we use small networks with $n = 30$ nodes and $K = 3$ blocks. The $K = 3$ blocks consist of 10 nodes (balanced case) or 5, 10, 15 nodes (unbalanced case).
II. A large-scale simulation study to assess the block recovery in Step 1 of the two-step estimation approach, using $K = 25, 50, 75, 100$ blocks of size $25, 50, 75, 100$.
III. A large-scale simulation study to assess how the block recovery in Step 1 of the two-step estimation approach is affected by the sparsity of between-block subgraphs, using $K = 25$ blocks of size 25.
IV. A large-scale simulation study to assess the parameter recovery in Step 2 of the two-step estimation approach, using $K = 25, 50, 75, 100$ blocks of size $25, 50, 75, 100$.

In Step 1 of the two-step estimation approach, we use the variational approach described in Section 4. We compare the variational approach to the spectral clustering method described in Lei and Rinaldo (2015). Spectral clustering is an alternative to the variational approach in Step 1 of the two-step estimation approach. In Step 2 of the two-step estimation approach, we use maximum pseudolikelihood estimators, which are more scalable than Monte Carlo maximum likelihood estimators and facilitate simulation studies with up to 10,000 nodes. In each scenario, we generate 500 graphs from the model having between-block edge terms and within-block edge and transitive edge terms, as described in Section 2. To select sensible values of the parameter vector $\boldsymbol{\theta}$, note that the probabilistic behavior of random graphs with local dependence is sensitive to the choice of parameter values, and so is the block recovery. The same applies to stochastic block models: e.g., when the probabilities of edges within and between blocks are too low, we may not be able to recover the block structure with high probability (Zhang and Zhou, 2016; Gao et al., 2017). We have therefore selected the parameter vector $\boldsymbol{\theta}$ based on the following considerations: We would like to ensure that, with high probability, the model generates graphs that resemble real-world networks in terms of sufficient statistics $s(\boldsymbol{X})$. In principle, we could select $\boldsymbol{\theta}$ by inspecting the expectation of $s(\boldsymbol{X})$. The problem is that the expectation is not available in closed form. We have therefore selected $\boldsymbol{\theta}$ based on simulating graphs, and checking whether the simulated graphs resemble real-world networks.

In simulation study I, to allow blocks of different sizes to have different parameters, we use size-dependent between-block edge parameters $\theta_{B,k,l} = -.882 \log n$ $(l < k = 1, \ldots, K)$ and within-block edge and transitive edge parameters $\theta_{W,k,k,1} = -.434 \log n_k(\boldsymbol{z})$ and $\theta_{W,k,k,2} = .217 \log n_k(\boldsymbol{z})$ $(k = 1, \ldots, K)$, where $n_k(\boldsymbol{z})$ is the size of block k under $\boldsymbol{z} \in \mathbb{Z}$. The size-dependent parameterization is motivated by the sparsity of random graphs: e.g., if edges $X_{i,j}$ are independent Bernoulli(μ) random variables, it is tempting to believe that there exist constants $c > 0$ and $0 < \alpha < 1$ such that the expected number of edges of each node $(n-1)\,\mu$ is bounded above by $c\, n^{\alpha}$, because real-world networks are sparse. As a consequence, μ should be of order n^{θ} and $\eta = \text{logit}(\mu)$ should be of order $\log n^{\theta} = \theta \log n$, where $\theta = \alpha - 1 < 0$. In more general models with edge terms as well as other model terms, all model terms should scale as the edge term, so that no model term can dominate any other model term. These considerations suggest that parameters should scale with the log number of nodes. In simulations studies II and IV, the sizes of blocks are identical, so we use between-block edge parameters $\theta_{B,k,l} = .5 - \log n$ $(l < k = 1, \ldots, K)$ and within-block edge and transitive edge parameters $\theta_{W,k,k,1} = -1.5$ and $\theta_{W,k,k,2} = .5$ $(k = 1, \ldots, K)$. In simulation study III, we use the same within-block parameters, but between-block edge parameters $\theta_{B,k,l} = .5 - \alpha \log n$ $(l < k = 1, \ldots, K)$, with α varying from .5 and 1.

In all scenarios, we assess block recovery by using Yule's ϕ-coefficient:

$$\phi(\boldsymbol{z}^{\star}, \boldsymbol{z}) \;=\; \frac{n_{0,0}\, n_{1,1} - n_{0,1}\, n_{1,0}}{\sqrt{(n_{0,0} + n_{0,1})(n_{1,0} + n_{1,1})(n_{0,0} + n_{1,0})(n_{0,1} + n_{1,1})}},$$

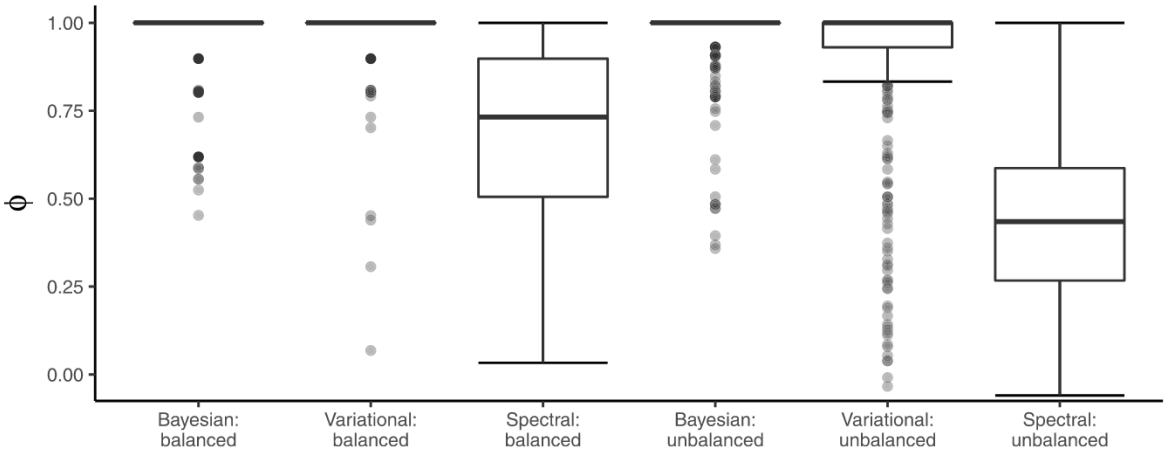

Fig. 1. Block recovery in terms of Yule's ϕ-coefficient. 500 graphs with $n = 30$ nodes and $K = 3$ blocks with 10 nodes (balanced case) or 5, 10, 15 nodes (unbalanced case) were generated. The figure compares the block recovery of the Bayesian approach, the variational approach, and spectral clustering. The Bayesian approach is the gold standard. The variational approach is the default in Step 1 of the two-step estimation approach. Spectral clustering is an alternative to the variational approach in Step 1.

Table 2
Average computing time in seconds of the Bayesian approach, the variational approach, and spectral clustering. 500 graphs with $n = 30$ nodes and $K = 3$ blocks with 10 nodes (balanced case) or 5, 10, 15 nodes (unbalanced case) were generated. The Bayesian approach is the gold standard. The variational approach is the default in Step 1 of the two-step estimation approach. Spectral clustering is an alternative to the variational approach in Step 1. In Step 2 of the two-step estimation approach, maximum pseudolikelihood estimates are computed. The computing times of the variational approach and spectral clustering mentioned above are the total computing times, including both Step 1 and Step 2.

	Bayesian approach	Variational approach	Spectral clustering
$n = 30$, $K = 3$, balanced	42,579.6	23.83	.18
$n = 30$, $K = 3$, unbalanced	32,480.9	24.00	.18

where

$$n_{a,b} \equiv n_{a,b}(\boldsymbol{z}^\star, \boldsymbol{z}) = \sum_{i<j}^{n} \mathbb{1}(\mathbb{1}(\boldsymbol{z}_i^\star = \boldsymbol{z}_j^\star) = a)\, \mathbb{1}(\mathbb{1}(\boldsymbol{z}_i = \boldsymbol{z}_j) = b), \quad a, b \in \{0, 1\}.$$

Here, $\mathbb{1}(.)$ is an indicator function, which is 1 if its argument is true and is 0 otherwise. The quantity $n_{0,0}(\boldsymbol{z}^\star, \boldsymbol{z})$ is the number of pairs of nodes that are assigned to distinct blocks under both $\boldsymbol{z}^\star$ and $\boldsymbol{z}$, while the quantity $n_{1,1}(\boldsymbol{z}^\star, \boldsymbol{z})$ is the number of pairs of nodes that are assigned to the same block under both $\boldsymbol{z}^\star$ and $\boldsymbol{z}$. The sum of the other two quantities, $n_{0,1}(\boldsymbol{z}^\star, \boldsymbol{z}) + n_{1,0}(\boldsymbol{z}^\star, \boldsymbol{z})$, is the number of pairs of nodes on which $\boldsymbol{z}^\star$ and $\boldsymbol{z}$ disagree. If $\boldsymbol{z}^\star$ and $\boldsymbol{z}$ agree on all pairs of nodes, Yule's ϕ-coefficient is 1. In fact, Yule's ϕ-coefficient is bounded above by 1, and is invariant to the labeling of the blocks.

The results of the small-scale simulation study I are shown in Fig. 1. The results suggest that the two-step estimation approach is almost as good as the Bayesian approach in terms of block recovery in the balanced case ($K = 3$ blocks of size 10), but is worse in the unbalanced case ($K = 3$ blocks of size 5, 10, 15). The worse performance in the unbalanced case may be due to the fact that there are smaller blocks in the unbalanced case than in the balanced case, and recovering small blocks is more challenging than recovering large blocks. That said, the advantage of the Bayesian approach over the two-step estimation approach in terms of block recovery in the unbalanced case is limited, and comes at excessive costs: Table 2 reveals that the computing time of the Bayesian approach is thousands of times higher than the computing time of the two-step estimation approach.

The large-scale simulation studies II, III, and IV shed light on the performance of Steps 1 and 2 of the two-step estimation approach in large networks. The results of simulation study II, shown in Fig. 2, reveal that the variational approach used in Step 1 of the two-step estimation approach outperforms spectral clustering in terms of block recovery in most scenarios. Simulation study III shows how the block recovery in Step 1 of the two-step estimation approach is affected by between-block subgraph sparsity. According to Fig. 3, the more sparse between-block subgraphs are, the more the dense within-block subgraphs "stand out", improving block recovery. Last, but not least, simulation study IV helps assess parameter recovery in Step 2 of the two-step estimation approach. Fig. 4 shows that maximum pseudolikelihood estimates are close to the data-generating parameters, and more so when the blocks are small and the number of blocks is large.

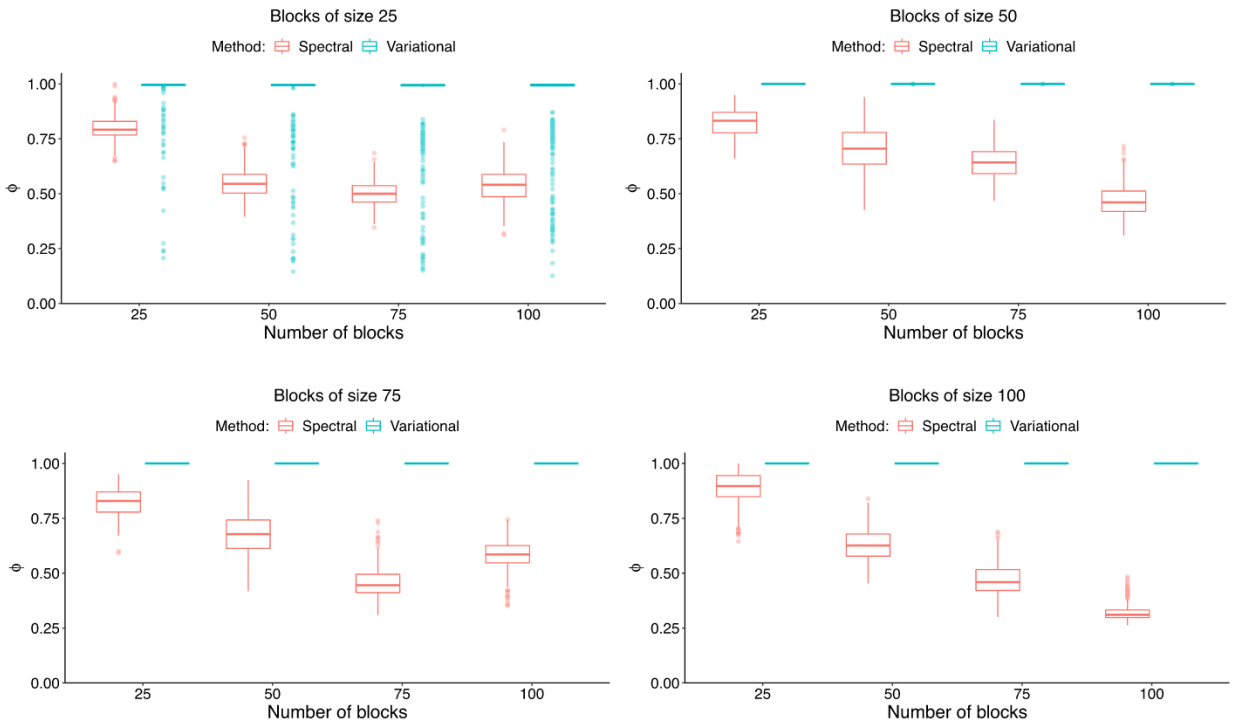

Fig. 2. Block recovery in terms of Yule's ϕ-coefficient. 500 graphs with $K = 25, 50, 75, 100$ blocks of size $25, 50, 75, 100$ were generated from the model with between-block edge parameters $\theta_{B,k,l} = .5 - \log n$ $(l < k = 1, \ldots, K)$ and within-block edge and transitive edge parameters $\theta_{W,k,k,1} = -1.5$ and $\theta_{W,k,k,2} = .5$ $(k = 1, \ldots, K)$. The figure compares two alternative approaches to recovering the block structure in Step 1 of the two-step estimation approach, the variational approach and spectral clustering.

6. Amazon product recommendation network

We use the two-step estimation approach to shed light on the structure of a large Amazon product recommendation network that is not captured by stochastic block models. The data on the Amazon product recommendation network were collected by Yang and Leskovec (2015) and can be downloaded from the website

http://snap.stanford.edu/data/com-Amazon.html

The network consists of products listed at www.amazon.com. Two products i and j are connected by an edge if i and j are frequently purchased together according to the "Customers Who Bought This Item Also Bought" feature at www.amazon.com. Amazon assigns all products to categories, which we consider to be ground-truth blocks. We use a subset of the network consisting of the top 500 non-overlapping categories with 10 to 80 products, where the ranking of categories is based on Yang and Leskovec (2015). The resulting network consists of 10,448 products and 33,537 edges.

To model the Amazon product recommendation network, we take advantage of curved exponential-family random graph models with within-block edge and geometrically weighted degree and edgewise shared partner terms (Snijders et al., 2006; Hunter and Handcock, 2006; Hunter et al., 2008). The resulting models are more general than the motivating example used in Sections 2 and 3 – the model with between-block edge terms and within-block edge and transitive edge terms – and ensure that, for each pair of products, the added value of additional edges and triangles within blocks decays. In fact, transitive edge terms are special cases of geometrically weighted edgewise shared partner terms, and both of them are well-behaved alternatives to the ill-behaved triangle terms mentioned in Section 2. A full-fledged discussion of those models is beyond the scope of our paper. We refer the interested reader to the seminal papers of Snijders et al. (2006), Hunter and Handcock (2006), and Hunter et al. (2008).

The natural parameters of the within-block edge terms are given by

$$\eta_{W,k,1}(\boldsymbol{\theta}, \boldsymbol{z}) \;=\; \theta_1 \, \log n_k(\boldsymbol{z}),$$

where the logarithmic term $\log n_k(\boldsymbol{z})$ arises from sparsity considerations and is a simple form of a size-dependent parameterization, as explained in Section 5. The within-block geometrically weighted degree terms are based on the number of products with t edges in block $\mathcal{A}_k$. The natural parameters of within-block geometrically weighted degree terms are given by

$$\eta_{W,k,2,t}(\boldsymbol{\theta}, \boldsymbol{z}) \;=\; \theta_2 \, \log n_k(\boldsymbol{z}) \, \exp(\theta_3) \left[1 - (1 - \exp(-\theta_3))^t\right], \quad t = 1, \ldots, n_k(\boldsymbol{z}) - 1.$$

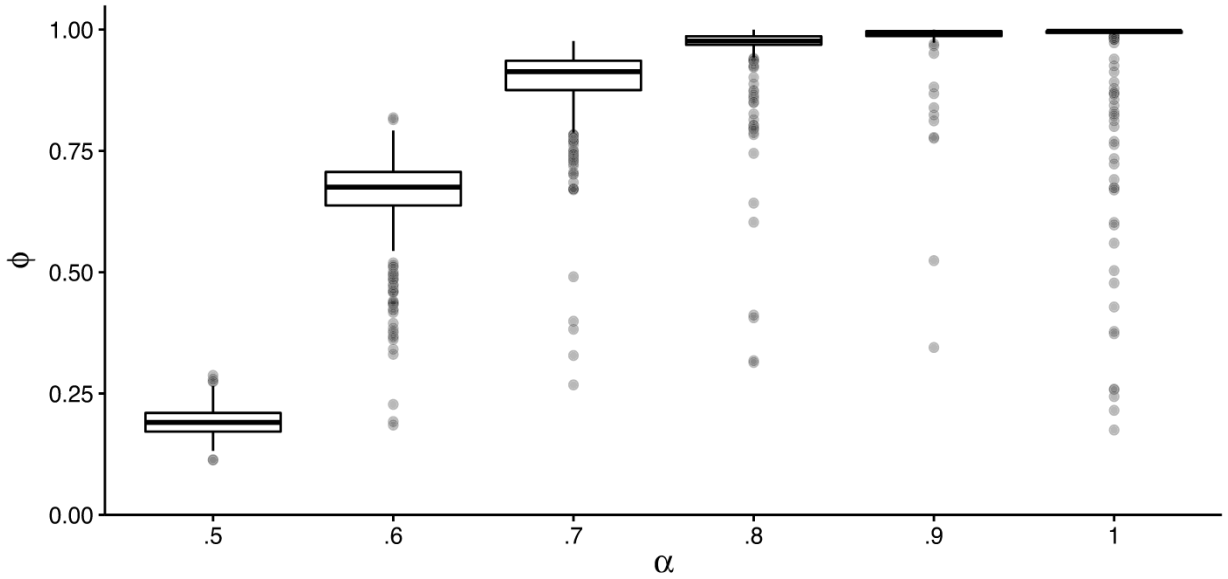

Fig. 3. Block recovery in terms of Yule's ϕ-coefficient as a function of between-block subgraph sparsity. 500 graphs with $K = 25$ blocks of size 25 were generated from the model with between-block edge parameters $\theta_{B,k,l} = .5 - \alpha \log n$ $(l < k = 1, \ldots, K)$ and within-block edge and transitive edge parameters $\theta_{W,k,k,1} = -1.5$ and $\theta_{W,k,k,2} = .5$ $(k = 1, \ldots, K)$, with α varying from .5 to 1. The figure demonstrates that the more sparse between-block subgraphs are, the more the dense within-block subgraphs "stand out", improving block recovery.

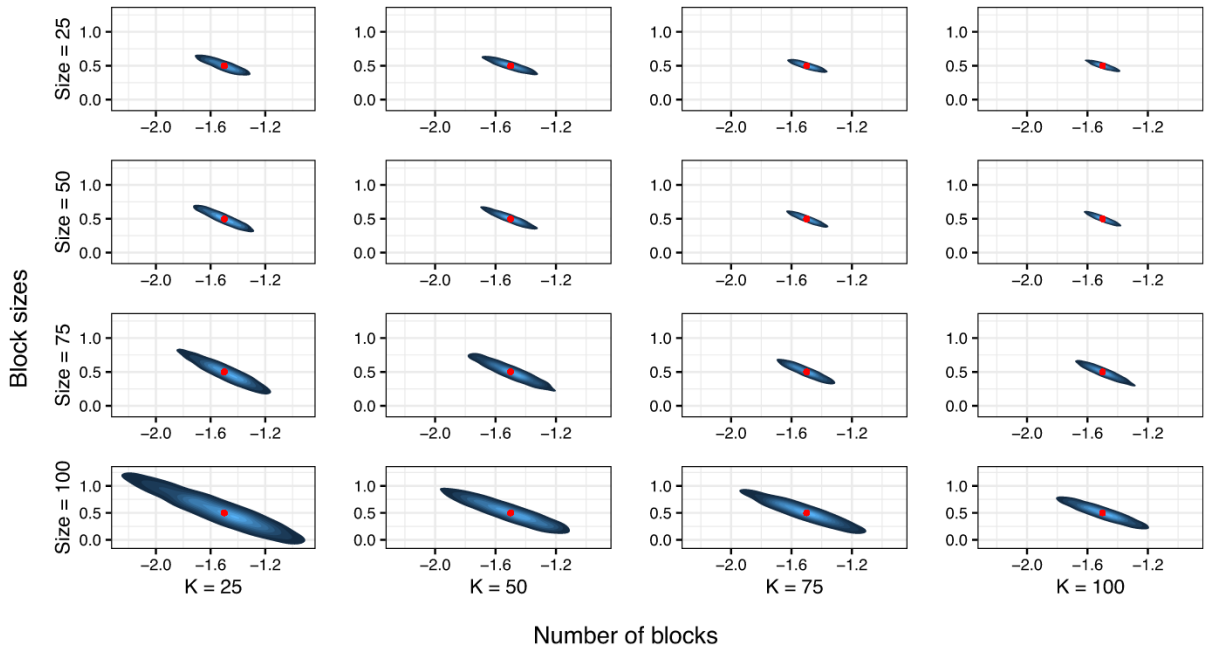

Fig. 4. Maximum pseudolikelihood estimates of within-block parameters. 500 graphs with $K = 25, 50, 75, 100$ blocks of size $25, 50, 75, 100$ were generated from the model with between-block edge parameters $\theta_{B,k,l} = .5 - \log n$ $(l < k = 1, \ldots, K)$ and within-block edge and transitive edge parameters $\theta_{W,k,k,1} = -1.5$ and $\theta_{W,k,k,2} = .5$ $(k = 1, \ldots, K)$. The horizontal axis corresponds to $\theta_{W,k,k,1} = -1.5$, while the vertical axis corresponds to $\theta_{W,k,k,2} = .5$ $(k = 1, \ldots, K)$. The red circle in each plot indicates the data-generating within-block parameter vector $(\theta_{W,k,k,1}, \theta_{W,k,k,2}) = (-1.5, .5)$.

The within-block geometrically weighted edgewise shared partner terms are based on the number of connected pairs of products i and j in block $\mathcal{A}_k$ such that i and j have t shared partners in block $\mathcal{A}_k$. The natural parameters of the within-block geometrically weighted edgewise shared partner terms are given by

$$\eta_{W,k,3,t}(\boldsymbol{\theta}, \boldsymbol{z}) \;=\; \theta_4 \, \log n_k(\boldsymbol{z}) \, \exp(\theta_5) \left[1 - (1 - \exp(-\theta_5))^t\right], \quad t = 1, \ldots, n_k(\boldsymbol{z}) - 2.$$

Table 3
Monte Carlo maximum likelihood estimates and standard errors (S.E.) of $\theta_1, \ldots, \theta_6$ estimated from the Amazon product recommendation network with 10,448 products; note that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$ should not be confused with the size-dependent natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$. The parameters θ_1 and θ_6 are the base weights of the within- and between-block edge terms, respectively. The parameters θ_2 and θ_4 are the base weights of the within-block degree and edgewise shared partner terms, whereas θ_3 and θ_5 control the rate of decay of the added value of additional edges and edgewise shared partners, respectively.

Term	Estimate	S.E.	Estimate	S.E.
Within-block edges θ_1	−.369	.002	−1.410	.008
Within-block degrees θ_2 (base parameter)			.742	.020
Within-block degrees θ_3 (decay parameter)			.910	.023
Within-block shared partners θ_4 (base parameter)			.303	.005
Within-block shared partners θ_5 (decay parameter)			1.106	.012
Between-block edges θ_6	−1.199	.004	−1.199	.004

To reduce computing time, it is convenient to truncate the two geometrically weighted model terms by setting $\eta_{W,k,2,t}(\boldsymbol{\theta}, \boldsymbol{z}) = 0$ $(t = 21, \ldots, n_k(\boldsymbol{z}) - 1)$ and $\eta_{W,k,3,t}(\boldsymbol{\theta}, \boldsymbol{z}) = 0$ $(t = 13, \ldots, n_k(\boldsymbol{z}) - 2)$. The two thresholds 21 and 13 are motivated by the fact that no product has 21 or more edges and less than 1% of all pairs of products has 13 or more edgewise shared partners. Last, but not least, the natural parameters of the between-block edge terms are given by

$$\eta_{B,k,l}(\boldsymbol{\theta}, \boldsymbol{z}) = \theta_6 \log n, \quad l < k.$$

The resulting exponential family is a curved exponential family (Hunter and Handcock, 2006), because the natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$ of the exponential family is a nonlinear function of $\boldsymbol{\theta}$ given $\boldsymbol{z} \in \mathbb{Z}$. The natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$ depends on the sizes of blocks, because we do not want to force small and large blocks to have the same natural parameters, as explained in Section 5.

The curved exponential-family model specified above can capture an excess in the expected number of triangles within blocks relative to stochastic block models, while ensuring that, for each pair of products, the added value of additional edges and triangles within blocks decays (Stewart et al., 2019). An excess in the expected number of triangles within blocks relative to stochastic block models can arise when, e.g., (a) three products are similar (e.g., three books on the same topic); (b) three products are dissimilar but complement each other (e.g., a bicycle helmet, head light, and tail light); and (c) three products, either similar or dissimilar, were produced by the same source (e.g., three books written by the same author).

Since we know the number of ground-truth blocks, we set $K = 500$ and estimate the block structure by using the two-step estimation approach, using the variational approach in Step 1 and Monte Carlo maximum likelihood estimates in Step 2. To assess the performance of the two-step estimation approach in terms of block recovery, we use Yule's ϕ-coefficient. Yule's ϕ-coefficient turns out to be .964, which indicates near-perfect recovery of the ground-truth block structure. An inspection of the estimated block structure reveals that, out of the 500 categories of products in the Amazon product recommendation network, products in 15 categories are misclassified. Some of the small categories are merged with large categories, while some unusual products of large blocks are misclassified as well. Some of the products are unusual in the sense of having few edges to other products in the same category, while others are unusual in the sense of having many edges to other products in the same category.

The Monte Carlo maximum likelihood estimates and standard errors of $\theta_1, \ldots, \theta_6$ are shown in Table 3. The parameters $\theta_1, \ldots, \theta_6$ of the geometrically weighted terms can be interpreted in terms of log odds of conditional probabilities of edges, given all other edges (Hunter, 2007). A worked-out example can be found in Stewart et al. (2019). Table 3 suggests that there is evidence for transitivity. The observed tendency towards transitivity has advantages in practice: It enables Amazon to recommend co-purchases of brandnew products and existing products, even when there are no data on past co-purchases of these products. For example, when a brandnew product i is introduced (e.g., a novel) and product i is known to be related to existing product j (e.g., a novel by the same author), and product j tends to be co-purchased with product k (e.g., a classic novel), then Amazon could recommend co-purchases of products i and j as well as products i and k, despite the fact that there are no data on past co-purchases of products i and k and there is no direct connection between products i and k (although there in indirect connection via product j).

To demonstrate that the curved exponential-family random graph model considered here can capture structural features of networks that simpler models – such as stochastic block models – cannot capture, we compare the goodness-of-fit of the curved exponential-family random graph model to the goodness-of-fit of stochastic block models. Since the two models impose the same probability law on between-block subgraphs, it is natural to compare the two models in terms of goodness-of-fit with respect to within-block subgraphs. We assess the goodness-of-fit of the two models in terms of the within-block geodesic distances of pairs of products, i.e., the length of the shortest path between pairs of products in the same block; the numbers of within-block dyadwise shared partners, i.e., the number of unconnected or connected pairs of products with i shared partners in the same block; the numbers of within-block edgewise shared partners, i.e., the number of connected pairs of products with i shared partners in the same block; and the number of within-block transitive edges, i.e., the number of pairs of products with at least one shared partner in the same block. Figs. 5 and 6 compare the

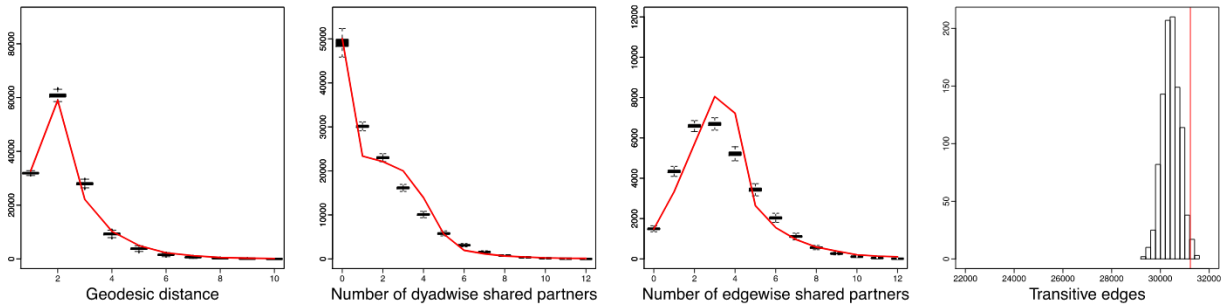

Fig. 5. Amazon product recommendation network with 10,448 products: goodness-of-fit of curved exponential-family random graph model. The red lines indicate observed values of statistics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

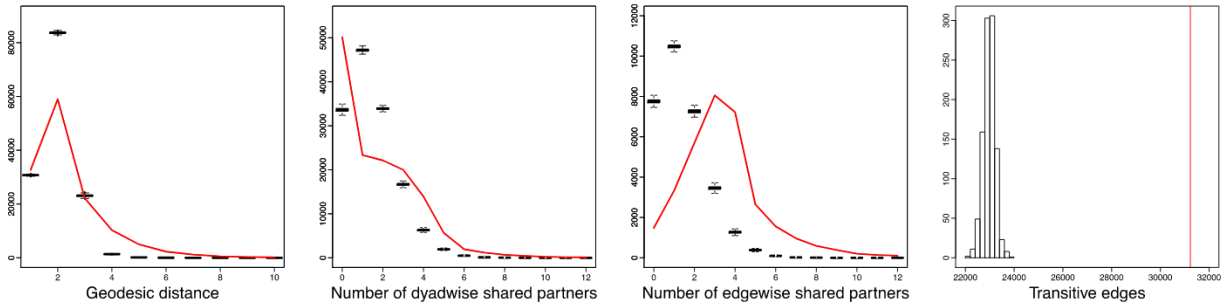

Fig. 6. Amazon product recommendation network with 10,448 products: goodness-of-fit of stochastic block models. The red lines indicate observed values of statistics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

goodness-of-fit of the two models based on 1,000 graphs simulated from the estimated models. The figures suggest that the curved exponential-family random graph model considered here is superior to the stochastic block model in terms of both connectivity and transitivity.

7. Discussion

The two-step estimation approach proposed here enables large-scale estimation of models with local dependence and unknown block structure provided that the number of blocks K is known. An important direction of future research are methods for selecting K when K is unknown. We note that even in the special case of stochastic block models, the issue of selecting K has not received much attention, with the notable exception of recent work by Saldana et al. (2017), Wang and Bickel (2017), and others. Whether – and how – the developed methods can be extended to models with local dependence is an open question, but having scalable methods for selecting K would doubtless be useful in practice.

The proposed methods are implemented in `R` packages `hergm` (Schweinberger and Luna, 2018). A stable and user-friendly version will be released in the near future.

Acknowledgments

We acknowledge support from the National Science Foundation (NSF), United States of America in the form of NSF awards DMS-1513644 (SB, JS, MS) and DMS-1812119 (JS, MS).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.107029.

References

Amini, A.A., Chen, A., Bickel, P.J., Levina, E., 2013. Pseudo-likelihood methods for community detection in large sparse networks. Ann. Statist. 41, 2097–2122.

Atchade, Y.F., Lartillot, N., Robert, C., 2013. Bayesian computation for statistical models with intractable normalizing constants. Braz. J. Probab. Stat. 27, 416–436.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 192–225.

Bickel, P.J., Chen, A., 2009. A nonparametric view of network models and Newman-Girvan and other modularities. Proc. Nat. Acad. Sci. 106, 21068–21073.
Bickel, P.J., Chen, A., Levina, E., 2011. The method of moments and degree distributions for network models. Ann. Statist. 39, 2280–2301.
Bickel, P.J., Choi, D., Chang, X., Zhang, H., 2013. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. Ann. Statist. 41, 1922–1943.
Binkiewicz, N., Vogelstein, J.T., Rohe, K., 2017. Covariate-assisted spectral clustering. Biometrika 104, 361–377.
Brown, L., 1986. Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Hayworth, CA, USA.
Byshkin, M., Stivala, A., Mira, A., Robins, G., Lomi, A., 2018. Fast maximum likelihood estimation via equilibrium expectation for large network data. Sci. Rep. 8, 2045–2322.
Caimo, A., Friel, N., 2011. Bayesian inference for exponential random graph models. Social Networks 33, 41–55.
Celisse, A., Daudin, J.J., Pierre, L., 2012. Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electron. J. Stat. 6, 1847–1899.
Chatterjee, S., Diaconis, P., 2013. Estimating and understanding exponential random graph models. Ann. Statist. 41, 2428–2461.
Choi, D.S., Wolfe, P.J., Airoldi, E.M, 2012. Stochastic blockmodels with growing number of classes. Biometrika 99, 273–284.
Comets, F., 1992. On consistency of a class of estimators for exponential families of Markov random fields on the lattice. Ann. Statist. 20, 455–468.
Daudin, J.J., Picard, F., Robin, S., 2008. A mixture model for random graphs. Stat. Comput. 18, 173–183.
van Duijn, M.A.J., Gile, K., Handcock, M.S, 2009. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. Social Networks 31, 52–62.
Erdős, P., Rényi, A., 1959. On random graphs. Publ. Math. 6, 290–297.
Erdős, P., Rényi, A., 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17–61.
Fienberg, S., 2012. A brief history of statistical models for network analysis and open challenges. J. Comput. Graph. Statist. 21, 825–839.
Frank, O., Strauss, D., 1986. Markov graphs. J. Amer. Statist. Assoc. 81, 832–842.
Gao, C., Ma, Z., Zhang, A.Y., Zhou, H.H, 2017. Achieving optimal misclassification proportion in stochastic block models. J. Mach. Learn. Res. 18, 1980–2024.
Häggström, O., Jonasson, J., 1999. Phase transition in the random triangle model. J. Appl. Probab. 36, 1101–1115.
Handcock, M.S, 2003. Assessing Degeneracy in Statistical Models of Social Networks. Tech. Rep., Center for Statistics and the Social Sciences, University of Washington, http://www.csss.washington.edu/Papers.
Handcock, M.S., Raftery, A.E., Tantrum, J.M, 2007. Model-based clustering for social networks. J. R. Stat. Soc. A 170, 301–354.
Harris, J.K, 2013. An Introduction to Exponential Random Graph Modeling. Sage, Thousand Oaks, California.
Hoff, P.D, 2020. Additive and multiplicative effects network models. Stat. Sci. in press.
Hoff, P.D., Raftery, A.E., Handcock, M.S, 2002. Latent space approaches to social network analysis. J. Amer. Statist. Assoc. 97, 1090–1098.
Holland, P.W., Leinhardt, S., 1970. A method for detecting structure in sociometric data. Am. J. Sociol. 76, 492–513.
Holland, P.W., Leinhardt, S., 1972. Some evidence on the transitivity of positive interpersonal sentiment. Am. J. Sociol. 77, 1205–1209.
Holland, P.W., Leinhardt, S., 1976. Local structure in social networks. Sociol. Methodol. 1–45.
Hummel, R.M., Hunter, D.R., Handcock, M.S, 2012. Improving simulation-based algorithms for fitting ERGMs. J. Comput. Graph. Statist. 21, 920–939.
Hunter, D.R., 2007. Curved exponential family models for social networks. Social Networks 29, 216–230.
Hunter, D.R., Goodreau, S.M., Handcock, M.S, 2008. Goodness of fit of social network models. J. Amer. Statist. Assoc. 103, 248–258.
Hunter, D.R., Handcock, M.S, 2006. Inference in curved exponential family models for networks. J. Comput. Graph. Statist. 15, 565–583.
Hunter, D.R., Krivitsky, P.N., Schweinberger, M., 2012. Computational statistical methods for social network models. J. Comput. Graph. Statist. 21, 856–882.
Hunter, D.R., Lange, K., 2004. A tutorial on MM algorithms. Amer. Statist. 58, 30–38.
Jin, I.H., Liang, F., 2013. Fitting social network models using varying truncation stochastic approximation MCMC algorithm. J. Comput. Graph. Statist. 22, 927–952.
Jonasson, J., 1999. The random triangle model. J. Appl. Probab. 36, 852–876.
Kolaczyk, E.D., 2009. Statistical Analysis of Network Data: Methods and Models. Springer-Verlag, New York.
Krivitsky, P.N, 2017. Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. Comput. Statist. Data Anal. 107, 149–161.
Lazega, E., Snijders, T.A.B. (Eds.), 2016. Multilevel Network Analysis for the Social Sciences. Springer-Verlag, Switzerland.
Lei, J., Rinaldo, A., 2015. Consistency of spectral clustering in stochastic block models. Ann. Statist. 43, 215–237.
Liang, F., Jin, I.H., Song, Q., Liu, J.S, 2016. An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. J. Amer. Statist. Assoc. 111, 377–393.
Lusher, D., Koskinen, J., Robins, G., 2013. Exponential Random Graph Models for Social Networks. Cambridge University Press, Cambridge, UK.
Mele, A., 2017. A structural model of dense network formation. Econometrica 85, 825–850.
Nowicki, K., Snijders, T.A B., 2001. Estimation and prediction for stochastic blockstructures. J. Amer. Statist. Assoc. 96, 1077–1087.
Okabayashi, S., Geyer, C.J, 2012. Long range search for maximum likelihood in exponential families. Electron. J. Stat. 6, 123–147.
Priebe, C.E., Sussman, D.L., Tang, M., Vogelstein, J.T, 2012. Statistical inference on errorfully observed graphs. J. Amer. Statist. Assoc. 107, 1119–1128.
Rohe, K., Chatterjee, S., Yu, B., 2011. Spectral clustering and the high-dimensional stochastic block model. Ann. Statist. 39, 1878–1915.
Rohe, K., Qin, T., Fan, H., 2014. The highest-dimensional stochastic block model with a regularized estimator. Statist. Sinica 24, 1771–1786.
Saldana, D.F., Yu, Y., Feng, Y., 2017. How many communities are there? J. Comput. Graph. Statist. 26, 171–181.
Salter-Townshend, M., White, A., Gollini, I., Murphy, T.B, 2012. Review of statistical network analysis: models, algorithms, and software. Stat. Anal. Data Min. 5, 243–264.
Schweinberger, M., 2011. Instability, sensitivity, and degeneracy of discrete exponential families. J. Amer. Statist. Assoc. 106, 1361–1370.
Schweinberger, M., 2020. Consistent structure estimation of exponential-family random graph models with block structure. Bernoulli 26, 1205–1233.
Schweinberger, M., Handcock, M.S, 2015. Local dependence in random graph models: characterization, properties and statistical inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 77, 647–676.
Schweinberger, M., Krivitsky, P.N., Butts, C.T., Stewart, J., 2020. Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios. Statist. Sci. in press.
Schweinberger, M., Luna, P., 2018. HERGM: Hierarchical exponential-family random graph models. J. Stat. Softw. 85, 1–39.
Schweinberger, M., Stewart, J., 2020. Concentration and consistency results for canonical and curved exponential-family models of random graphs. Ann. Statist. 48, 374–396.
Sewell, D.K., Chen, Y., 2015. Latent space models for dynamic networks. J. Amer. Statist. Assoc. 110, 1646–1657.
Smith, A.L., Asta, D.M., Calder, C.A, 2019. The geometry of continuous latent space models for network data. Statist. Sci. 34, 428–453.
Snijders, T.A.B., 2002. Markov chain Monte Carlo estimation of exponential random graph models. J. Soc. Struct. 3, 1–40.

Snijders, T.A.B., 2007. Contribution to the discussion of Handcock, M.S., Raftery, A.E., and J.M. Tantrum, Model-based clustering for social networks. J. R. Stat. Soc. Ser. A 170, 322–324.
Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S, 2006. New specifications for exponential random graph models. Sociol. Methodol. 36, 99–153.
Stewart, J., Schweinberger, M., 2020. Scalable estimation of random graph models with dependent edges and parameter vectors of increasing dimension. Tech. Rep., Department of Statistics, Rice University.
Stewart, J., Schweinberger, M., Bojanowski, M., Morris, M., 2019. Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. Social Networks 59, 98–119.
Strauss, D., 1986. On a general class of models for interaction. SIAM Rev. 28, 513–527.
Strauss, D., Ikeda, M., 1990. Pseudolikelihood estimation for social networks. J. Amer. Statist. Assoc. 85, 204–212.
Tan, L.S.L., Friel, N., 2020. Bayesian variational inference for exponential random graph models. J. Comput. Graph. Statist. in press.
Thiemichen, S., Kauermann, G., 2017. Stable exponential random graph models with non-parametric components for large dense networks. Social Networks 49, 67–80.
Vu, D.Q., Hunter, D.R., Schweinberger, M., 2013. Model-based clustering of large networks. Ann. Appl. Stat. 7, 1010–1039.
Wang, Y.X., Bickel, P.J, 2017. Likelihood-based model selection for stochastic block models. Ann. Stat. 45, 500–528.
Yang, J., Leskovec, J., 2015. Defining and evaluating network communities based on ground-truth. Knowl. Inf. Syst. 42, 181–213.
Zhang, A.Y., Zhou, H.H, 2016. Minimax rates of community detection in stochastic block models. Ann. Statist. 44, 2252–2280.