Sankhyā : The Indian Journal of Statistics
2020, Volume 82-A, Part 2, pp. 448-476
© 2019, Indian Statistical Institute

Posterior Contraction Rates for Stochastic Block Models

Prasenjit Ghosh, Debdeep Pati and Anirban Bhattacharya
Texas A&M University, College Station, USA

Abstract

With the advent of structured data in the form of social networks, genetic circuits and protein interaction networks, statistical analysis of networks has gained popularity over recent years. The stochastic block model constitutes a classical cluster-exhibiting random graph model for networks. There is a substantial amount of literature devoted to proposing strategies for estimating and inferring parameters of the model, both from classical and Bayesian viewpoints. Unlike the classical counterpart, there is a dearth of theoretical results on the accuracy of estimation in the Bayesian setting. In this article, we undertake a theoretical investigation of the posterior distribution of the parameters in a stochastic block model. In particular, we show that one obtains near-optimal rates of posterior contraction with routinely used multinomial-Dirichlet priors on cluster indicators and uniform or general Beta priors on the probabilities of the random edge indicators. Our theoretical results are corroborated through a small scale simulation study.

AMS (2000) subject classification. Primary 62G07, 62G20; secondary 60K35.
Keywords and phrases. Bayesian asymptotics, Stochastic block models, Clustering, Multinomial-Dirichlet, Networks, Posterior contraction, Random graphs

1 Introduction

Data available in the form of networks are increasingly becoming common in applications ranging from brain connectivity, protein interactions, web applications and social networks to name a few, motivating an explosion of activity in the statistical analysis of networks in recent years (Goldenberg et al., 2010). Estimating large networks offers unique challenges in terms of structured dimension reduction and estimation in stylized domains, necessitating new tools for inference. A rich variety of probabilistic models have been studied for network estimation, ranging from the classical Erdos and Renyi graphs (Erdős and Rényi, 1961), exponential random graph models (Holland and Leinhardt, 1981), stochastic block models (Holland et al., 1983), Markov Graphs (Frank and Strauss, 1986) and latent space models (Hoff et al., 2002) to name a few.

In a network with n nodes, there are $O(n^2)$ possible connections between pairs of nodes, the exact number depending on whether the network is directed/undirected and whether self-loops are permitted. A common goal of the parametric models mentioned previously is to parsimoniously represent the $O(n^2)$ probabilities of connections between pairs of nodes in terms of fewer parameters. The stochastic block model achieves this by clustering the nodes into $k \ll n$ groups, with the probability of an edge between two nodes solely dependent on their cluster memberships. The block model originated in the mathematical sociology literature (Holland et al., 1983), with subsequent widespread applications in statistics (Wang and Wong, 1987; Snijders and Nowicki, 1997; Nowicki and Snijders, 2001). In particular, the clustering property of block models offers a natural way to find *communities* within networks, inspiring a large literature on community detection (Bickel and Chen, 2009; Newman, 2012; Zhao et al., 2012; Karrer and Newman, 2011; Zhao et al., 2011; Amini et al., 2013). Various modifications of the stochastic block model have also been proposed, including the mixed membership stochastic block model (Airoldi et al., 2009) and degree-corrected stochastic block model (Dasgupta et al., 2004; Karrer and Newman, 2011).

Statistical accuracy of parameter estimates for inference in stochastic block models is of growing interest, with one of the objects of interest being the $n \times n$ matrix of probabilities of edges between pairs of nodes, which we shall denote by $\theta = (\theta_{ij})$. Using a singular-value thresholding approach, Chatterjee (2014) obtained a $\sqrt{k/n}$ rate for estimating θ with respect to the squared ℓ_2 distance in a k-component stochastic block model. In a recent technical report, Gao et al. (2015) obtained an improved $k^2/n^2 + \log k/n$ rate by considering a least-squares type estimator. They also showed that the resulting rate is minimax-optimal; interestingly the minimax rate comprises of two parts which (Gao et al., 2015) refer to as the *nonparametric* and *clustering* rates respectively. Among other related work, Bickel et al. (2013) provided conditions for asymptotic normality of maximum likelihood estimators in stochastic block models.

In this article, we consider a Bayesian formulation of a stochastic block model where θ is equipped with a hierarchical prior and study the contraction of the posterior distribution assuming the data to be generated from a stochastic block model. We show that one obtains the minimax rate of posterior contraction with essentially automatic prior choices, such as multinomial-Dirichlet priors on cluster indicators and uniform[1] on the

[1]Our result continues to hold for general Beta priors on the edge-inclusion probabilities.

probability of the random edge indicators. Such priors are commonly used and there is a sizable literature (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001; Golightly and Wilkinson, 2005; McDaid et al., 2013) on posterior sampling and inference in the stochastic block model. The theoretical development of our present work assumes the knowledge of the number of clusters *a priori*. In a different perspective, when such prior knowledge is unavailable, Geng et al. (2018) proposed an efficient Markov Chain Monte Carlo (MCMC) algorithm to simultaneously estimate the number of unknown clusters and clustering structure. While preparing this manuscript, we also came to know about some recent studies relating to various theoretical properties of such stochastic block models. For instance, Gao et al. (2018) considers a general unified framework of structured linear models that covers many complex statistical problems such as stochastic block models, bi-clustering, sparse linear regression, regression under group sparsity, multi-task learning and dictionary learning. The authors of this paper study the posterior contraction rate of their newly proposed elliptical Laplace distribution under this general set up. Refer also to Channarond et al. (2012), Suwan et al. (2016), van der Pas and van der Vaart (2018), & Hayashi et al. (2016), among others for recent works on the theoretical investigation of various aspects of the SBM.

Theoretical investigation of the posterior distribution in block models offers some unique challenges relative to the small but growing literature on posterior contraction in high-dimensional *sparse* problems (Castillo and van der Vaart 2012, 2015; Pati et al. 2014; Banerjee and Ghosal, 2014). When a large subset of the parameters are exactly or approximately zero, the sparsity assumption can be exploited to reduce the complexity of the model space to derive tests for the true parameter versus the complement of a neighborhood of the true parameter (Castillo and van der Vaart, 2012; Pati et al., 2014). It is now well appreciated that constructing such tests plays a crucial role in posterior asymptotics (Schwartz, 1965; Barron, 1988, 1999; Ghosal et al., 2000). In the present setting, we exploit the parsimonious structure of the parameter space as a result of clustering of n nodes into $k << n$ communities to derive such tests. This also enables us to reduce the "effective" number of parameters (the edge probabilities) to be estimated from $O(n^2)$ to $O(k^2+n)$. This dimension reduction is enabled by exploiting the structure of the model unlike the traditional notion of sparsity in typical sparse high-dimensional studies where a subset of the parameters are zero or negligible in magnitude.

The remainder of the paper is organized as follows. Some notations are introduced in Section 2. We provide an overview of the stochastic block models in Section 3. Our main theoretical results on posterior contraction

are stated in Section 4. While the proof of Theorem 4.1 is given at the end of Section 4.2, proof of other main theoretical results are deferred to the Appendix. A small-scale simulation study is presented in Section 5 and some additional simulation results are also given in the Appendix. We conclude the paper with some discussions in Section 6.

2 Preliminaries

For $\mathcal{S} \subset \mathbb{R}$, we shall denote the set of all $d \times d$ matrices with entries in $\mathcal{S}$ by $\mathcal{S}^{d\times d}$. For any $B = (B_{ll'}) \in \mathbb{R}^{d\times d}$, we denote the Euclidean (equivalently Frobenius) norm of B by $\|B\| = \sqrt{\sum_{l=1}^{d}\sum_{l'=1}^{d} B_{ll'}^2}$. Given $X^* \in \mathbb{R}^{d\times d}, W \in \mathbb{R}_+^{d\times d}$, let $\xi_{d^2}(X^*; W)$ denote the unit ellipsoid with *center* X^* and *weight* W given by

$$\xi_{d^2}(X^*; W) = \left\{ X \in \mathbb{R}^{d\times d} : \sum_{l=1}^{d}\sum_{l'=1}^{d} W_{ll'}(X_{ll'} - X^*_{ll'})^2 \le 1 \right\}. \tag{2.1}$$

Viewed as a subset of $\mathbb{R}^{d^2}$, the Euclidean volume of $\xi_{d^2}(X^*; W)$, denoted by $|\xi_{d^2}(X^*; W)|$, is

$$|\xi_{d^2}(X^*; W)| = \frac{\pi^{d^2}}{\Gamma(d^2/2+1)} \prod_{l=1}^{d}\prod_{l'=1}^{d} W_{ll'}^{-1/2}. \tag{2.2}$$

Given sequences $\{a_n\}, \{b_n\}$, $a_n \lesssim b_n$ indicates there exists a constant $K > 0$ such that $a_n \le K b_n$ for all large n. We say $a_n \asymp b_n$ when $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given any function f and some subset A in its domain, we denote by $f(A)$ the image of A under f. Throughout, C, C' denote positive constants whose values might change from one line to the next.

3 Stochastic Block Models

Let $A = (A_{ij}) \in \{0,1\}^{n\times n}$ denote the adjacency matrix of a network with n nodes, with $A_{ij} = 1$ indicating the presence of an edge from node i to node j and $A_{ij} = 0$ indicating a lack thereof. To keep the subsequent notation clean, we shall consider directed networks with self-loops so that A_{ij} and A_{ji} need not be the same and A_{ii} can be both 0 and 1. Our theoretical results can be modified to undirected networks with or without self-loops in a straightforward fashion; refer to Section 4.1 for further discussion.

Let θ_{ij} denote the probability of an edge from node i to j, with $A_{ij} \sim$ Bernoulli(θ_{ij}) independently for $1 \le i, j \le n$. A stochastic block model postulates that the nodes are clustered into communities, with the probability

of an edge between two nodes solely dependent on their community memberships. Specifically, let $z_i \in \{1, \ldots, k\}$ denote the cluster membership of the ith node and $Q = (Q_{rs}) \in [0,1]^{k\times k}$ be a matrix of probabilities, with Q_{rs} indicating the probability of an edge from any node i in cluster r to any node j in cluster s. With these notations, a k-component stochastic block model is given by

$$A_{ij} \sim \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}. \tag{3.1}$$

We use $\mathbb{E}_\theta/\mathbb{P}_\theta$ to denote an expectation/probability under the sampling mechanism (3.1).

The stochastic block model clearly imposes a parsimonious structure on the node probabilities $\theta = (\theta_{ij})$ when $k \ll n$, reducing the effective number of parameters from $O(n^2)$ to $O(k^2 + n)$. To describe the parameter space for θ, we need to introduce some notations. For $k \leq n$, let $\mathcal{Z}_{n,k} = \{z = (z_1, \ldots, z_n) : z_i \in \{1, \ldots, k\}, 1 \leq i \leq n\}$ denote all possible clusterings of n nodes into k clusters.

For any $1 \leq r \leq k$, $z^{-1}(r)$ is used as a shorthand for $\{1 \leq i \leq n : z_i = r\}$; the nodes belonging to cluster r. When z is clear from the context, we shall use $n_r = |z^{-1}(r)|$ to denote the number of nodes in cluster r; clearly $\sum_{r=1}^k n_r = n$. For the theoretical development in this paper, it is assumed that $n_r \geq 1$ for all $r = 1, \ldots, k$, that is, each cluster is assumed to be non-empty containing at least one observation.

With these notations, the parameter space Θ_k for θ is given by

$$\Theta_k = \{\theta \in [0,1]^{n\times n} : \theta_{ij} = Q_{z_i z_j},\ z \in \mathcal{Z}_{n,k},\ Q \in [0,1]^{k\times k}\}. \tag{3.2}$$

For any $z \in \mathcal{Z}_{n,k}$ and $Q \in [0,1]^{k\times k}$, we denote the corresponding $\theta \in \Theta_k$ by $\theta^{z,Q}$, so that $\theta_{ij}^{z,Q} = Q_{z_i z_j}$. In fact, $(z, Q) \mapsto \theta^{z,Q}$ is a surjective map from $\mathcal{Z}_{n,k} \times [0,1]^{k\times k} \to \Theta_k$, though it is clearly not injective.

Given $z \in \mathcal{Z}_{n,k}$, let $A_{[rs]}$ denote the $n_r \times n_s$ sub matrix of A consisting of entries A_{ij} with $z_i = r$ and $z_j = s$. The joint likelihood of A under model (3.1) can be expressed as

$$P(A \mid z, Q) = \prod_{r=1}^k \prod_{s=1}^k P(A_{[rs]} \mid z, Q), \quad P(A_{[rs]} \mid z, Q) = \prod_{i:z_i=r} \prod_{j:z_j=s} Q_{rs}^{A_{ij}} (1 - Q_{rs})^{1-A_{ij}}. \tag{3.3}$$

A Bayesian specification of the stochastic block model can be completed by assigning independent priors to z and Q, which in turn induces a prior on Θ_k via the mapping $(z, Q) \mapsto \theta^{z,Q}$. We generically use $p(z, Q) = p(z)p(Q)$ to denote the joint prior on z and Q. The induced prior on Θ_k will be denoted by $\Pi(\theta)$ and the corresponding posterior given data $A = (A_{ij})$ will

be denoted by $\Pi_n(\theta \mid A)$. The following fact is useful and heavily used in the sequel: for any $U \subset \Theta_k$,

$$\Pi(U) = \sum_{z \in \mathcal{Z}_{n,k}} \Pi(U \mid z)\, p(z) = \sum_{z \in \mathcal{Z}_{n,k}} p(Q : \theta^{z,Q} \in U)\, p(z), \tag{3.4}$$

where the second equality uses the independence of z and Q. Specific choices of $p(z)$ and $p(Q)$ are discussed below.

We assume independent $U(0,1)$ prior on the Q_{rs}'s. We consider a hierarchical prior on z where each node has probability π_r of being allocated to the rth cluster independently of the other nodes, and the vector of probabilities $\pi = (\pi_1, \ldots, \pi_k)$ follows a Dirichlet$(\alpha_1, \ldots, \alpha_k)$ prior. Here $\alpha_1, \ldots, \alpha_k$ are fixed hyper-parameters that do not depend on k or n; a default choice is $\alpha_r = 1/2$ for all $r = 1, \ldots, k$. We further assume the number of clusters k to be known. Model (3.1) along with the prior specified above can be expressed hierarchically as follows:

$$Q_{rs} \overset{\text{ind}}{\sim} U(0,1), \quad r, s = 1, \ldots, k, \tag{3.5}$$

$$P(z_i = r \mid \pi) = \pi_r, \quad r = 1, \ldots, k, i = 1, \ldots, n, \tag{3.6}$$

$$\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k), \tag{3.7}$$

$$A_{ij} \mid z, Q \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}. \tag{3.8}$$

A hierarchical specification as in (or very similar to) (3.5)–(3.8) has been commonly used in the literature; see for example, Snijders and Nowicki (1997), Nowicki and Snijders (2001), Golightly and Wilkinson (2005), & McDaid et al. (2013). Analytic marginalizations can be carried out due to the conjugate nature of the prior, facilitating posterior sampling (McDaid et al., 2013). In particular, using standard multinomial-Dirichlet conjugacy, the marginal prior of z can be written as

$$p(z) = \frac{\Gamma(\sum_{r=1}^k \alpha_r)}{\Gamma(n + \sum_{r=1}^k \alpha_r)} \prod_{r=1}^k \frac{\Gamma(n_r + \alpha_r)}{\Gamma(\alpha_r)}, \quad z \in \mathcal{Z}_{n,k}, \tag{3.9}$$

where we recall that $n_r = \sum_{i=1}^n \mathbb{1}(z_i = r)$. The following lemma provides an upper bound to the prior ratio $\{p(z)/p(z_0)\}$ which is used subsequently in the proof of our main theorem.

Lemma 3.1. *Assume $z_0 \in \mathcal{Z}_{n,k}$ with $n_{0r} = \sum_{i=1}^n \mathbb{1}(z_{0i} = r) \geq 1$ for all $r = 1, \ldots k$. Then, $\max_{z \in \mathcal{Z}_{n,k}} p(z)/p(z_0) \leq e^{Cn \log k}$, where C is a positive constant.*

PROOF. Fix $z \in \mathcal{Z}_{n,k}$. From Eq. 3.9, $p(z)/p(z_0) = \prod_{r=1}^{k} \Gamma(n_r + \alpha_r)/\Gamma(n_{0r} + \alpha_r)$. Then

$$\log\{p(z)/p(z_0)\} = \sum_{r=1}^{k} \log \Gamma(n_r + \alpha_r) - \sum_{r=1}^{k} \log \Gamma(n_{0r} + \alpha_r).$$

The first term is maximized over $z \in \mathcal{Z}_{n,k}$ when $n_r = n$ for some r and $n_s = 0$ for all $s \neq r$. Further, replacing n_{0r} by n/k for all $r = 1, \ldots, k$ only decreases the second term in the above display. Hence, letting $\alpha_{(k)} = \max\{\alpha_1, \ldots, \alpha_k\}$ and $\alpha_{(1)} = \min\{\alpha_1, \ldots, \alpha_k\}$,

$$\log\{p(z)/p(z_0)\} \leq \log \Gamma(n + \alpha_{(k)}) - k \log \Gamma(n/k + \alpha_{(1)}).$$

Using the standard two sided bound (Abramowitz and Stegun, 1964), we obtain $\log \Gamma(z) = \log(2\pi)/2 + (z - 1/2)\log(z) - z + R(z)$ with $0 < R(z) < (12z)^{-1}$ for $z > 0$, the dominating term in the right hand side of the above display being $n \log\{(n+\alpha_{(k)})/(n/k+\alpha_{(1)})\} \lesssim Cn \log k$, concluding the proof.

4 Posterior Contraction Rates in Stochastic Block Models

We are interested in contraction properties of the posterior $\Pi_n(\cdot \mid A)$ assuming the true data-generating parameter $\theta^0 \in \Theta_k$. To measure the discrepancy in the estimation of $\theta^0 \in \Theta_k$, the mean squared error has been used in the frequentist literature,

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\hat{\theta}_{ij} - \theta^0_{ij})^2 = \frac{1}{n^2} \left\| \hat{\theta} - \theta^0 \right\|^2, \tag{4.1}$$

where $\hat{\theta}$ is an estimator of θ^0. Chatterjee (2014) proposed estimating θ^0 using a low rank decomposition of the adjacency matrix A followed by a singular value decomposition to obtain a convergence rate of $\sqrt{k/n}$. More recently, Gao et al. (2015) considered a least squares type approach which can be related to maximum likelihood estimation where the Bernoulli likelihood is replaced by a Gaussian likelihood. They obtained a rate of $k^2/n^2 + \log k/n$, which they additionally showed to be the minimax rate over Θ_k, i.e.,

$$\inf_{\hat{\theta}} \sup_{\theta^0 \in \Theta_k} \mathbb{E}_{\theta_0} \frac{1}{n^2} \left\| \hat{\theta} - \theta^0 \right\|^2 \asymp \frac{k^2}{n^2} + \frac{\log k}{n}. \tag{4.2}$$

Interestingly, the minimax rate has two components, k^2/n^2 and $\log k/n$. Gao et al. (2015) refer to the k^2/n^2 term in the minimax rate as the *nonparametric*

rate, since it arises from the need to estimate k^2 unknown elements in Q from n^2 observations. The second part, $\log k/n$, is termed as the *clustering rate*, which appears since the clustering configuration z is unknown and needs to be estimated from the data. Observe that the clustering rate grows logarithmically in k. Parameterizing $k = n^\zeta$ with $\zeta \in [0, 1]$, the interplay between the two components becomes clearer (refer to equation 2.6 of Gao et al. 2015); in particular, the clustering rate dominates when k is small and the nonparametric rate dominates when k is large.

To evaluate Bayesian procedures from a frequentist standpoint, one seeks for the minimum possible sequence $\epsilon_n \to 0$ such that the posterior probability assigned to the complement of an ϵ_n-neighborhood (blown up by a constant factor) of θ^0 receives vanishingly small probabilities. The smallest such ϵ_n is called the *posterior contraction rate* (Ghosal et al., 2000). There is now a growing body of literature showing that Bayesian procedures achieve the frequentist minimax rate of posterior contraction (up to a logarithmic term) in models where the parameter dimension grows with the sample size; see Bontemps (2011), Castillo and van der Vaart (2012), Pati et al. (2014), Banerjee and Ghosal (2014), van der Pas et al. (2014), & Castillo et al. (2015) for some flavor of the recent literature.

We now state the main result of this article where we derive the contraction rate of the posterior arising from the hierarchical formulation (3.5)–(3.8).

Theorem 4.1. *Assume $A = (A_{ij})$ is generated from a k-component stochastic block model (3.1) with the true data-generating parameter $\theta^0 = (\theta^0_{ij}) \in \Theta_k$, where Θ_k is as in (3.2). Further assume that there exists a small constant $\delta \in (0, 1/2)$ such that $\theta^0_{ij} \in (\delta, 1-\delta)$ for all $i, j = 1, \ldots, n$. Suppose the hierarchical Bayesian model (3.5)–(3.8) is fitted. Then, with $\epsilon_n^2 = k^2\{\log n + \log(\delta^{-1})\}/n^2 + \log k/n$, and a sufficiently large constant $M > 0$,*

$$\mathbb{E}_{\theta^0}\Pi_n\left\{\frac{1}{n^2}\left\|\theta - \theta^0\right\|^2 > M^2\epsilon_n^2 \mid A\right\} \leq \exp\{-M^2n^2\epsilon_n^2\} + \frac{1}{Cn^2\epsilon_n^2}, \quad (4.3)$$

for some $C > 0$ and for all $n \geq 1$.

Remark 4.2. *Since $\theta^0 \in \Theta_k$, following the discussion after (3.2), there exists $z^0 \in \mathcal{Z}_{n,k}$ and $Q^0 \in [0, 1]^{k\times k}$ such that $\theta^0 = \theta^{z^0,Q^0}$. The condition of the theorem posits that all entries of Q^0 lie in $(\delta, 1-\delta)$. As long as $\delta \geq n^{-a}$ for any $a > 0$, the posterior contraction rate is the same (up to a constant) as in the case of fixed δ. The assumption $\theta^0 \in \Theta_k$ also implicitly implies that all*

the clusters have at least one observation, i.e., i.e., $n_r^0 = \sum_{i=1}^n \mathbb{1}(z_i^0 = r) \geq 1$ *for all* $r = 1, \ldots, k$*; otherwise there exists* $l < k$ *such that* $\theta^0 \in \Theta_l$*.*

A proof of Theorem 4.1 can be found towards the end of Section 4.2 after some important auxiliary results which are instrumental in deriving the main theoretical results of this paper. Theorem 4.1 shows that as long as $\delta \geq n^{-a}$ for any $a > 0$, the posterior contracts at a (near) minimax rate of $k^2 \log n/n^2 + \log k/n$. The nonparametric component of the rate is slightly hurt by a logarithmic term; appearance of such an additional logarithmic term is common in Bayesian nonparametrics.

It would be noteworthy that in Theorem 4.1 a uniform $\mathrm{U}(0,1)$ prior is assigned to the edge probabilities Q_{rs}'s, while in a similar independent work, van der Pas and van der Vaart (2018) considered a more general $\mathrm{Beta}(\beta_1, \beta_2)$ distribution for Q's that includes the uniform prior as a special case. While our main goal of inference is the recovery of the edge probabilities, van der Pas and van der Vaart (2018) focused on detection of the community memberships. A pertinent question that would be natural to ask in this context is whether our posterior contraction results can be extended further for a more general Beta prior. The following result, namely, Corollary 4.3 provides an affirmative answer to the aforesaid question. In particular, it says that, for recovery of the edge probabilities Q_{rs}', the posterior obtained from a more general $\mathrm{Beta}(\beta_1, \beta_2)$ prior contracts at the same rate as obtained for the uniform prior as in Theorem 4.1. As a matter of fact, it turns out that our general scheme of arguments for deriving the contraction rates works equally well even for this Beta prior.

Corollary 4.3. *Consider the set up of Theorem 4.1, where* $Q_{rs} \overset{ind}{\sim}$ *Beta* (β_1, β_2)*, for* $r,\ s = 1, \ldots, k$*, in* Eq. 3.5 *instead of an* $U(0,1)$ *prior. Then, with* $\epsilon_n^2 = k^2\{\log n + \log(\delta^{-1})\}/n^2 + \log k/n$*, and a sufficiently large constant* $M > 0$ *(depending on* (β_1, β_2)*)*

$$\mathbb{E}_{\theta^0} \Pi_n \left\{ \frac{1}{n^2} \left\| \theta - \theta^0 \right\|^2 > M^2 \epsilon_n^2 \mid A \right\} \leq \exp\{-M^2 n^2 \epsilon_n^2\} + \frac{1}{C n^2 \epsilon_n^2},$$

for some $C > 0$ *and for all* $n \geq 1$*.*

Proof of Corollary 4.3 above follows along exactly the same line of arguments as that of Theorem 4.1 and is given in the Appendix. An inspection of the proof of Theorem 4.1 reveals that the only technical difference between the proofs of the aforesaid results lies in a careful exploitation of a volume argument used in the proof of Theorem 4.1 under the more general $\mathrm{Beta}(\beta_1, \beta_2)$ prior for every possible choice of (β_1, β_2), while rest of the arguments remain unaltered.

4.1. Undirected Networks Theorem 4.1 can be extended to the case of undirected networks with or without self-loops. For technical simplification, we consider the case when there are no self-loops. We highlight the key differences in the data generation and the prior specification below. Let $z_i \in \{1, \ldots, k\}$ denote the cluster membership of the ith node and $Q = (Q_{rs}) \in [0,1]^{k \times k}$ be a *symmetric* matrix of probabilities, with $Q_{rs} = Q_{sr}$ indicating the probability of an edge between node i in cluster r and any node j in cluster s. Then an *undirected* version of Eq. 3.1 can be obtained by letting

$$A_{ij} \sim \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}, \quad 1 \le i < j \le n, \tag{4.4}$$

and $A_{ii} = \theta_{ii} = 0$ for $i = 1, \ldots, n$. The prior distributions are appropriately modified as:

$$Q_{rs} \overset{\text{ind}}{\sim} U(0,1), \quad 1 \le r \le s \le k \tag{4.5}$$

$$P(z_i = k \mid \pi) = \pi_k, \quad i = 1, \ldots, n, \tag{4.6}$$

$$\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k). \tag{4.7}$$

We modify the discrepancy measure as

$$\frac{1}{n^2} \sum \sum_{1 \le i < j \le n} (\hat{\theta}_{ij} - \theta^0_{ij})^2 \tag{4.8}$$

where θ, θ^0 are in the parameter space

$$\Theta^u_k = \{\theta \in [0,1]^{n \times n} : \theta_{ij} = Q_{z_i z_j}, 1 \le i \ne j \le n; \theta_{ii} = 0, 1 \le i \le n, z \in \mathcal{Z}_{n,k}, \\ Q \in [0,1]^{k \times k}, Q_{rs} = Q_{sr}, 1 \le r \le s \le k\}. \tag{4.9}$$

Then the following version of Theorem 4.1 is true for undirected networks:

Theorem 4.4. *Assume $A = (A_{ij})$ is generated as in* Eq. 4.4 *with $\theta^0 = (\theta^0_{ij}) \in \Theta^u_k$, where Θ^u_k is as in* Eq. 4.9. *Further assume that there exists a small constant $\delta \in (0, 1/2)$ such that $\theta^0_{ij} \in (\delta, 1-\delta)$ for all $1 \le i \le j \le n$. Suppose the hierarchical Bayesian model* (4.5)–(4.7) *is fitted. Then, with $\epsilon^2_n = k^2\{\log n + \log(\delta^{-1})\}/n^2 + \log k/n$, and a sufficiently large constant $M > 0$, the conclusion* (4.3) *is true.*

A sketch of the proof of Theorem 4.4 is given in the Appendix.

4.2. Geometry of Θ_k In this section, we derive a number of auxiliary results aimed at understanding the geometry of the parameter space Θ_k. These results are useful in proving our main concentration result presented in Theorem 4.1.

We first state a testing lemma which harnesses the ability of the likelihood to separate points in the parameter space.

Lemma 4.5. *Assume $\theta^0 \neq \theta^1 \in \Theta_k$ and let $E = \{\theta \in [0,1]^{n\times n} : \left\|\theta - \theta^1\right\| \leq \left\|\theta^1 - \theta^0\right\|/2\}$ be an Euclidean ball of radius $\|\theta^1 - \theta^0\|/2$ around θ^1 inside $[0,1]^{n\times n}$. Based on $A_{ij} \overset{ind}{\sim}$ Bernoulli(θ_{ij}) for $i,j = 1,\ldots,n$, consider testing $H_0 : \theta = \theta^0$ versus $H_1 : \theta \in E$. There exists a test function Φ such that*

$$\mathbb{E}_{\theta^0}(\Phi) \leq \exp\{-C_1 \left\|\theta^1 - \theta^0\right\|^2\}, \quad \sup_{\theta\in E} \mathbb{E}_\theta(1-\Phi) \leq \exp\{-C_2 \left\|\theta^1 - \theta^0\right\|^2\}, \tag{4.10}$$

for constants $C_1, C_2 > 0$ independent of n, θ^1 and θ^0.

Proof. Define the test function Φ as

$$\Phi = \mathbb{1}\left\{\sum_{i=1}^n\sum_{j=1}^n (\theta^1_{ij} - \theta^0_{ij})(A_{ij} - \theta^0_{ij}) > \left\|\theta^1 - \theta^0\right\|^2/4\right\},$$

where $\mathbb{1}(\cdot)$ denotes the indicator of a set. We show below that this test has the desired error rates (4.10).

We first bound the type-I error $\mathbb{E}_{\theta^0}(\Phi)$. Noting that under $\mathbb{P}_{\theta^0}$, $(A_{ij} - \theta^0_{ij})$ are independent zero mean random variables with $|A_{ij} - \theta^0_{ij}| < 1$, we use a version of Hoeffding's inequality (refer to Proposition 5.10 of Vershynin 2012) to conclude that,

$$\begin{aligned}
\mathbb{E}_{\theta^0}(\Phi) &= \mathbb{P}_{\theta^0}\left\{\sum_{i=1}^n\sum_{j=1}^n (\theta^1_{ij} - \theta^0_{ij})(A_{ij} - \theta^0_{ij}) > \left\|\theta^1 - \theta^0\right\|^2/4\right\} \\
&\leq \exp\left\{-C_1\frac{\left\|\theta^1 - \theta^0\right\|^4}{\left\|\theta^1 - \theta^0\right\|^2}\right\} = \exp\left\{-C_1\left\|\theta^1 - \theta^0\right\|^2\right\}
\end{aligned}$$

for a constant $C_1 > 0$ independent of n, θ^1 and θ^0.

We next bound the type-II error $\sup_{\theta\in E}\mathbb{E}_\theta(1-\Phi)$. Fix $\theta \in E$. We have,

$$\begin{aligned}
\mathbb{E}_\theta(1-\Phi) &= \mathbb{P}_\theta\left\{\sum_{i=1}^n\sum_{j=1}^n (\theta^1_{ij} - \theta^0_{ij})(A_{ij} - \theta^0_{ij}) < \left\|\theta^1 - \theta^0\right\|^2/4\right\} \\
&= \mathbb{P}_\theta\left\{\sum_{i=1}^n\sum_{j=1}^n (\theta^1_{ij} - \theta^0_{ij})(A_{ij} - \theta_{ij}) < \left\|\theta^1 - \theta^0\right\|^2/4 - \left\langle\theta^1 - \theta^0, \theta - \theta^0\right\rangle\right\}, \qquad (4.11)
\end{aligned}$$

where we abbreviate $\langle \theta', \theta'' \rangle = \sum_{i=1}^n \sum_{j=1}^n \theta'_{ij}\theta''_{ij}$. Bound

$$\begin{aligned}
&\left\langle \theta^1 - \theta^0, \theta - \theta^0 \right\rangle \\
&= \left\langle \theta^1 - \theta^0, \theta^1 - \theta^0 \right\rangle - \left\langle \theta^1 - \theta^0, \theta^1 - \theta \right\rangle \\
&\geq \left\|\theta^1 - \theta^0\right\|^2 - \left\|\theta^1 - \theta^0\right\|^2/2 = \left\|\theta^1 - \theta^0\right\|^2/2,
\end{aligned}$$

where the penultimate step used the Cauchy–Schwarz inequality along with the fact that $\left\|\theta - \theta^1\right\| \leq \left\|\theta^1 - \theta^0\right\|/2$. Substituting in Eq. 4.11 and noting that under $\mathbb{P}_\theta$, $(A_{ij} - \theta_{ij})$ are independent zero mean bounded random variables, another application of Hoeffding's inequality yields

$$\begin{aligned}
\mathbb{E}_\theta(1 - \Phi) &\leq \mathbb{P}_\theta \left\{ \sum_{i=1}^n \sum_{j=1}^n (\theta^1_{ij} - \theta^0_{ij})(A_{ij} - \theta_{ij}) < -\left\|\theta^1 - \theta^0\right\|^2/4 \right\} \\
&\leq \exp\left\{ -C_2 \frac{\left\|\theta^1 - \theta^0\right\|^4}{\left\|\theta^1 - \theta^0\right\|^2} \right\} = \exp\left\{ -C_2 \left\|\theta^1 - \theta^0\right\|^2 \right\}
\end{aligned}$$

for some constant $C_2 > 0$ independent of n and θ. Taking a supremum over $\theta \in E$ yields the desired result.

Our next result is concerned with the structure of a specific type of Euclidean balls inside Θ_k. Recall that $\theta^{z,Q}$ denotes the element of Θ_k with $\theta^{z,Q}_{ij} = Q_{z_i z_j}$. For $z \in \mathcal{Z}_{n,k}$, let

$$\Theta_k(z) = \left\{ \theta^{z,Q} : Q \in [0,1]^{k \times k} \right\} \tag{4.12}$$

denote a slice of Θ_k along z. In other words, given z, $\Theta_k(z)$ is the image of the map $Q \mapsto \theta^{z,Q}$ in Θ_k. Suppose $\theta^* = \theta^{z^*,Q^*} \in \Theta_k$, and consider a ball $B(z)$ in $\Theta_k(z)$ centered at θ^* of the form $B(z) = \{\theta \in \Theta_k(z) : \|\theta - \theta^*\| < t\}$ for some $t > 0$. If $z^* = z$, then it is straightforward to observe that

$$\left\|\theta^{z,Q} - \theta^{z^*,Q^*}\right\|^2 = \sum_{r=1}^k \sum_{s=1}^k n_r n_s (Q_{rs} - Q^*_{rs})^2, \tag{4.13}$$

where we recall that $n_r = \sum_{i=1}^n \mathbb{1}(z_i = r)$ for $r = 1, \ldots, k$. Therefore, although a subset of $[0,1]^{n \times n}$, $B(z)$ can be identified with a k^2-dimensional ellipsoid in $[0,1]^{k \times k}$. When $z^* \neq z$, one no longer has a nice identity as above and the geometry of $B(z)$ is more difficult to describe. However, we show below in Lemma 4.6 that $B(z)$ is always contained inside a set $\widetilde{B}(z)$ in $\Theta_k(z)$ which can be identified with a k^2-dimensional ellipsoid in $[0,1]^{k \times k}$. Recall our convention for describing ellipsoids from Eq. 2.1.

Lemma 4.6. *Fix $z^* \in \mathcal{Z}_{n,k}, Q^* \in [0,1]^{k\times k}$, and let $\theta^* = \theta^{z^*,Q^*}$. For $z \in \mathcal{Z}_{n,k}$ and $t > 0$, let $B(z) = \{\theta \in \Theta_k(z) : \|\theta - \theta^*\| < t\}$. Set $W_{rs} = n_r n_s/t^2$ and $W = (W_{rs})$, where $n_r = \sum_{i=1}^n \mathbb{1}(z_i = r)$ for $r = 1, \ldots, k$. Then, $B(z) \subseteq \widetilde{B}(z)$, where*

$$\widetilde{B}(z) = \left\{\theta^{z,Q} : Q \in \xi_{k^2}(\bar{Q}^*, W) \cap [0,1]^{k\times k}\right\} \tag{4.14}$$

for some $\bar{Q}^* \in [0,1]^{k\times k}$ depending on Q^*, z^* and z. In particular, if $z^* = z$, then $\bar{Q}^* = Q^*$ and the containment becomes equality, i.e., $B(z) = \widetilde{B}(z)$.

Proof. Define $\bar{\theta} = \arg\min_{\theta\in\Theta_k(z)} \|\theta - \theta^*\|^2$. According to the definition of $\bar{\theta}$, we have from the Pythagorean identity

$$\|\theta - \theta^*\|^2 = \left\|\theta - \bar{\theta}\right\|^2 + \left\|\bar{\theta} - \theta^*\right\|^2.$$

for $\theta \in \Theta_k(z)$. Therefore, $\left\|\theta - \bar{\theta}\right\| \le \|\theta - \theta^*\|$, which implies $\{\theta \in \Theta_k(z) : \|\theta - \theta^*\| \le t\} \subset \{\theta \in \Theta_k(z) : \left\|\theta - \bar{\theta}\right\| \le t\}$. Since $\bar{\theta} \in \Theta_k(z)$, there exists $\bar{Q}^* \in [0,1]^{k\times k}$ such that $\bar{\theta}_{ij} = \bar{Q}^*_{z_i z_j}$. This completes the proof of the first part. When $z = z^*$, the proof of the second part is completed by noting that

$$\|\theta - \theta^*\|^2 = \sum_{r=1}^k \sum_{s=1}^k n_r n_s (Q_{rs} - Q^*_{rs})^2.$$

Remark 4.7. *From* Eq. 2.1, *$\xi_{k^2}(\bar{Q}^*, W)$ in Lemma 4.6 is the collection of all Q satisfying $\sum_{r=1}^k \sum_{s=1}^k n_r n_s (Q_{rs} - \bar{Q}^*_{rs})^2 < t^2$. The last part of Lemma 4.6 is consistent with the discussion preceding* (4.13)*. When $z^* = z$,* Eq. 4.13 *implies that $B(z)$ consists of all $\theta^{z,Q}$ with $Q \in [0,1]^{k\times k}$ satisfying $\sum_{r=1}^k \sum_{s=1}^k n_r n_s (Q_{rs} - Q^*_{rs})^2 < t^2$.*

Corollary 4.8. *Inspecting the proof of Lemma 4.6, the condition $Q \in [0,1]^{k\times k}$ is only used to show that $\bar{Q}^* \in [0,1]^{k\times k}$. If we let Q to be unrestricted, then the containment relation continues to hold as subsets of $\mathbb{R}^{k\times k}$, i.e.,*

$$\left\{\theta^{z,Q} : Q \in \mathbb{R}^{k\times k}, \left\|\theta^{z,Q} - \theta^{z^*,Q^*}\right\| < t\right\} \subseteq \left\{\theta^{z,Q} : Q \in \xi_{k^2}(\bar{Q}^*, W)\right\}, \tag{4.15}$$

with equality when $z^* = z$.

Lemma 4.6 crucially exploits the lower dimensional structure underlying the parameter space Θ_k and is used subsequently multiple times. First, recall from Eq. 3.4 that one needs a handle on $p(Q : \theta^{z,Q} \in U)$ to bound

the prior probability of $U \subset \Theta_k$. In particular, if $U = \{\|\theta - \theta^0\| < t\}$, then $p(Q : \theta^{z,Q} \in U)$ equals the volume of $U \cap \Theta_k(z)$, which can be suitably bounded by the volume of the bounding k^2 dimensional ellipsoid. Second, a handle on the size of balls in Θ_k facilitates calculating the complexity of the model space (in terms of metric entropy) which is pivotal in proving the posterior contraction; in particular, to extend the test function in Lemma 4.5 to construct test functions against more complex alternatives in Lemma 4.9 below. Once again, the dimensionality reduction is key to preventing the metric entropy from growing too fast.

Lemma 4.9. *Recall ϵ_n from Theorem 4.1. Assume $\theta^0 \in \Theta_k$ and for $l \geq 1$, let $U_{l,n} = \left\{\theta \in \Theta_k : ln\epsilon_n \leq \|\theta - \theta^0\| < (l+1)n\epsilon_n\right\}$. Based on $A_{ij} \overset{ind}{\sim}$ Bernoulli(θ_{ij}) for $i, j = 1, \ldots, n$, consider testing $H_0 : \theta = \theta^0$ versus $H_1 : \theta \in U_{l,n}$. There exists a test function $\Phi_{l,n}$ such that*

$$\mathbb{E}_{\theta^0}(\Phi_{l,n}) \leq \exp(-C_1 l^2 n^2 \epsilon_n^2), \qquad \sup_{\theta \in U_{l,n}} \mathbb{E}_\theta(1 - \Phi_{l,n}) \leq \exp(-C_2 l^2 n^2 \epsilon_n^2), \quad (4.16)$$

for constants $C_1, C_2 > 0$ independent of n.

PROOF. Since $\theta^0 \in \Theta_k$, there exists $z^0 \in \mathcal{Z}_{n,k}$ and $Q^0 \in [0,1]^{k \times k}$ with $\theta^0 = \theta^{z^0, Q^0}$. For $z \in \mathcal{Z}_{n,k}$, define $U_{l,n}(z) = U_{l,n} \cap \Theta_k(z)$, where $\Theta_k(z)$ is as in Eq. 4.12. Clearly,

$$U_{l,n}(z) = \left\{\theta^{z,Q} : Q \in [0,1]^{k \times k},\ ln\epsilon_n \leq \left\|\theta^{z,Q} - \theta^{z^0,Q^0}\right\| < (l+1)n\epsilon_n\right\}, \quad (4.17)$$

and $U_{l,n} \subset \cup_{z \in \mathcal{Z}_{n,k}} U_{l,n}(z)$. We first use Lemma 4.5 to construct tests against $U_{l,n}(z)$ for fixed z. Our desired test is obtained by taking the maximum of all such test functions.

Fix $z \in \mathcal{Z}_{n,k}$. Let $\mathcal{N}_{l,n}(z) = \{\theta_{l,n,h} \in U_{l,n}(z) : h \in I_{l,n}(z)\}$ be a *maximal* $ln\epsilon_n/2$-separated set inside $U_{l,n}(z)$ for some index set $I_{l,n}(z)$; i.e., $\mathcal{N}_{l,n}(z)$ is such that $\|\theta^1 - \theta^2\| \geq ln\epsilon_n/2$ for all $\theta^1 \neq \theta^2 \in \mathcal{N}_{l,n}(z)$, and no subset of $U_{l,n}(z)$ containing $\mathcal{N}_{l,n}(z)$ has this property. We provide a volume argument to determine an upper bound for $|I_{l,n}(z)|$, the cardinality of $\mathcal{N}_{l,n}(z)$. The separation property implies that Euclidean balls of radius $ln\epsilon_n/4$ centered at the points in $\mathcal{N}_{l,n}(z)$ are disjoint. Since $B_h^+ := \left\{\theta^{z,Q} : Q \in \mathbb{R}^{k \times k}, |\theta^{z,Q} - \theta_{l,n,h}| < ln\epsilon_n/4\right\}$ is contained inside an Euclidean ball of radius $ln\epsilon_n/4$ centered at $\theta_{l,n,h}$, the sets B_h^+ are disjoint as h varies over $I_{l,n}(z)$. By the triangle inequality, all B_h^+s lie inside $B^+ = \left\{\theta^{z,Q} : Q \in \mathbb{R}^{k \times k}, \|\theta^{z,Q} - \theta^0\| \leq (5l/4 + 1)n\epsilon_n\right\}$, since $\|\theta^{z,Q} - \theta^0\| \leq \|\theta^{z,Q} - \theta_{l,n,h}\| + \|\theta_{l,n,h} - \theta^0\| \leq (l+1)n\epsilon_n + ln\epsilon_n/4$.

It should be noted that the sets B_h^+s and B^+ are constructed in a way that Q is not restricted to be inside $[0,1]^{k\times k}$. This allows us to invoke Corollary 4.8 to identify B_h^+ and B^+ with appropriate ellipsoids in $\mathbb{R}^{k^2}$ and simplify volume calculations. First, since $\theta_{l,n,h} \in \Theta_k(z)$ for each h, it follows from (the equality part of) Corollary 4.8 that $B_h^+ = \{\theta^{z,Q} : Q \in \xi_{k^2}(\bar{Q}_h, \widetilde{W})\}$ with $\bar{Q}_h$ constructed as in the proof of Lemma 4.6 and $\widetilde{W}_{rs} = n_r n_s/\{(ln\epsilon_n)^2\}$. The equality is crucially used below; also note that $\widetilde{W}$ does not depend on h. Invoking Corollary 4.8 one more time, we obtain $B^+ \subset \{\theta^{z,Q} : Q \in \xi_{k^2}(\bar{Q}^0, W)\}$, with $W_{rs} = n_r n_s/[\{(5l/4+1)n\epsilon_n\}^2]$. We conclude that the Euclidean ellipsoids $\xi_{k^2}(\bar{Q}_h, \widetilde{W})$ are disjoint as h varies over $I_{l,n}(z)$ and all of them are contained in $\xi_{k^2}(\bar{Q}^0, W)$. Comparing volumes,

$$|\xi_{k^2}(\bar{Q}_h, \widetilde{W})||I_{l,n}(z)| \le |\xi_{k^2}(\bar{Q}^0, W)|.$$

Using the volume formula in Eq. 2.2 and canceling out common terms, we finally have

$$|I_{l,n}(z)| \le \left\{\frac{(5l/4+1)}{l/2}\right\}^{k^2} \le 9^{k^2}. \tag{4.18}$$

We are now in a position to construct the test. The maximality of $\mathcal{N}_{l,n}(z)$ implies that $\mathcal{N}_{l,n}(z)$ is an $ln\epsilon_n/2$-net of $U_{l,n}(z)$, i.e., the sets $E_{l,n,z,h} = \{\theta \in [0,1]^{n\times n} : \|\theta - \theta_{l,n,h}\| < ln\epsilon_n/2\}$ cover $U_{l,n}(z)$ as h varies. For each $\theta_{l,n,h} \in \mathcal{N}_{l,n}(z)$, consider testing $H_0 : \theta = \theta^0$ versus $H_1 : \theta \in E_{l,n,z,h}$ using the test function from Lemma 4.5. Lemma 4.5 is applicable since $\|\theta^0 - \theta_{l,n,h}\| \ge ln\epsilon_n$; let $\Phi_{l,n,z,h}$ denote the corresponding test with type-I and II errors bounded above by $e^{-Cl^2n^2\epsilon_n^2}$. Define $\Phi_{l,n} = \max_{z\in\mathcal{Z}_{n,k}} \max_{h\in I_{l,n}(z)} \Phi_{l,n,z,h}$. For any $\theta \in U_{l,n}$, there exists $z \in \mathcal{Z}_{n,k}$ and $h \in I_{l,n}(z)$ such that $\theta \in E_{l,n,z,h}$, so that $\mathbb{E}_\theta(1-\Phi_{l,n}) \le \mathbb{E}_\theta(1-\Phi_{l,n,z,h}) \le e^{-Cl^2n^2\epsilon_n^2}$. Taking supremum over $\theta \in U_{l,n}$ delivers the desired type-II error. Further, the type-I error of $\Phi_{l,n}$ can be bounded as

$$\mathbb{E}_{\theta^0}(\Phi_{l,n}) \le \sum_{z\in\mathcal{Z}_{n,k}} \sum_{h\in I_{l,n}(z)} \mathbb{E}_{\theta^0}(\Phi_{l,n,z,h}) \le k^n 9^{k^2} e^{-Cl^2n^2\epsilon_n^2}, \tag{4.19}$$

since $|\mathcal{Z}_{n,k}| = k^n$ and by Eq. 4.18, $|I_{l,n}(z)| \le 9^{k^2}$ for all z. The conclusion then follows since $n^2\epsilon_n^2 = k^2\{\log n + \log(\delta^{-1})\} + n\log k \gtrsim k^2 + n\log k$.

As commented earlier, below we present the proof of Theorem 4.1 already stated in Section 4 of this paper.

Proof of Theorem 4.1. Let $\mathbb{E}_0/\mathbb{P}_0$ denote an abbreviation of $\mathbb{E}_{\theta^0}/\mathbb{P}_{\theta^0}$. Since $\theta^0 \in \Theta_k$, there exists some $z^0 \in \mathcal{Z}_{n,k}$ and $Q^0 \in [0,1]^{k\times k}$ with

$\theta^0 = \theta^{z^0,Q^0}$. Recall $\epsilon_n^2 = k^2\{\log n + \log(\delta^{-1})\}/n^2 + \log k/n$ and define $U_n = \left\{\theta \in \Theta_k : \|\theta - \theta^0\|^2 > M^2 n^2 \epsilon_n^2\right\}$ for some large constant $M > 0$ to be chosen later. Letting $f_{\theta_{ij}}(A_{ij}) = \theta_{ij}^{A_{ij}}(1-\theta_{ij})^{1-A_{ij}}$ denote the Bernoulli(θ_{ij}) likelihood evaluated at A_{ij}, the posterior probability assigned to U_n can be written as

$$\Pi_n(U_n \mid A) = \frac{\int_{U_n} \prod_{i=1}^n \prod_{j=1}^n \frac{f_{\theta_{ij}}(A_{ij})}{f_{\theta_{ij}^0}(A_{ij})}\, p(dz, dQ)}{\int_{\Theta_k} \prod_{i=1}^n \prod_{j=1}^n \frac{f_{\theta_{ij}}(A_{ij})}{f_{\theta_{ij}^0}(A_{ij})}\, p(dz, dQ)} = \frac{\mathcal{N}_n}{\mathcal{D}_n}, \tag{4.20}$$

where $\mathcal{N}_n$ and $\mathcal{D}_n$ respectively denote the numerator and denominator of the fraction in Eq. 4.20. Let $\mathcal{F}_n$ denote the σ-field generated by $\tilde{A} = (\tilde{A}_{ij})$, with $\tilde{A}_{ij}$ independently distributed as Bernoulli(θ_{ij}^0); the true data generating distribution. We first claim that there exists a set $\mathcal{A}_n \in \mathcal{F}_n$ where we can bound $\mathcal{D}_n$ from below with large probability under $\mathbb{P}_0$ in Lemma 4.10. The proof is adapted from Lemma 10 of Ghosal and van der Vaart (2007).

Lemma 4.10. *Assume θ^0 satisfies the conditions of Theorem 4.1. Then, there exists a set $\mathcal{A}_n$ in the σ-field $\mathcal{F}_n$ with $\mathbb{P}_0(\mathcal{A}_n) \geq 1 - C/(n^2\epsilon_n^2)$ for some $C > 0$, such that within $\mathcal{A}_n$,*

$$\mathcal{D}_n \geq e^{-Cn^2\epsilon_n^2}\Pi\left(\left\|\theta - \theta^0\right\|^2 < n^2\delta^2\epsilon_n^2\right).$$

PROOF. Let $f_{\theta_{ij}}(A_{ij})$ denote the likelihood for Bernoulli(θ_{ij}) evaluated at A_{ij}. Letting $B_{l,n} = \{\theta \in \Theta_k : l^2\epsilon_n^2 \leq (1/n^2)\|\theta - \theta^0\|^2 \leq (l+1)^2\epsilon_n^2\}$. Define

$$B_n(\theta^0;\epsilon_n) := \left\{\theta : \sum_{1\leq i,j\leq n} E_{\theta_{ij}^0} \log \frac{f_{\theta_{ij}^0}(A_{ij})}{f_{\theta_{ij}}(A_{ij})} \leq n^2\epsilon_n^2, \sum_{1\leq i,j\leq n} E_{\theta_{ij}^0}\left\{\log \frac{f_{\theta_{ij}^0}(A_{ij})}{f_{\theta_{ij}}(A_{ij})}\right\}^2 \leq n^2\epsilon_n^2\right\}$$

and $\mathcal{A}_n = \{A : \int \prod_{1\leq i,j\leq n} f_{\theta_{ij}}(A_{ij})/f_{\theta_{ij}^0}(A_{ij})p(dz, dQ) \geq e^{-n^2\epsilon_n^2}\Pi(B_n(\theta^0;\epsilon_n)\}$. The following fact is a straightforward modification of Lemma 5 of Ghosal and Roy (2006). Let $0 < \delta < 1/2$, $\delta < \alpha, \beta < 1 - \delta$. Then there exists a constant L such that

$$\alpha\left(\log\frac{\alpha}{\beta}\right)^m + (1-\alpha)\left(\log\frac{1-\alpha}{1-\beta}\right)^m \leq \frac{L(\alpha-\beta)^2}{\delta^2}, \quad m = 1, 2.$$

Since $\delta < \theta_{ij}^0 \leq 1 - \delta$ for $1 \leq i, j \leq n$, it follows from the above fact that $B_n(\theta^0;\epsilon_n) \supset \{\theta : \left\|\theta - \theta^0\right\|^2 \leq n^2\delta^2\epsilon_n^2\}$. It now follows from Lemma 10 of

Ghosal and van der Vaart (2007) that $\mathbb{P}_0(\mathcal{A}_n) \geq 1 - C/(n^2\epsilon_n^2)$ for some $C > 0$.

In view of Lemma 4.10, it is sufficient to provide an upper bound to

$$\lim_{n\to\infty} \mathbb{E}_0 \left\{ \Pi_n(U_n \mid A)\mathbb{1}_{\mathcal{A}_n} \right\}.$$

For $l \geq M$, let $U_{l,n} = \left\{ \theta \in \Theta_k : l^2 n^2 \epsilon_n^2 \leq \left\|\theta - \theta^0\right\|^2 < (l+1)^2 n^2 \epsilon_n^2 \right\}$ denote an annulus in Θ_k centered at θ^0 with inner and outer Euclidean radii $ln\epsilon_n$ and $(l+1)n\epsilon_n$ respectively. Using a standard testing argument (see, for example, the proof of Proposition 5.1 in Castillo and van der Vaart (2012) 2012) in conjunction with Lemma 4.10, one arrives at

$$\mathbb{E}_0 \left\{ \Pi_n(U_n \mid A)\mathbb{1}_{\mathcal{A}_n^{\rfloor}} \right\} \leq \sum_{l=M}^{\infty} \left\{ \mathbb{E}_0(\Phi_{l,n}) + \frac{\Pi(U_{l,n}) \sup_{\theta \in U_{l,n}} \mathbb{E}_\theta(1 - \Phi_{l,n})}{e^{-Cn^2\epsilon_n^2} \Pi(\|\theta - \theta^0\|^2 < n^2\delta^2\epsilon_n^2)} \right\} \tag{4.21}$$

where $\Phi_{l,n}$ is the test function constructed in Lemma 4.9 for testing $H_0 : \theta = \theta^0$ versus $H_1 : \theta \in U_{l,n}$ with error rates as in Eq. 4.16. Recall $U_{l,n}(z) = U_{l,n} \cap \Theta_k(z)$ and its equivalent representation in Eq. 4.17 from the proof of Lemma 4.9. Since $U_{l,n} \subseteq \cup_{z \in \mathcal{Z}_{n,k}} U_{l,n}(z)$, from Eq. 3.4,

$$\Pi(U_{l,n}) \leq \sum_{z \in \mathcal{Z}_{n,k}} \Pi \left\{ U_{l,n}(z) \right\} \leq |\mathcal{Z}_{n,k}| \max_{z \in \mathcal{Z}_{n,k}} p(z), \tag{4.22}$$

where $p(z)$ is the prior probability (3.9) of z under the Dirichlet-multinomial prior.

Next, consider the term $\Pi(\left\|\theta - \theta^0\right\|^2 < n^2\delta^2\epsilon_n^2)$ in the denominator of the expression for $\beta_{l,n}$. Bound $\Pi(\left\|\theta - \theta^0\right\|^2 < n^2\delta^2\epsilon_n^2) \geq \Pi(\left\|\theta - \theta^0\right\|^2 < n^2\delta^2\epsilon_n^2 \mid z = z^0)p(z^0)$ and using Lemma 4.6 once again,

$$\Pi\left(\left\|\theta - \theta^0\right\|^2 < n^2\delta^2\epsilon_n^2 \mid z = z^0\right) = p\left\{ Q : \sum_{r=1}^{k}\sum_{s=1}^{k} n_{0r}n_{0s}(Q_{rs} - Q_{rs}^0)^2 < n^2\delta^2\epsilon_n^2 \right\}. \tag{4.23}$$

The probability in the right hand side of the above display is the volume of the *intersection* of an ellipsoid with $[0,1]^{k\times k}$, and therefore we cannot simply replace the probability by the volume of the ellipsoid. Instead, we embed an appropriate rectangle inside the intersection of the ellipsoid and $[0,1]^{k\times k}$. We claim that

$$\prod_{r=1}^{k}\prod_{s=1}^{k}[Q_{rs}^0 - \delta\epsilon_n/2, Q_{rs}^0 + \delta\epsilon_n/2] \subset \left\{ Q \in [0,1]^{k\times k} : \sum_{r=1}^{k}\sum_{s=1}^{k} n_{0r}n_{0s}(Q_{rs} - Q_{rs}^0)^2 < n^2\delta^2\epsilon_n^2 \right\}. \tag{4.24}$$

First, based on our assumption that all entries of Q^0 are bounded away from 0 and 1 and the fact that $\epsilon_n \to 0$, it is immediate that the rectangle

is contained in $[0,1]^{k\times k}$ for sufficiently large n. Second, for any Q with $|Q_{rs} - Q^0_{rs}| \le \delta\epsilon_n/2$ for all $1 \le r, s \le k$, we have

$$\sum_{r=1}^{k}\sum_{s=1}^{k} n_{0r}n_{0s}(Q_{rs} - Q^0_{rs})^2 \le \frac{\delta^2\epsilon_n^2}{4}\sum_{r=1}^{k}\sum_{s=1}^{k} n_{0r}n_{0s} = \frac{n^2\delta^2\epsilon_n^2}{4},$$

thereby proving the claim in Eq. 4.24. Now we can bound $\Pi(\|\theta - \theta^0\|^2 < n^2\delta^2\epsilon_n^2 \mid z = z^0)$ from below by the volume of the rectangle, which equals $(\epsilon_n\delta)^{k^2}$. Since $n_{0r} \ge 1$ for all $r = 1, \ldots, k$, invoke Lemma 3.1 to bound $\max_{z\in\mathcal{Z}_{n,k}} p(z)/p(z_0) \le e^{Cn\log k}$. Combining this with error rates (4.16) in (4.21) we obtain,

$$\mathbb{E}_0\left\{\Pi_n(U_n \mid A)\mathbb{1}_{\mathcal{A}^c_l}\right\} \le \sum_{l=M}^{\infty}\left\{e^{-C_1^2 l^2 n^2\epsilon_n^2} + \frac{e^{-C_2^2 l^2 n^2\epsilon_n^2}e^{C_3 n\log k}}{(\delta\epsilon_n)^{k^2}}\right\}. \tag{4.25}$$

For $n^2\epsilon_n^2 = k^2\{\log n + \log(\delta^{-1})\} + n\log k$, the right hand size of Eq. 4.25 converges to zero for all M larger than a suitable constant.

5 Simulation Studies

In this section, we consider a small-scale simulation study to examine the accuracy in estimating θ as the number of nodes n in the network increases. We simulate 100 replicates of an SBM network using $k = 3, 4, 5$ equi-sized communities, with $n = 120$, 150, and 200. The off-diagonal entries of Q are set to 0.1 and all the diagonal entries are set to a constant $\rho > 0.1$. The smaller the value of ρ is, the more vague the block structure is in the network.

For each n, we consider $\rho = 0.3, 0.5, 0.8$. The true community assignment z_0 is set to $\{(1)_{n/3}, (2)_{n/3}, (3)_{n/3}\}$, where $(x)_k$ denotes the vector obtained by replicated x, k times. We consider (i) an SBM with $k = 3$, (ii) an SBM with $k = 4$ and (iii) an SBM with $k = 5$; note the number of communities is mis-specified in (ii). The following Gibbs sampler is employed to sample from the posterior distribution of the parameters.

5.1. Gibbs sampling for fixed k (directed networks) Define

$$\begin{aligned} n_r &= \sum_{i=1}^{n} I(z_i = r), \quad r = 1, \ldots, k. \\ n_{rs} &= \sum\sum_{1\le i\ne j\le n} I(z_i = r, z_j = s) = n_r n_s - n_r I(r = s). \\ A[rs] &= \sum\sum_{(i,j):z_i=r, z_j=s} A_{ij}, \quad r = 1, \ldots, k, s = 1, \ldots, k. \end{aligned}$$

Then the full-conditional distributions of π and Q can be obtained as

$$\begin{aligned} \pi \mid - &\sim \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_k + n_k) \\ Q_{rs} \mid - &\sim \text{Beta}(1 + A[rs], 1 + n_{rs} - A[rs]). \end{aligned}$$

Observe that

$$P(z_i = l \mid z_{-i}, A, \pi, Q) \propto P(A \mid z, \pi, Q) P(z \mid \pi) P(\pi) P(Q).$$

Keeping the terms involving z_i,

$$P(A \mid z, \pi, Q) \propto \left\{ \prod_{j \neq i} Q_{z_i z_j}^{A_{ij}} (1 - Q_{z_i z_j})^{1 - A_{ij}} \right\} \times \left\{ \prod_{k \neq i} Q_{z_k z_i}^{A_{ki}} (1 - Q_{z_k z_i})^{1 - A_{ki}} \right\}, \quad P(z \mid \pi) \propto \pi_{z_i}.$$

Hence,

$$P(z_i = l \mid z_{-i}, A, \pi, Q) \propto \pi_{z_i} \times \left\{ \prod_{j \neq i} Q_{z_i z_j}^{A_{ij}} (1 - Q_{z_i z_j})^{1 - A_{ij}} \right\} \times \left\{ \prod_{k \neq i} Q_{z_k z_i}^{A_{ki}} (1 - Q_{z_k z_i})^{1 - A_{ki}} \right\}.$$

The Gibbs sampler proceeds by cycling through $\pi \mid -, Q_{rs} \mid -$ and $z_i \mid z_{-i}, A, \pi, Q$. We set $\alpha_j = 1, j = 1, \ldots, k$ and ran the MCMC for 3000 iterations with a burn-in of 1000. The posterior mean $\hat{\theta}$ of θ and the posterior mode of z post burn-in are obtained as the Bayes estimates. As a measure of discrepancy between θ and θ_0, we compute mean squared error (MSE): $(1/n^2)\|\theta - \theta_0\|^2$ and for z and z_0, we obtained the Rand-index (RI) where $\text{RI} = \#\,\text{mismatched pairs}/\binom{n}{2}$. The results are summarized in Tables 1 & 2. It is evident that for a fixed ρ, MSE decreases and RI increases as n increases. On the other hand, for a fixed n, as ρ increases, the clustering pattern is more evident in the network leading to improved accuracy in estimating z_0.

Interestingly, the results for known k are very similar with that of the misspecified k indicating the robustness of the Bayesian approach. It is possible that a phenomenon similar to overfitted Gaussian mixtures (Rousseau and Mengersen, 2011) is at work here.

6 Discussion

In this article, we presented a theoretical investigation of posterior contraction in stochastic block models. One crucial assumption in our current results is that the true number of clusters k is known. Geng et al. (2018) studied inference in a stochastic block model with an unknown number of clusters within a Bayesian non-parametric framework. Their objective was two-fold : (i) simultaneous estimation of the number of clusters and the cluster structure and (ii) consistent cluster detection. Towards that, they

Table 1: MSE ($\times 10^2$) and standard error ($\times 10^3$) comparison over 100 replicates

n	120			150			200		
ρ	0.3	0.5	0.8	0.3	0.5	0.8	0.3	0.5	0.8
$\mathrm{SBM}_{\mathbf{k=3}}$	0.24	0.26	0.76	0.22	0.45	0.71	0.11	0.19	0.91
	(2.01)	(5.55)	(18.82)	(2.18)	(7.29)	(17.69)	(1.89)	(5.31)	(19.42)
$\mathrm{SBM}_{\mathbf{k=4}}$	0.44	1.08	3.32	0.39	1.03	3.16	0.37	1.08	3.06
	(1.01)	(2.49)	(8.44)	(1.16)	(2.04)	(6.27)	(1.23)	(2.74)	(1.30)
$\mathrm{SBM}_{\mathbf{k=5}}$	0.50	1.27	3.85	0.46	1.32	3.96	0.43	1.30	3.94
	(0.72)	(0.76)	(0.50)	(0.72)	(1.63)	(4.06)	(0.72)	(1.24)	(2.95)

Table 2: Rand Index and standard error comparison ($\times 10^3$) over 100 replicates

n	120			150			200		
ρ	0.3	0.5	0.8	0.3	0.5	0.8	0.3	0.5	0.8
$\mathrm{SBM}_{\mathbf{k=3}}$	0.88	0.96	0.97	0.88	0.94	0.97	3	0.98	0.96
	(110.64)	(86.97)	(78.55)	(116.86)	(101.46)	(78.41)	(946.47)	(67.34)	(86.23)
$\mathrm{SBM}_{\mathbf{k=4}}$	0.71	0.86	0.86	0.74	0.87	0.87	0.77	0.86	0.87
	(113.90)	(48.59)	(41.38)	(118.92)	(42.89)	(30.61)	(11.76)	(38.07)	(17.77)
$\mathrm{SBM}_{\mathbf{k=5}}$	0.49	0.80	0.82	0.58	0.80	0.81	0.61	0.82	0.82
	(187.51)	(43.76)	(35.74)	(139.42)	(69.51)	(50.83)	(131.39)	(46.06)	(39.79)

employed a mixture of finite mixtures (MFM) as a prior distribution for k. A natural and interesting extension of our present work would be to theoretically explore the situation when k remains unknown and an MFM or a mixture of Dirichlet processes prior is used to adaptively learn about k. We leave this as an important research problem in future. In a very recent technical report, Gao et al. (2018) provided general conditions for optimal posterior contraction rates in stochastic block models adaptively for all values of $k \in \{1, 2, \ldots, n\}$ using Laplace-type priors on Q and a complexity prior on k. Their proposed elliptical Laplace prior distribution is theoretically interesting and accommodates many statistical problems in a unified way. Contrary to that, we worked with a more natural and easily implementable uniform prior specification which is widely used in network analysis problems. An interesting direction is to develop a fully Bayesian approach with the more commonly used uniform prior on Q and a complexity prior on k and to show that the corresponding procedure yields optimal rates of posterior contraction adaptively for all values of $k \in \{1, 2, \ldots, n\}$. Such an approach can be connected to nonparametric estimation of networks (Bickel and Chen, 2009) where one typically assumes a more flexible way of data generation; $A_{ij} \mid \xi_i, \xi_j \sim \text{Bernoulli}\{f(\xi_i, \xi_j)\}$, where f is a function from $[0,1]^2 \to [0,1]$, called a *graphon* and ξ_is are i.i.d. random variables on $[0,1]$. It is well known (refer, for example, to Szemerédi 1975; Lovász and Szegedy 2006; Airoldi et al. 2013; Gao et al. 2015) that one can approximate a sufficiently smooth graphon using elements of Θ_k. When the smoothness of the graphon is unknown, the prior on k should facilitate the posterior to concentrate in the appropriate region. Using such approximation results and modifying our Theorem 4.1, it may be possible to derive posterior contraction rates for estimating a graphon.

References

Abramowitz, M. and Stegun, I. (1964). Handbook of mathematical functions: with formulas, graphs, and mathematical tables. No. 55. Courier Corporation.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2009). Mixed membership stochastic blockmodels. In: Advances in Neural Information Processing Systems. pp. 33–40.

Airoldi, E., Costa, T. and Chan, S. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In: Advances in Neural Information Processing Systems. pp. 692–700.

Amini, A. A., Chen, A., Bickel, P. J. and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* **41**, 4, 2097–2122.

BANERJEE, S. and GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics* **8**, 2, 2111–2137.

BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Univ.

BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* **27**, 2, 536–561.

BICKEL, P. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences* **106**, 50, 21068–21073.

BICKEL, P. J., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics* **41**, 4, 1922–1943.

BONTEMPS, D. (2011). Bernstein–von mises theorems for gaussian regression with increasing number of regressors. *The Annals of Statistics* **39**, 5, 2557–2584.

CASTILLO, I. and VAN DER VAART, A. W. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40**, 4, 2069–2101.

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43**, 5, 1986–2018. doi: https://doi.org/10.1214/15-AOS1334.

CHANNAROND, A., DAUDIN, J. -J. and ROBIN, S. (2012). Classification and estimation in the stochastic block model based on the empirical degrees. *Electronic Journal of Statistics* **6**, 2574–2601.

CHATTERJEE, S. (2014). Matrix estimation by universal singular value thresholding. *The Annals of Statistics* **43**, 1, 177–214.

DASGUPTA, A., HOPCROFT, J. E. and MCSHERRY, F. (2004). *Spectral analysis of random graphs with skewed degree distributions*. IEEE, p. 602–610.

ERDŐS, P. and RÉNYI, A. (1961). On the evolution of random graphs. *Bull. Inst. Internat. Statist* **38**, 4, 343–347.

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *Journal of the American Statistical association* **81**, 395, 832–842.

GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics* **43**, 6, 2624–2652.

GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2018). *A general framework for bayes structured linear models*. arXiv:1506.02174.

GENG, J., BHATTACHARYA, A. and PATI, D. (2018). Probabilistic community detection with unknown number of communities. *Journal of American Statistical Association (to appear)*.

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28**, 2, 500–531.

GHOSAL, S. and ROY, A. (2006). Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 2413–2429.

GHOSAL, S. and VAN DER VAART, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35**, 1, 192–223.

GOLDENBERG, A., ZHENG, A., FIENBERG, S. and AIROLDI, E. (2010). A survey of statistical network models. *Foundations and Trends®, in Machine Learning* **2**, 2, 129–233.

GOLIGHTLY, A. and WILKINSON, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61**, 3, 781–788.

HAYASHI, K., KONISHI, T. and KAWAMOTO, T. (2016). *A tractable fully bayesian method for the stochastic block model.* arXiv:1602.02256.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical association* **97**, 460, 1090–1098.

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical association* **76**, 373, 33–50.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5**, 2, 109–137.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83**, 1, 016107.

LOVÁSZ, L. and SZEGEDY, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B* **96**, 6, 933–957.

MCDAID, A., MURPHY, T. B., FRIEL, N. and HURLEY, N. (2013). Improved bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis* **60**, 12–31.

NEWMAN, M. E. J. (2012). Communities, modules and large-scale structure in networks. *Nature Physics* **8**, 1, 25–31.

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

PATI, D., BHATTACHARYA, A., PILLAI, N. S. and DUNSON, D. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics* **42**, 3, 1102–1130.

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 5, 689–710.

SCHWARTZ, L. (1965). On bayes procedures. *Probability Theory and Related Fiel* **4**, 1, 10–26.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14**, 1, 75–100.

SUWAN, S., LEE, D. S., TANG, R., SUSSMAN, D. L., TANG, M. and PRIEBE, C. E. (2016). Empirical bayes estimation for the stochastic block model. *Electronic Journal of Statistics* **10**, 1, 761–782.

SZEMERÉDI, E. (1975). On sets of integers containing no k elements in arithmetic progression. *Acta Arith* **27**, 199-245, 2.

VAN DER PAS, S., KLEIJN, B. and VAN DER VAART, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2, 2585–2618.

VAN DER PAS, S. L. and VAN DER VAART, A. W. (2018). Bayesian community detection. *Bayesian Analysis* **13**, 3, 767–796.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*, 210–268.

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

ZHAO, Y., LEVINA, E. and ZHU, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences* **108**, 18, 7321–7326.

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40**, 4, 2266–2292.

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A. Appedix

A.1 Proof of Corollary 4.3

Following exactly the same set of arguments as in the proof of Theorem 4.1, we have for all sufficiently large n,

$$\prod_{r=1}^{k}\prod_{s=1}^{k}[Q_{rs}^0-\delta\epsilon_n/2, Q_{rs}^0+\delta\epsilon_n/2] \subset \left\{Q : \sum_{r=1}^{k}\sum_{s=1}^{k} n_{0r}n_{0s}(Q_{rs}-Q_{rs}^0)^2 < n^2\delta^2\epsilon_n^2\right\}. \quad (4.26)$$

Since for each (r,s), $Q_{rs}^0 \in (\delta, 1-\delta)$, the prior probability of the embedded rectangle $\prod_{r=1}^{k}\prod_{s=1}^{k}[Q_{rs}^0-\delta\epsilon_n/2, Q_{rs}^0+\delta\epsilon_n/2]$ can be bounded below as follows:

$$\begin{aligned} p\left(\prod_{r=1}^{k}\prod_{s=1}^{k}[Q_{rs}^0-\delta\epsilon_n/2, Q_{rs}^0+\delta\epsilon_n/2]\right) &\geq (\delta\epsilon_n)^{k^2}\prod_{r=1}^{k}\prod_{s=1}^{k}\inf p\left([Q_{rs}^0-\delta\epsilon_n/2, Q_{rs}^0+\delta\epsilon_n/2]\right) \\ &\geq (\delta\epsilon_n)^{k^2}\left\{\inf p[\delta(1-\epsilon_n/2), 1-\delta(1-\epsilon_n/2)]\right\}^{k^2} \\ &= (\mathrm{Beta}(\beta_1,\beta_2))^{-k^2}(\delta\epsilon_n)^{k^2} \times \\ &\quad \left\{\inf_{q\in[\delta(1-\epsilon_n/2),1-\delta(1-\epsilon_n/2)]} q^{\beta_1-1}(1-q)^{\beta_2-1}\right\}^{k^2} \end{aligned} \quad (4.27)$$

where $\mathrm{Beta}(\beta_1,\beta_2)$ denotes the standard Beta function with parameters (β_1,β_2). Next we observe that

$$\inf_{q\in[\delta(1-\epsilon_n/2),1-\delta(1-\epsilon_n/2)]} q^{\beta_1-1}(1-q)^{\beta_2-1} \geq \psi_{\delta,\epsilon_n}(\beta_1,\beta_2), \quad (4.28)$$

where for each fixed (δ,ϵ_n), the function $\psi_{\delta,\epsilon_n}\colon (0,\infty)^2 \to (0,\infty)$ is defined as

$$\psi_{\delta,\epsilon_n}(\beta_1,\beta_2) := \begin{cases} \{\delta(1-\epsilon_n/2)\}^{\beta_1+\beta_2-2} & \text{if } \beta_1\geq 1, \beta_2\geq 1 \\ \{1-\delta(1-\epsilon_n/2)\}^{\beta_1-1}\{\delta(1-\epsilon_n/2)\}^{\beta_2-1} & \text{if } \beta_1<1, \beta_2\geq 1 \\ \{\delta(1-\epsilon_n/2)\}^{\beta_1-1}\{1-\delta(1-\epsilon_n/2)\}^{\beta_2-1} & \text{if } \beta_1\geq 1, \beta_2<1 \\ \{1-\delta(1-\epsilon_n/2)\}^{\beta_1+\beta_2-2} & \text{if } \beta_1<1, \beta_2<1. \end{cases}$$

Using Eqs. 4.26–4.28, and following exactly the same line of arguments as in the proof of Theorem 4.1, we obtain

$$\mathbb{E}_0\left\{\Pi_n(U_n \mid A)1_{\mathcal{A}_n^c}\right\} \leq \sum_{l=M}^{\infty}\left\{e^{-C_1^2 l^2 n^2 \epsilon_n^2} + \frac{e^{-C_2^2 l^2 n^2 \epsilon_n^2} e^{C_3 n \log k}}{(\delta\epsilon_n)^{k^2} C(\beta_1, \beta_2, \delta, \epsilon_n)^{k^2}}\right\}, \tag{4.29}$$

for some constant $C(\beta_1, \beta_2, \delta, \epsilon_n) = (\text{Beta}(\beta_1, \beta_2))^{-k^2}\left(\psi_{\delta,\epsilon_n}(\beta_1, \beta_2)\right)^{k^2} > 0$. Now, for every possible choice of the pair (β_1, β_2), we note that $\log C(\beta_1, \beta_2, \delta, \epsilon_n) \sim k^2$. For instance, suppose $\beta_1 \geq 1, \beta_2 \geq 1$. Then, as $\delta \in (0, 1/2)$ is fixed and $\epsilon_n \to 0$ as $n \to \infty$, $\log C(\beta_1, \beta_2, \delta, \epsilon_n) = -k^2 \log \text{Beta}(\beta_1, \beta_2) + (\beta_1 + \beta_2 - 2)k^2 \log(\delta(1 - \epsilon_n/2)) \sim k^2$. Therefore, for $n^2\epsilon_n^2 = k^2\{\log n + \log(\delta^{-1})\} + n \log k$, $\log C(\beta_1, \beta_2, \delta, \epsilon_n) = o(n^2\epsilon_n^2)$ as $n \to \infty$. Thus, choosing a large enough constant $M > 0$ (depending on (β_1, β_2)), it follows that the above sum in Eq. 4.29 converges to zero for all large values of M which concludes the argument.

A.2 Proof of Theorem 4.4

Observe that the posterior distribution in the case of directed networks can be written as

$$\Pi_n(U_n \mid A) = \frac{\int_{U_n} \prod\prod_{1\leq i<j\leq n} \frac{f_{\theta_{ij}}(A_{ij})}{f_{\theta_{ij}^0}(A_{ij})}\, p(dz, dQ)}{\int_{\Theta_k} \prod\prod_{1\leq i<j\leq n} \frac{f_{\theta_{ij}}(A_{ij})}{f_{\theta_{ij}^0}(A_{ij})}\, p(dz, dQ)}. \tag{4.30}$$

Observe that the discrepancy measure in Eq. 4.8 can also be written as

$$\frac{1}{n^2}\sum\sum_{1\leq i<j\leq n}(\hat{\theta}_{ij} - \theta_{ij}^0)^2 = \frac{1}{2n^2}\left\|\hat{\theta} - \theta^0\right\|^2.$$

for $\theta, \theta^0 \in \Theta_k^u$, defined in Eq. 4.9. Hence, it is straightforward to obtain versions of Lemmata 4.5, 4.6, 4.9 and 4.10 as well as Corollary 4.8 for parameters $\theta \in \Theta_k^u$. The conclusion then follows by replicating arguments (4.21)-(4.25).

A.3 Additional Simulations Results

Below we present an additional small scale simulation study where we simulate 100 replicates of an SBM network using $k = 3$ and 5 equi-sized communities with $n = 30$, 60, and 90 and $\rho = 0.3, 0.5$. We summarize these additional results into Tables 3 and 4 below.

Table 3: MSE ($\times 10^2$) and standard error ($\times 10^3$) comparison over 100 replicates

n	30		60		90	
ρ	0.3	0.5	0.3	0.5	0.3	0.5
$\mathrm{SBM}_{k=3}$	0.83 (0.12)	1.42 (4.77)	0.74 (1.3)	0.24 (4.55)	0.33 (1.78)	0.12 (4.16)
$\mathrm{SBM}_{k=5}$	0.82 (0.22)	1.27 (5.45)	0.72 (1.4)	0.56 (1.31)	0.19 (1.77)	0.02 (0.01)

Table 4: Rand Index and standard error ($\times 10^3$) comparison over 100 replicates

n	30		60		90	
ρ	0.3	0.5	0.3	0.5	0.3	0.5
$\mathrm{SBM}_{k=3}$	0.31 (0.00)	0.78 (132.31)	0.44 (183.65)	0.96 (83.68)	0.83(100.33)	0.98(53.90)
$\mathrm{SBM}_{k=5}$	0.34 (72.67)	0.82 (101.76)	0.46 (181.76)	0.99 (31.95)	0.91 (101.58)	1 (0.00)

Prasenjit Ghosh,
Debdeep Pati and
Anirban Bhattacharya
Department of Statistics,
Texas A&M University,
College Station,
Texas, USA
E-mail: prasenjit@stat.tamu.edu;
debdeep@stat.tamu.edu;
anirbanb@stat.tamu.edu

Paper received: 15 December 2018.