Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Probabilistic Community Detection With Unknown Number of Communities

Junxian Geng, Anirban Bhattacharya & Debdeep Pati

To cite this article: Junxian Geng, Anirban Bhattacharya & Debdeep Pati (2019) Probabilistic Community Detection With Unknown Number of Communities, Journal of the American Statistical Association, 114:526, 893-905, DOI: 10.1080/01621459.2018.1458618

To link to this article: https://doi.org/10.1080/01621459.2018.1458618

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 526, 893–905: Theory and Methods
https://doi.org/10.1080/01621459.2018.1458618

Taylor & Francis
Taylor & Francis Group

Probabilistic Community Detection With Unknown Number of Communities

Junxian Geng[a], Anirban Bhattacharya[b], and Debdeep Pati[b]

[a]Department of Statistics, Florida State University, Tallahassee, FL; [b]Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

A fundamental problem in network analysis is clustering the nodes into groups which share a similar connectivity pattern. Existing algorithms for community detection assume the knowledge of the number of clusters or estimate it a priori using various selection criteria and subsequently estimate the community structure. Ignoring the uncertainty in the first stage may lead to erroneous clustering, particularly when the community structure is vague. We instead propose a coherent probabilistic framework for simultaneous estimation of the number of communities and the community structure, adapting recently developed Bayesian nonparametric techniques to network models. An efficient Markov chain Monte Carlo (MCMC) algorithm is proposed which obviates the need to perform reversible jump MCMC on the number of clusters. The methodology is shown to outperform recently developed community detection algorithms in a variety of synthetic data examples and in benchmark real-datasets. Using an appropriate metric on the space of all configurations, we develop nonasymptotic Bayes risk bounds even when the number of clusters is unknown. Enroute, we develop concentration properties of nonlinear functions of Bernoulli random variables, which may be of independent interest in analysis of related models. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received March 2016
Revised January 2018

KEYWORDS
Bayesian nonparametrics; Clustering consistency; MCMC; Mixture models; Model selection; Network analysis

1. Introduction

Data available in the form of networks are increasingly becoming common in modern applications ranging from brain remote activity, protein interactions, web applications, social networks to name a few. Accordingly, there has been an explosion of activities in the statistical analysis of networks in recent years; see Goldenberg et al. (2010) for a review of various application areas and statistical models. Among various methodological and theoretical developments, the problem of community detection has received widespread attention. Broadly speaking, the aim there is to cluster the network nodes into groups which share a similar connectivity pattern, with sparser intergroup connections compared to more dense within-group connectivities; a pattern which is observed empirically in a variety of networks (Goldenberg, Libai, and Muller 2001). Various statistical approaches have been proposed for community detection and extraction. These include hierarchical clustering (see Newman 2004 for a review), spectral clustering (White and Smyth 2005; Zhang, Wang, and Zhang 2007; Rohe, Chatterjee, and Yu 2011), and algorithms based on optimizing a global criterion over all possible partitions, such as normalized cuts (Shi and Malik 2000) and network modularity (Newman and Girvan 2004).

From a model-based perspective, the stochastic block model (SBM; Holland, Laskey, and Leinhardt 1983) and its various extensions (Airoldi et al. 2008; Karrer and Newman 2011) enable formation of communities in networks. A generic formulation of an SBM starts with clustering the nodes into groups, with the edge probabilities $\mathbb{E}A_{ij} = \theta_{ij}$ solely dependent on the cluster memberships of the connecting nodes. A realization of a network from an SBM is shown in Figure 1; formation of a community structure is clearly evident. This clustering property of SBMs has inspired a large literature on community detection (Bickel and Chen 2009; Karrer and Newman 2011; Zhao, Levina, and Zhu 2011; Newman 2012; Zhao, Levina, and Zhu 2012; Amini et al. 2013).

A primary challenge in community detection is the estimation of both the number of communities and the clustering configurations. Essentially all existing community detection algorithms assume the knowledge of the number of communities (Airoldi et al. 2009; Bickel and Chen 2009; Amini et al. 2013) or estimate it a priori using either of cross-validation, hypothesis testing, BIC, or spectral methods (Daudin, Picard, and Robin 2008; Latouche, Birmele, and Ambroise 2012; Wang and Bickel 2015; Le and Levina 2015). Such two-stage procedures ignore uncertainty in the first stage and are prone to increased erroneous cluster assignments when there is inherent variability in the number of communities. Although model-based methods are attractive for inference and quantifying uncertainty, fitting block models from a frequentist point of view, even with the number of communities known, is a nontrivial task especially for large networks, since in principle the problem of optimizing over all possible label assignments is NP-hard.

Bayesian inference offers a natural solution to this problem by providing a probabilistic framework for simultaneous inference of the number of clusters and the clustering configurations. However, the case of unknown number of communities poses a stiff computational challenge even in a fully Bayes framework. Snijders and Nowicki (1997) and Nowicki and Snijders (2001)

CONTACT Junxian Geng jgeng@stat.fsu.edu Department of Statistics, Florida State University, Tallahassee, FL 32306.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

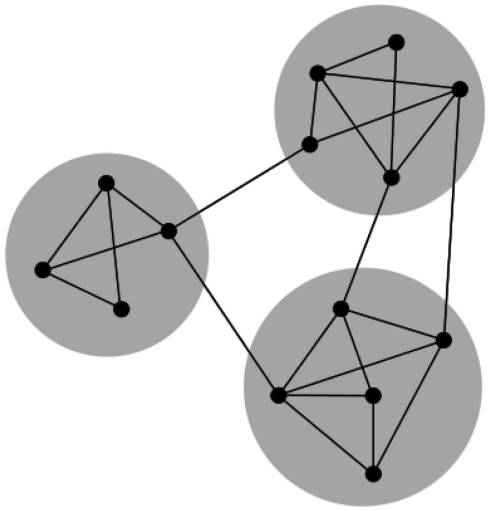

Figure 1. A sketch of a network displaying community structure, with three groups of nodes with dense internal edges and sparser edges among groups.

developed a Markov chain Monte Carlo (MCMC) algorithm to estimate the parameters in an SBM for a given number of communities. Often, a frequentist estimate of k is first determined through a suitable criterion, for example, integrated likelihood (Daudin, Picard, and Robin 2008; Zanghi, Ambroise, and Miele 2008; Latouche, Birmele, and Ambroise 2012), composite likelihood BIC (Saldana, Yu, and Feng 2015), etc., with a subsequent Bayesian model fitted with the estimated number of components. In a fully Bayesian framework, a prior distribution is assigned on the number of communities which is required to be updated at each iteration of an MCMC algorithm. This calls for complicated search algorithms in variable dimensional parameter space such as the reversible jump MCMC algorithm (Green 1995), which are difficult to implement and automate, and are known to suffer from lack of scalability and mixing issues. McDaid et al. (2013) proposed an algorithm by "collapsing" some of the nuisance parameters, which allows them to implement an efficient algorithm based on the allocation sampler of Nobile and Fearnside (2007). However, the parameter (k) indicating the number of components still cannot be marginalized out within the Gibbs sampler requiring complicated Metropolis moves to simultaneously update the clustering configurations and k.

In this article, we consider a Bayesian formulation of an SBM (Snijders and Nowicki 1997; Nowicki and Snijders 2001; McDaid et al. 2013) with standard conjugate Dirichlet-Multinomial prior on the community assignments and Beta priors on the edge probabilities. Our contribution is two-folds. First, we allow simultaneous learning of the number of communities and the community memberships via a prior on the number of communities k. A seemingly automatic choice to allow uncertainty in the number of communities is to use a Bayesian nonparametric approach such as the Chinese restaurant process (CRP; Pitman 1995). While it has been empirically observed that CRPs often have the tendency to create tiny extraneous clusters, it has only been recently established that CRPs lead to inconsistent estimation of the number of clusters in a fairly general setting (Miller and Harrison 2017). We instead adapt the mixture of finite mixture (MFM) approach of Miller and Harrison (2017) which alleviates the drawback of CRP by automatic model-based pruning of the tiny extraneous clusters leading to consistent estimate of the number of clusters. Moreover, MFM admits a clustering scheme similar to the CRP which is exploited to develop an efficient MCMC algorithm. In particular, we analytically marginalize over the number of communities to obtain an efficient Gibbs sampler and avoid resorting to complicated reversible jump MCMC algorithms or allocation samplers. We exhibit the efficacy of our proposed MFM-SBM approach over existing two-stage approaches and the CRP prior through various simulation examples. We envision simple extensions of MFM-SBM to degree corrected SBM (Karrer and Newman 2011) and mixed membership block model (Airoldi et al. 2008), which will be reported elsewhere.

Our second contribution is to develop a framework for consistent community detection, where we derive nonasymptotic bounds on the posterior probability of the true configuration. As a consequence, we can show that the marginal posterior distribution on the set of community assignments increasingly concentrates (in an appropriate sense) on the true configuration with increasing number of nodes. This is a stronger statement than claiming that the true configuration is the maximum a posteriori model with the highest posterior probability. Although there is now a well-established literature on posterior convergence in density estimation and associated functionals in Bayesian nonparametric mixture models (see, e.g., Kruijer et al. 2010 and references therein), there are no existing results on clustering consistency in network models or beyond to best of our knowledge. In fact, the question of consistency of the number of mixture components has only been resolved very recently (Rousseau and Mengersen 2011; Miller and Harrison 2017). Clustering consistency is clearly a stronger requirement and significantly more challenging to obtain than consistency of the number of mixture components. We exploit the conjugate nature of the Bayesian SBM to obtain the marginal likelihoods for each cluster configuration, and subsequently use probabilistic bounds on the log-marginal likelihood ratios to deliver our nonasymptotic bound. We hope our results on selection consistency have a broader appeal to the Bayesian model selection community; see in particular the second paragraph in Section 4 for a detailed discussion.

The rest of the article is organized as follows. We start with a brief review of the SBM in Section 2. The Bayesian methods for simultaneous inference on the number of clusters and the clustering configurations are discussed in Section 3 and the Gibbs sampler is provided in Section 3.1. The theory for consistent community detection is developed in Section 4. Simulation studies and comparisons with existing methods are provided in Section 5 and illustration of our method on a benchmark real dataset is in Section 6. Additional simulations exploring sensitivity, convergence diagnostics, and robustness, and proofs of all technical results, are provided in a separate supplemental document. The supplemental document additionally contains a second real data example.

2. Stochastic Block Models

We use $\mathcal{A} = (A_{ij}) \in \{0, 1\}^{n\times n}$ to 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. We consider undirected networks without self-loops so that $A_{ij} = A_{ji}$ and $A_{ii} = 0$. The sampling algorithms presented here can be trivially modified to directed networks with or without self-loops. The theory would require some additional work in case of directed networks though conceptually a straightforward modification of the current results should go through.

The probability of an edge from node i to j is denoted by θ_{ij}, with $A_{ij} \sim \text{Bernoulli}(\theta_{ij})$ independently for $1 \le i < j \le n$. In a k-component SBM, the nodes are clustered into communities, with the probability of an edge between two nodes solely dependent on their community memberships. Specifically,

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

where $z_i \in \{1, \ldots, k\}$ denotes the community membership of the ith node and $Q = (Q_{rs}) \in [0,1]^{k \times k}$ is a symmetric matrix of probabilities, with $Q_{rs} = Q_{sr}$ indicating the probability of an edge between any node i in cluster r and any node j in cluster s.

Let $\mathcal{Z}_{n,k} = \{(z_1, \ldots, z_n) : z_i \in \{1, \ldots, k\}, 1 \le i \le n\}$ denote all possible clusterings of n nodes into k clusters. 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 (1) can be expressed as

$$\begin{aligned} P(A \mid z, Q, k) &= \prod_{1 \le r \le s \le k} P(A_{[rs]} \mid z, Q), \\ P(A_{[rs]} \mid z, Q, k) &= \prod_{1 \le i < j \le n : z_i = r, z_j = s} Q_{rs}^{A_{ij}} (1 - Q_{rs})^{1 - A_{ij}}. \end{aligned} \tag{2}$$

A common Bayesian specification of the SBM when k is given can be completed by assigning independent priors to z and Q. We generically use $p(z, Q) = p(z)p(Q)$ to denote the joint prior on z and Q. When K (the true number of clusters) is unknown, a natural Bayesian solution is to place a prior on k. This is described in Section 3.

3. Bayesian Community Detection in SBM

A natural choice of a prior distribution on $(z_1, z_2, \ldots, z_n)$ that allows automatic inference on the number of clusters k is the CRP (Aldous 1985; Pitman 1995; Neal 2000). A CRP is described through the popular Chinese restaurant metaphor: imagine customers arriving at a Chinese restaurant with infinitely many tables with the index of the table having a one-one correspondence with the cluster label. The first customer is seated at the first table, so that $z_1 = 1$. Then z_i, $i = 2, \ldots, n$ are defined through the following conditional distribution (also called a Pólya urn scheme, Blackwell and MacQueen 1973)

$$P(z_i = c \mid z_1, \ldots, z_{i-1}) \propto \begin{cases} |c|, & \text{at an existing table labeled } c \\ \alpha, & \text{if } c \text{ is a new table.} \end{cases} \tag{3}$$

The above prior for $\{z_i\}$ can also be defined through a stochastic process where at any positive-integer time n, the value of the process is a partition $\mathcal{C}_n$ of the set $\{1, 2, 3, \ldots, n\}$, whose probability distribution is determined as follows. At time $n = 1$, the trivial partition $\{\{1\}\}$ is obtained with probability 1. At time $n + 1$ the element $n + 1$ is either (i) added to one of the blocks of the partition $\mathcal{C}_n$, where each block is chosen with probability $|c|/(n+1)$ where $|c|$ is the size of the block, or (ii) added to the partition $\mathcal{C}_n$ as a new singleton block, with probability $1/(n+1)$. Marginally, the distribution of z_i is given by the stick-breaking formulation of a Dirichlet process (Sethuraman 1994):

$$z_i \sim \sum_{h=1}^{\infty} \pi_h \delta_h, \quad \pi_h = \nu_h \prod_{l<h} (1 - \nu_l), \quad \nu_h \sim \text{Beta}(1, \alpha). \tag{4}$$

Let $t = |\mathcal{C}_n|$ denote the number of blocks in the partition $\mathcal{C}_n$. Under (3), one can obtain the probability of block-sizes $s = (s_1, s_2, \ldots, s_t)$ of a partition $\mathcal{C}_n$ as

$$p_{\text{DP}}(s) \propto \prod_{j=1}^{t} s_j^{-1}. \tag{5}$$

It is clear from (5) that CRP assigns large probabilities to clusters with relatively smaller size. A striking consequence of this has been recently discovered (Miller and Harrison 2017) where it is shown that the CRP produces extraneous clusters in the posterior leading to inconsistent estimation of the *number of clusters* even when the sample size grows to infinity. Miller and Harrison (2017) proposed a modification of the CRP based on a mixture of finite mixtures (MFM) model to circumvent this issue:

$$\begin{aligned} k &\sim p(\cdot), \quad (\pi_1, \ldots, \pi_k) \mid k \sim \text{Dir}(\gamma, \ldots, \gamma), \\ z_i \mid k, \pi &\sim \sum_{h=1}^{k} \pi_h \delta_h, \quad i = 1, \ldots, n, \end{aligned} \tag{6}$$

where $p(\cdot)$ is a proper p.m.f on $\{1, 2, \ldots, \}$ and δ_h is a point-mass at h. Miller and Harrison (2017) showed that the joint distribution of $(z_1, \ldots, z_n)$ under (6) admit a Pólya urn scheme akin to CRP:

1. Initialize with a single cluster consisting of element 1 alone: $\mathcal{C}_1 = \{\{1\}\}$,
2. For $n = 2, 3, \ldots,$ place element n in
   (a) an existing cluster $c \in \mathcal{C}_{n-1}$ with probability $\propto |c| + \gamma$
   (b) a new cluster with probability $\propto \frac{V_n(t+1)}{V_n(t)} \gamma$
   where $t = |\mathcal{C}_{n-1}|$.

$V_n(t)$ is a coefficient of partition distribution that need to be precomputed in this model,

$$V_n(t) = \sum_{n=1}^{+\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p(k),$$

where $k_{(t)} = k(k-1)\ldots(k-t+1)$, and $(\gamma k)^{(n)} = \gamma k(\gamma k + 1)\ldots(\gamma k + n - 1)$. (By convention, $x^{(0)} = 1$ and $x_{(0)} = 1$).

Compared to the CRP, the introduction of new tables is slowed down by the factor $V_n(|\mathcal{C}_{n-1}| + 1)/V_n(|\mathcal{C}_{n-1}|)$, thereby allowing a model-based pruning of the tiny extraneous clusters. An alternative way to understand this is to look at the probability of block-sizes $s = (s_1, s_2, \ldots, s_t)$ of a partition $\mathcal{C}_n$ with $t = |\mathcal{C}_n|$ under MFM. As opposed to (5), the probability of the cluster-sizes $(s_1, \ldots, s_t)$ under MFM is

$$p_{\text{MFM}}(s) \propto \prod_{j=1}^{t} s_j^{\gamma - 1}. \tag{7}$$

From (5) and (7), it is easy to see that MFM assigns comparatively smaller probability to clusters with small sizes. The parameter γ controls the relative size of the clusters; small γ favors lower entropy π's, while large γ favors higher entropy π's.

Adapting MFM to the SBM setting, our model and prior can be expressed hierarchically as

$$\begin{aligned} &k \sim p(\cdot), \text{ where} p(\cdot) \text{is a p.m.f on } \{1, 2, \ldots\} \\ &Q_{rs} = Q_{sr} \overset{\text{ind}}{\sim} \text{Beta}(a, b), \quad r, s = 1, \ldots, k, \\ &\text{pr}(z_i = j \mid \pi, k) = \pi_j, \quad j = 1, \ldots, k, \ i = 1, \ldots, n, \\ &\pi \mid k \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \\ &A_{ij} \mid z, Q, k \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = Q_{z_i z_j}, \quad 1 \le i < j \le n. \end{aligned} \tag{8}$$

A default choice of $p(\cdot)$ is a Poisson(1) distribution truncated to be positive (Miller and Harrison 2017), which is assumed through the rest of the article. We refer to the hierarchical model above as MFM-SBM. While MFM-SBM admits a CRP representation, an important distinction from infinite mixture models hinges on the fact that for any given prior predictive realization, one draws a value of k and as n grows the individuals are distributed into the k clusters. On the other hand, the number of clusters keeps growing with n for the infinite mixture models.

3.1. Gibbs Sampler

Our goal is to sample from the posterior distribution of the unknown parameters $k, z = (z_1, \ldots, z_n) \in \{1, \ldots, k\}^n$ and $Q = (Q_{rs}) \in [0, 1]^{k \times k}$. Miller and Harrison (2017) developed the MFM approach for clustering in mixture models, where their main trick was to analytically marginalize over the distribution of k. While MFM-SBM is different from a standard Bayesian mixture model, we could still exploit the Pólya urn scheme for MFMs to analytically marginalize over k and develop an efficient Gibbs sampler. The sampler is presented in Algorithm 1 in Appendix A of the supplemental document, which efficiently cycles through the full conditional distribution of Q and $z_i \mid z_{-i}$ for $i = 1, 2, \ldots, n$, where $z_{-i} = z \backslash \{z_i\}$. The marginalization over k allows us to avoid complicated reversible jump MCMC algorithms or even allocation samplers. In practice, one way to initialize the number of clusters is to use a frequentist approach (e.g., Le and Levina 2015). For the initialization of cluster configurations, we randomly assign all observations into those clusters.

4. Consistent Community Detection

In this section, we provide theoretical justification to the proposed approach by showing that marginal posterior distribution on the space of community assignments concentrates on the truth exponentially fast as the number of nodes increases. At the very onset, some clarification is required regarding the mode of convergence, since the community assignments are only identifiable up to arbitrary labeling of the community indicators within each community. For example, in a network of 5 nodes with 2 communities, consider two community assignments z and z', with $z_1 = z_3 = z_5 = 1$ and $z_2 = z_4 = 2$; and $z'_1 = z'_3 = z'_5 = 2$ and $z'_2 = z'_4 = 1$. Clearly, although z and z' are different as 5-tuples, they imply the same community structure and the posterior cannot differentiate between z and z'. To bypass such *label switching* issues, we consider a permutation-invariant Hamming distance introduced in Zhang et al. (2016) as our loss function and bound the posterior expected loss (equivalently, the Bayes risk) with large probability under the true data-generating mechanism. The concentration of the posterior on the true community assignment (up to labeling) follows as a straightforward corollary of the Bayes risk bound.

Consistency results for our Bayesian procedure complement a series of recent frequentist work on consistent community detection (Bickel and Chen 2009; Zhao, Levina, and Zhu 2012; Gao et al. 2017; Abbe and Sandon 2015a, 2015c, 2015b; Zhang et al. 2016) among others. From a Bayesian viewpoint, our result contributes to a growing literature on consistency of Bayesian model selection procedures when the number of competing models grow exponentially relative to the sample size (Johnson and Rossell 2012; Narisetty et al. 2014; Castillo, Schmidt-Hieber, and van der Vaart 2015; Shin, Bhattacharya, and Johnson 2018). Our present problem has two key distinctions from these existing results which primarily focus on variable selection in (generalized) linear models: (a) the model space does not have a natural nested structure as in case of (generalized) linear models, which requires additional care in enumeration of the space of community assignments; and (b) the log-marginal likelihood differences between a putative community assignment and the truth is not readily expressible as a χ^2-statistic, necessitating careful analysis of such objects.

4.1. Preliminaries

We introduce some basic notations here that are required to state our main results. Notations that only appear in proofs are introduced at appropriate places in the supplemental document.

Throughout C, C', etc., denote constants that are independent of everything else but whose values may change from one line to the other. $\mathbb{1}(B)$ denotes the indicator function of set B. For two vectors $x = \{x_i\}$ and $y = \{y_i\}$ of equal length n, the Hamming distance between x and y is $d_H(x, y) = \sum_{i=1}^n \mathbb{1}(x_i \neq y_i)$. For any positive integer m, let $[m] := \{1, \ldots, m\}$. A community assignment of n nodes into $K < n$ communities is given by $z = (z_1, \ldots, z_n)^\top$ with $z_i \in [K]$ for each $i \in [n]$. Let $\mathcal{Z}_{n,K}$ denote the space of all such community assignments. For a permutation δ on $[K]$, define $\delta \circ z$ as the community assignment given by $\delta \circ z(i) = \delta(z_i)$ for $i \in [n]$. Clearly, $\delta \circ z$ and z provide the same clustering up to community labels. Define $\langle z \rangle$ to be the collection of $\delta \circ z$ for all permutations δ on $[K]$; we shall refer to $\langle z \rangle$ as the equivalence class of z. Define a permutation-invariant Hamming distance (see Zhang et al. 2016)

$$d(z, z') = \inf_{\delta} d_H(\delta \circ z, z'), \tag{9}$$

where the infimum is over all permutations of $[K]$. Note that $d(z, z') = 0$ if and only if z and z' are in the same equivalence class, that is, $\langle z \rangle = \langle z' \rangle$.

4.2. Homogeneous SBMs

To state our theoretical result, we restrict attention to *homogeneous SBMs*. An SBM is called homogeneous when the Q matrix in (1) has a compound-symmetry structure, with $Q_{rs} = q + (p-q)I(r=s)$, so that all diagonal entries of Q are p and all off-diagonal entries are q. Thus, the edge probabilities

$$\theta_{ij} = \begin{cases} p & \text{if } z_i = z_j, \\ q & \text{if } z_i \neq z_j. \end{cases}$$

For a homogeneous SBM, the likelihood function for p, q, z, k assumes the form

$$\begin{aligned} f(\mathcal{A} \mid z, p, q, k) &= \prod_{i<j} \theta_{ij}^{a_{ij}} (1-\theta_{ij})^{1-a_{ij}} \\ &= p^{A_\uparrow(z)}(1-p)^{n_\uparrow(z)-A_\uparrow(z)} q^{A_\downarrow(z)}(1-q)^{n_\downarrow(z)-A_\downarrow(z)}, \end{aligned} \tag{10}$$

where

$$n_\uparrow(z) = \sum_{i<j} \mathbb{1}(z_i = z_j), \quad A_\uparrow(z) = \sum_{i<j} a_{ij}\mathbb{1}(z_i = z_j), \tag{11}$$

$$n_\downarrow(z) = \sum_{i<j} \mathbb{1}(z_i \neq z_j), \quad A_\downarrow(z) = \sum_{i<j} a_{ij}\mathbb{1}(z_i \neq z_j). \tag{12}$$

Clearly, $n_\downarrow(z) = \binom{n}{2} - n_\uparrow(z)$.

As in Section 3, we consider independent $\mathrm{U}(0,1)$ priors on p and q. A key object is the *marginal likelihood* of z, denoted $\mathcal{L}(\mathcal{A} \mid z, k)$, obtained by integrating over the priors on p and q. Exploiting Beta-binomial conjugacy, we have,

$$\begin{aligned} \mathcal{L}(\mathcal{A} \mid z, k) &= \left\{ \int_0^1 p^{A_\uparrow(z)}(1-p)^{n_\uparrow(z)-A_\uparrow(z)} dp \right\} \\ &\quad \times \left\{ \int_0^1 q^{A_\downarrow(z)}(1-q)^{n_\downarrow(z)-A_\downarrow(z)} dq \right\} \\ &= \frac{1}{n_\uparrow(z)+1} \frac{1}{\binom{n_\uparrow(z)}{A_\uparrow(z)}} \frac{1}{n_\downarrow(z)+1} \frac{1}{\binom{n_\downarrow(z)}{A_\downarrow(z)}}. \end{aligned} \tag{13}$$

Letting $\Pi(z \mid k)$ denote the prior probability of the community assignment z conditional on k, its posterior probability $\Pi(z \mid k, \mathcal{A}) \propto \mathcal{L}(\mathcal{A} \mid z, k)\Pi(z \mid k)$. Observe that each one of $n_\uparrow(z)$, $n_\downarrow(z)$, $A_\uparrow(z)$, and $A_\downarrow(z)$ are labeling invariant, that is, they assume a constant value on $\langle z \rangle$, and hence so is $\mathcal{L}(\mathcal{A} \mid z, k)$. Hence, as long as the prior $\Pi(\cdot \mid k)$ is labeling invariant, the same can thus be concluded regarding the posterior $\Pi(\cdot \mid k, \mathcal{A})$. For example, the Dirichlet-multinomial prior (conditional on k) in (6) in Section 3 is labeling invariant.

4.3. Main Result for Known K Case

Our first set of results pertain to the case when the number of communities K is fixed and known. We assume the true network-generating model is a homogeneous SBM with K communities, and true within- and between-community edge probabilities p_0 and q_0, respectively. We note that unlike several existing results, we do not assume knowledge of p_0 and q_0. Let z_0 denote the true community assignment.

We state our assumptions on these quantities below.

(A1) Assume the number of nodes n is an integer multiple of K, with each community having an equal size of n/K. Without loss of generality, we assume that $z_{0i} = \lfloor (i-1)/K \rfloor + 1$ for $i = 1, \ldots, n$.

(A2) The true edge probabilities $p_0 \neq q_0$ satisfy $n\bar{D}(p_0, q_0)/K \to \infty$ as $n \to \infty$, where

$$\bar{D}(p_0, q_0) := \frac{(p_0 - q_0)^2}{(p_0 \vee q_0)\{1 - (p_0 \wedge q_0)\}} \tag{14}$$

with $\vee$ and $\wedge$ denoting maximum and minimum, respectively.

(A1) assumes a balanced network which is fairly common in the literature; see, for example, Zhang et al. (2016). Extension to the case where the community sizes are unequal but of the same order can be accomplished, albeit with substantially more tedious counting arguments. Condition (A2) is automatically satisfied if p_0 and q_0 do not vary with n. However, (A2) is much stronger in that one can accommodate *sparse networks* where p_0 and q_0 decay to zero. Indeed, parameterizing $p_0 = a/n$ and $q_0 = b/n$, the condition in (A2) amounts to $(a-b)^2/(a \vee b) \to \infty$. Recent information-theoretical results (Theorem 1.1 of Zhang et al. 2016, eq. (16) in Abbe and Sandon 2015a) show that the condition $(a-b)^2/(a \vee b) \to \infty$ is necessary for complete recovery of the community assignments. The quantity $\bar{D}(p_0, q_0)$ is closely related to Renyi divergence measures between Bernoulli(p_0) and Bernoulli(q_0) distributions that appear in the information-theoretical lower bounds.

We next state a Lipschitz-type condition on the log-prior mass on the community assignments.

(P1) Assume z_0 satisfies (A1). The prior Π on $\mathcal{Z}_{n,K}$ satisfies

$$\left| \log \Pi(z) - \log \Pi(z_0) \right| \leq CKd(z, z_0), \tag{15}$$

for all $z \in \mathcal{Z}_{n,K}$.

Remark 1. (P1) requires $\log \Pi(\cdot)$ to be Lipschitz continuous with respect to the distance d, with Lipschitz constant bounded by a multiple of K. (P1) is satisfied by the Dirichlet-multinomial prior in Section 3. Straightforward calculations yield, for the Dirichlet-multinomial prior with Dirichlet concentration parameter γ,

$$\frac{\Pi(z)}{\Pi(z_0)} = \prod_{h=1}^{K} \frac{\Gamma(n_h(z)+\gamma)}{\Gamma(n/K+\gamma)},$$

where, recall $n_h(z) = \sum_{i=1}^n \mathbb{1}(z_i = h)$. The inequality (15) follows from an application of the following two-sided bound for the gamma function: for any $x > 0$, $\log \Gamma(x) = (x - 1/2)\log x - x + R(x)$, with $0 < R(x) < (12x)^{-1}$.

Let $\mathbb{P}$ denote probability under the true data-generating mechanism. We now provide a bound to the posterior expected loss of $d(z, z_0)$, that is, $E[d(z, z_0) \mid \mathcal{A}]$, that holds with large $\mathbb{P}$-probability (w.r.t. $\mathcal{A}$), in Theorem 1. The proof is deferred to Appendix E of the supplemental document.

Theorem 1. Recall the permutation-invariant Hamming distance $d(\cdot, \cdot)$ from (9). Assume the true cluster assignment z_0 satisfies (A1), and the true within and between edge probabilities p_0 and q_0 satisfy (A2). Also, assume that the prior Π on $\mathcal{Z}_{n,K}$

satisfies (P1). Then,

$$E[d(z, z_0) \mid \mathcal{A}] \leq \exp\left\{-\frac{Cn\bar{D}(p_0, q_0)}{K}\right\},$$

holds with $\mathbb{P}$-probability at least $1 - e^{-C(\log n)^v}$ for some $v > 1$.

An immediate corollary of Theorem 1 is that the posterior almost surely concentrates on the true configuration z_0. To see this, let $\mathcal{C}$ denote the large $\mathbb{P}$-probability set in Theorem 1. We have, inside $\mathcal{C}$,

$$\begin{aligned}\Pi[\langle z\rangle = \langle z_0\rangle \mid \mathcal{A}] &= \Pi[d(z, z_0) = 0 \mid \mathcal{A}] \\ &= 1 - \Pi[d(z, z_0) > 1 \mid \mathcal{A}] \geq 1 - \exp\left\{-\frac{Cn\bar{D}(p_0, q_0)}{K}\right\},\end{aligned}$$

where the penultimate inequality follows from Markov's inequality. We summarize in the following Corollary which is a straightforward application of the first Borel–Cantelli Lemma.

Corollary 1. Suppose the conclusion of Theorem 1 holds. Then,

$$\begin{aligned}&\Pi[\langle z\rangle = \langle z_0\rangle \mid \mathcal{A}] \\ &\quad\geq 1 - \exp\left\{-\frac{Cn\bar{D}(p_0, q_0)}{K}\right\} \quad \text{almost surely } \mathbb{P} \text{ as } n \to \infty.\end{aligned}$$

Corollary 1 ensures that as $n \to \infty$, for almost every network sampled from $\mathbb{P}$, $\Pi[\langle z\rangle = \langle z_0\rangle \mid \mathcal{A}]$ is close to 1 at the same rate obtained in Theorem 1. This is possible since $\mathbb{P}(\mathcal{C}^c)$ decreases sufficiently fast to 0 as $n \to \infty$.

The proof of Theorem 1 is lengthy and thus provided in Appendix E of the supplemental document. We briefly comment on some of the salient aspects here. The key ingredient in proving Theorem 1 is to uniformly bound from below the difference in log-marginal marginal likelihood between the true community assignment z_0 and a putative community assignment z with $d(z, z_0) = r$. As a first step, we approximate the log-marginal likelihood $\log \mathcal{L}(\mathcal{A} \mid z)$ by $\widetilde{\ell}(z) := n_\uparrow(z)h\{A_\uparrow(z)/n_\uparrow(z)\} + n_\downarrow(z)h\{A_\downarrow(z)/n_\downarrow(z)\}$, where $h(x) = x\log x + (1-x)\log(1-x)$ for $x \in (0, 1)$. This is essentially a Laplace approximation of the log-marginal likelihood and the error in approximation can be bounded appropriately. We construct a set $\mathcal{C}$ with $\mathbb{P}(\mathcal{C}) \geq 1 - e^{-C(\log n)^v}$ in Proposition E.1 stated in the supplemental document such that within $\mathcal{C}$,

$$\widetilde{\ell}(z_0) - \widetilde{\ell}(z) \geq \frac{C\bar{D}(p_0, q_0)\, n\, d(z, z_0)}{K}, \tag{16}$$

for all $z \in \mathcal{Z}_{n,K}$. Equation (16) combined with the prior mass condition (P1) essentially delivers the proof of Theorem 1.

A couple of intertwined technical challenges show up in obtaining a concentration bound of the form (16). First, the random quantities $\widetilde{\ell}(z_0)$ and $\widetilde{\ell}(z)$ can be highly dependent, particularly when $d(z, z_0)$ is small, which rules out separately analyzing the concentration of each term around its expectation. However, a combined analysis of the difference is complicated by the presence of the nonlinear function h. We note that h is non-Lipschitz, and hence standard concentration inequalities for Lipschitz functions of several independent variables cannot be applied. We crucially exploit convexity of h to analyze the difference $\widetilde{\ell}(z_0) - \widetilde{\ell}(z)$. A careful combinatorial analysis of terms arising inside the bounds (Lemma E.1 in the supplemental document) along with concentration inequalities for sub-Gaussian random variables (Vershynin 2012) deliver the desired bound.

4.4. Main Result for Unknown K Case

We now partially aim to answer the question: if the true K is unknown and a prior is imposed on k which assigns positive mass to the true K, can we recover K and the true community assignment z_0 from the posterior? To best of our knowledge, this question has not been settled even for usual mixture models, and a complete treatment for SBMs is beyond the scope of this article. An inspection of the proof of Proposition E.1 in the supplemental document will reveal that the only place where the fact that both z and z_0 lie in $\mathcal{Z}_{n,K}$ has been used in Lemma E.1. The primary difficulty in extending the theoretical results in the previous subsection to the variable k case precisely lie in generalizing the combinatorial bounds in Lemma E.1. Recall the metric d in (9) is defined on $\mathcal{Z}_{n,K}$. To define $d(z_1, z_2)$ for $z_1 \in \mathcal{Z}_{n,K_1}$ and $z_2 \in \mathcal{Z}_{n,K_2}$, an option is to embed all the $Z_{n,k}$'s inside $\cap_{k=1}^{K_{max}} Z_{n,k}$, where $K_{\max}$ is an upper bound on the number of communities. This substantially complicates the analysis as one now has to take into account zero counts for one or more communities in obtaining the combinatorial bounds.

We consider the following simplified setting. Suppose the true K can be either 2 or 3. Given K, the network is generated exactly as in the previous subsection, that is, according to a homogeneous SBM with equal-sized communities satisfying (A1) and (A2). We do not assume knowledge of the true K, and use an MFM-SBM model with a prior on k supported on $\{2, 3\}$. We only require $\Pi(k)$ to have positive probability on both 2 and 3. We show below that the posterior of k concentrates on the true K, characterizing the rate of concentration.

Theorem 2. Assume the true cluster assignment z_0 satisfies (A1) with $K \in \{2, 3\}$, and the true within and between edge probabilities p_0 and q_0 satisfy (A2). Also, assume that the prior Π on $\mathcal{Z}_{n,k}$ satisfies (P1) conditional on k and $\Pi(k) > 0$ for $k \in \{2, 3\}$. Then,

$$\Pi(k = K \mid \mathcal{A}) \geq 1 - \exp\{-cn^q\},$$

for some constant $c > 0$, with $\mathbb{P}$-probability at least $1 - e^{-t_n}$ for $t_n \to \infty$ where $q = 1$ and $t_n = o(\sqrt{n})$ for $K = 2$ and $q = 2$ and $t_n = o(n)$ for $K = 3$.

The proof is deferred to Appendix F of the supplemental document. Theorem 2 is an illustration of model-selection consistency when the goal is to identify the number of clusters K. In the overfitted case when $K = 2$ and the model is fitted with $k = 3$, the posterior can successfully "empty-out" the extraneous cluster and recover the true number of clusters. The likelihood of the SBM can potentially derive strength from $O(n^2)$ edges as opposed to $O(n)$ data points in standard regression and mixture models. In the overfitted case when $K = 2$ and the model is fitted with $k = 3$, the marginal likelihood ratio corresponding to a given configuration z against the null z_0 becomes the weakest when the Rand index between z and z_0 is close to 1. In this case, the marginal likelihood ratio corresponding to $k = 3$ and

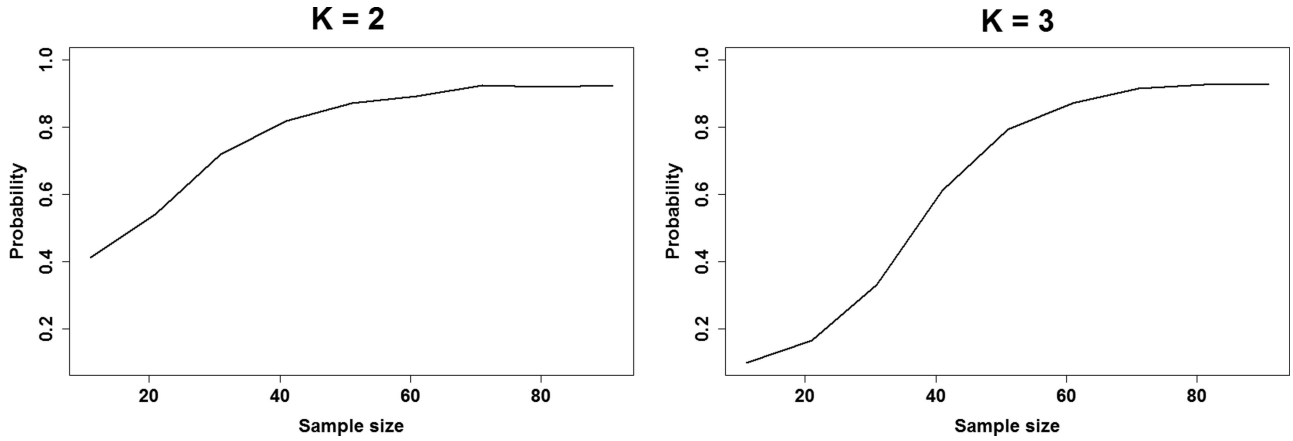

Figure 2. Growth rate of the posterior probability of the true number of components, $\Pi(k = K \mid \mathcal{A})$, as sample size n increases, under the setup of Theorem 2. Left panel corresponds to the case when $K = 2$, while the right panel corresponds to $K = 3$.

$K = 2$ is only exponentially small (e^{-n}) when the rand-index between the true configuration and fitted configuration is close to 1. Apparently, this may appear to impede model selection consistency since the model complexity is exponential in n. However, it turns out that the number of configurations for which the rand-index is sufficiently close to 1 is only polynomial in n. This is also aided by the Dirichlet-Multinomial formulation which restricts $\Pi(z \mid k)/\Pi(z_0 \mid K)$ for configurations close to z_0 to be at most polynomial in n. Hence, the Bayes factor is exponentially small in n delivering an exponential concentration of the posterior of k. This is a clear distinction with standard mixture or regression models (typically polynomial in n in such cases, Rousseau and Mengersen 2011; Drton and Plummer 2017). In the underfitted case, the Rand-Index between the true and the fitted configuration can never be close to 1 which makes separation between the log-marginal likelihoods of the order of n^2. This is strong enough to offset the exponential model complexity as well as the prior ratio leading to a posterior concentration rate of e^{-n^2}.

To empirically demonstrate the posterior probability bounds for $K = 2$ and $K = 3$ in Theorem 2, we conduct a small simulation study under the setup of the theorem. Figures 2 displays $\Pi(k = K \mid \mathcal{A})$ averaged over 100 replicated datasets plotted against n when $K = 2$ and $K = 3$, respectively, and $(p_0, q_0) = (0.5, 0.1)$. It is evident that $\Pi(k = K \mid \mathcal{A})$ approaches 1 at a faster rate for $K = 3$ than for $K = 2$.

5. Simulation Studies

In this section, we investigate the performance of the proposed MFM-SBM approach from a variety of angles. At the very onset, we outline the skeleton of the data-generating process followed throughout this section.

Step 1: Fix the number of nodes n and the true number of communities K.

Step 2: Generate the true clustering configuration $z_0 = (z_{01}, \ldots, z_{0n})$ with $z_{0i} \in \{1, \ldots, K\}$. To this end, we fix the respective community sizes $n_{01}, \ldots, n_{0K}$, and without loss of generality, let $z_{0i} = l$ for all $i = \sum_{j<l} n_{0,j} + 1, \ldots, \sum_{j<l} n_{0,j} + n_{0l}$ and $l = 1, \ldots, K$. We consider both balanced (i.e., $n_{0l} \sim \lfloor n/K \rfloor$ for all l) and unbalanced networks. In the unbalanced case, the community sizes are chosen as $n_{01} : \cdots : n_{0K} = 2 : \cdots : K + 1$.

Step 3: Construct the matrix Q in (1) with $q_{rs} = q + (p - q)I(r = s)$, so that all diagonal entries of Q are p and all off-diagonal entries are q. We fix $q = 0.10$ throughout and vary p subject to $p > 0.10$. Clearly, smaller values of p represent weaker clustering pattern.

Step 4: Generate the edges $A_{ij} \sim \text{Bernoulli}(Q_{z_{0i}z_{0j}})$ independently for $1 \le i < j \le n$.

The Rand index (Rand 1971) is used to measure the accuracy of clustering. Given two partitions $\mathcal{C}_1 = \{X_1, \ldots, X_r\}$ and $\mathcal{C}_2 = \{Y_1, \ldots, Y_s\}$ of $\{1, 2, \ldots, n\}$, let a, b, c, and d, respectively, denote the number of pairs of elements of $\{1, 2, \ldots, n\}$ that are (a) in a same set in $\mathcal{C}_1$ and a same set in $\mathcal{C}_2$, (b) in different sets in $\mathcal{C}_1$ and different sets in $\mathcal{C}_2$, (c) in a same set in $\mathcal{C}_1$ but in different sets in $\mathcal{C}_2$, and (d) in different sets in $\mathcal{C}_1$ and a same set in $\mathcal{C}_2$. The Rand index RI is

$$\text{RI} = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}.$$

Clearly, $0 \le \text{RI} \le 1$ with a higher value indicating a better agreement between the two partitions. In particular, $\text{RI} = 1$ indicates $\mathcal{C}_1$ and $\mathcal{C}_2$ are identical (modulo labeling of the nodes).

We also briefly discuss the estimation of k from the posterior. In our collapsed Gibbs sampler, k is marginalized out and hence we do not directly obtain samples from the posterior distribution of k. However, we can still estimate k based on the posterior distribution of $|z|$, the number of unique values (occupied components) in $(z_1, \ldots, z_n)$. This is asymptotically justified for mixtures of finite mixtures as in sec. 4.3.2 of Miller (2014) who showed that the (prior) posterior distribution of $|z|$ behaves very similarly to that for the number of components k when n is large. This approach also works well in finite samples as demonstrated below.

In all the simulation examples considered below, we employed Algorithm 1 with $\gamma = 1$ and $a = b = 1$ to fit the MFM-SBM model; we shall henceforth refer to this as the MFM-SBM algorithm. For all simulations, a truncated Poisson prior with mean 1 is assumed on k. We arbitrarily initialized our algorithm with nine clusters and randomly allocated the

cluster configurations in all the examples. We experimented with various other choices and did not find any evidence of sensitivity to the initialization; a detailed sensitivity analysis can be found in Appendix C of the supplemental document. In more complex real networks, a practical guideline for the truncated Poisson mean is to take an empirical Bayes approach and set it to the estimated number of clusters from a frequentist algorithm (such as BHM considered in the article).

5.1. Estimation Performance

We now study the accuracy of MFM-SBM in terms of estimating the number of communities as well as the community memberships. As benchmark for comparison, we consider two modularity-based methods available in the R Package `igraph` which first estimate the number of communities by some model selection criterion and subsequently optimize a modularity function to obtain the community allocations. The first competitor, called the leading eigenvector method (LEM; Newman 2006), finds densely connected subgraphs by calculating the leading nonnegative eigenvector of the modularity matrix of the graph. The second competitor, called the hierarchical modularity measure (HMM; Blondel et al. 2008), implements a multi-level modularity optimization algorithm for finding the community structure. Our experiments suggests that these two methods have the overall best performance among available methods in the R Package `igraph`. In addition to LEM and HMM, we also consider a couple of very recent spectral methods which have been developed solely for estimating the number of communities and have been shown to outperform a wide variety of existing approaches based on BIC, cross-validation, etc. These methods are based on the spectral properties of certain graph operators, namely, the nonbacktracking matrix (NBM) and the Bethe Hessian matrix (BHM). We also compare our algorithm to transdimensional MCMC algorithms like reversible jump MCMC or allocation samplers (Nobile and Fearnside 2007) that also allow the number of components to be inferred from data. We found the very recent preprint (Newman and Reinert 2016, ; MH-MCMC) that came out (C code publicly available) while this article was in submission which implements a similar idea to update k using Metropolis–Hastings moves and also uses a Dirichlet-multinomial prior.

We consider balanced networks with 100 nodes and different choices of K and p. We generate 100 independent datasets using the steps outlined at the beginning of the section and compare the different approaches based on the proportion of times the true K is recovered among the 100 replicates. For MFM-SBM, we used random initializations to run 10 MCMC chains in parallel for 250 iterations each, and took majority voting among the posterior modes of k from each chain to arrive at a final point estimate. The summaries from the 100 replicates are provided in Figures 3 and 4.

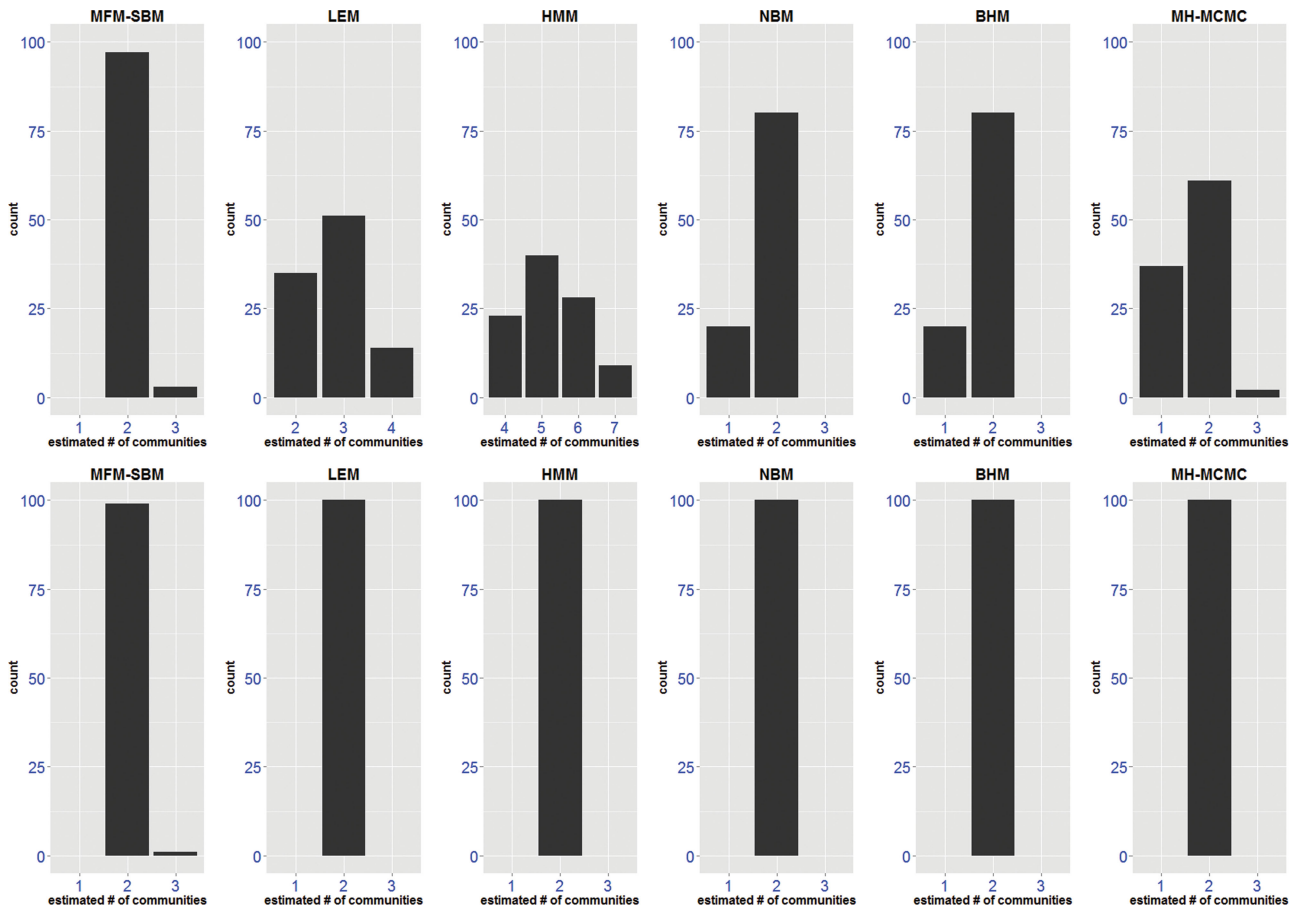

Figure 3. Balanced network with 100 nodes and 2 communities. Histograms of estimated number of communities across 100 replicates. The lower panel is the case when the community structure in the network is prominent ($p = 0.5$); the top panel is for a vague block structure ($p = 0.24$). From left to right: our method (MFM-SBM), leading eigenvector method (LEM), hierarchical modularity measure (HMM), nonbacktracking matrix (NBM), Bethe Hessian matrix (BHM), and MH-MCMC.

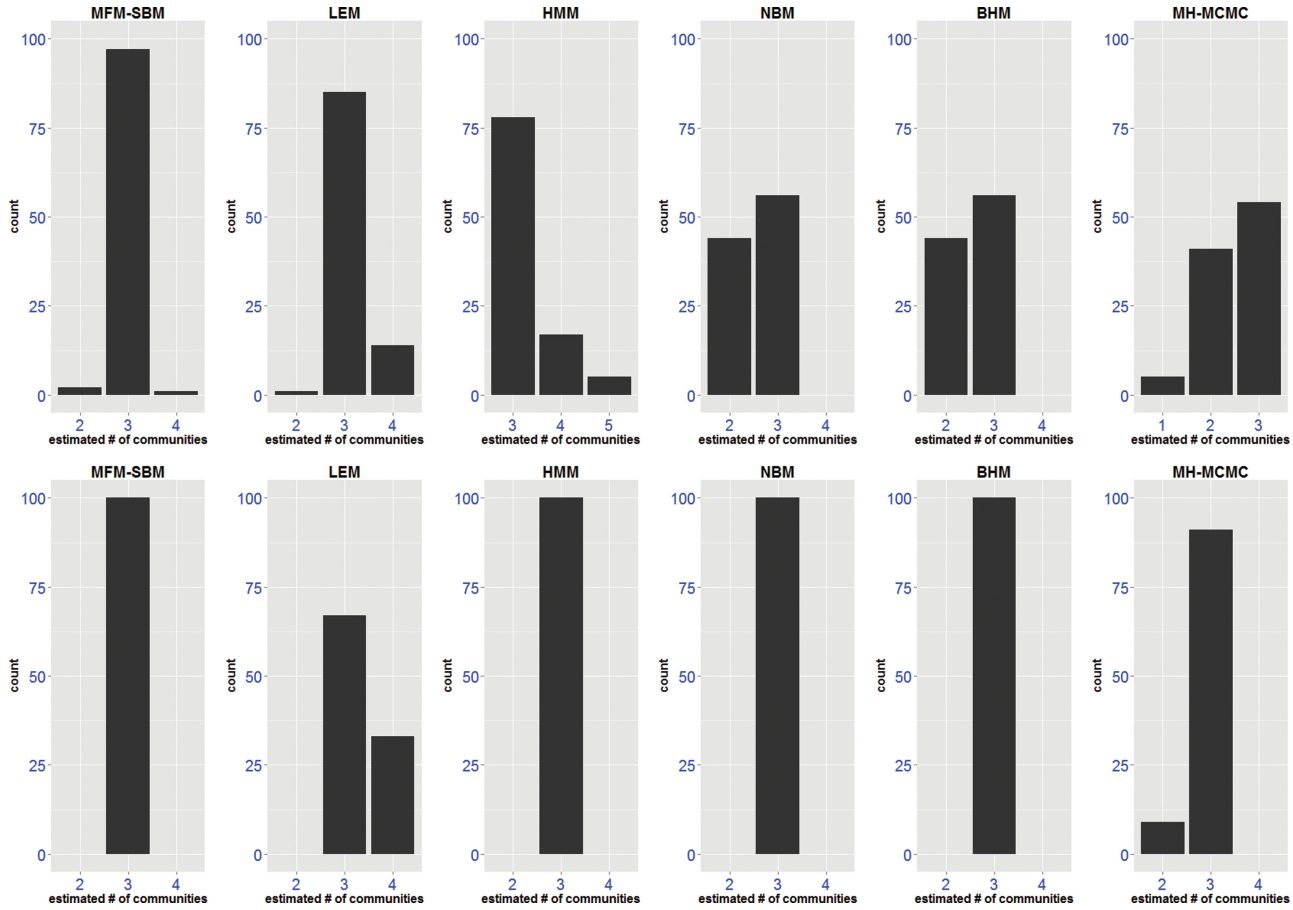

Figure 4. Balanced network with 100 nodes and 3 communities. Histograms of estimated number of communities across 100 replicates. The lower panel is the case when the community structure in the network is prominent ($p = 0.5$); the top panel is for a vague block structure ($p = 0.33$). From left to right: our method (MFM-SBM), leading eigenvector method (LEM), hierarchical modularity measure (HMM), nonbacktracking matrix (NBM), Bethe Hessian matrix (BHM), and MH-MCMC.

From the lower panels of Figures 3 and 4, we can see that when the community structure in the network is prominent ($p = 0.5$), all three methods have 100% accuracy. However, the situation is markedly different when the block structure is vague, as can be seen from the top panels of the respective figures. When the true number of communities is 2 and $p = 0.24$ (top panel of Figure 3), MFM-SBM comprehensively outperforms the competing methods. When $p = 0.33$ with three communities (top panel of Figure 4), our method continues to have the best performance.

We next proceed to compare the estimation performance in recovering the true community memberships using the Rand index as a discrepancy measure. For MFM-SBM, inference on the clustering configurations is obtained employing the modal clustering method of Dahl et al. (2009). Comparisons with LEM, HMM, and MH-MCMC are summarized in Table 1; NBM and BHM are excluded since they only estimate the number of communities. When the block structure is more vague (small p), MFM-SBM provides more accurate estimation of the community memberships.

We also conducted a thorough simulation study to assess robustness of the method to misspecification in Appendix B of the supplemental document.

Table 1. The value outside the parenthesis denotes the proportion of correct estimation of the number of clusters out of 100 replicates. The value inside the parenthesis denotes the average Rand index value when the estimated number of clusters is true. NAs indicate no correct estimation of the number of clusters out of all replicates.

(K, p)	MFM-SBM	LEM	HMM	MH-MCMC
$K = 2, p = 0.50$	0.99 (1.00)	1.00 (0.99)	1.00 (1.00)	1.00 (1.00)
$K = 2, p = 0.24$	0.97 (0.84)	0.35 (0.79)	NA (NA)	0.61 (0.78)
$K = 3, p = 0.50$	1.00 (1.00)	0.67 (0.96)	1.00 (0.99)	0.91 (0.99)
$K = 3, p = 0.33$	0.97 (0.93)	0.85 (0.79)	0.78 (0.89)	0.54 (0.93)

6. Benchmark Real Datasets

We consider two real datasets popularly considered in the literature (i) the dolphin social network data and the (ii) US political books network. Both can be found in *http://www-personal.umich.edu/mejn/netdata/*. We mention the analysis of the first dataset in Section 6.1 and defer the analysis of the second dataset to Appendix D of the supplemental document.

6.1. Community Detection in Dolphin Social Network Data

We consider the social network dataset (Lusseau et al. 2003) obtained from a community of 62 bottlenose dolphins (Tursiops spp.) over a period of seven years from 1994 to 2001. The nodes in the network represent the dolphins, and ties between nodes represent associations between dolphin pairs occurring more often than by random chance. A reference clustering of this undirected network with 62 nodes is in Figure 5 (refer to

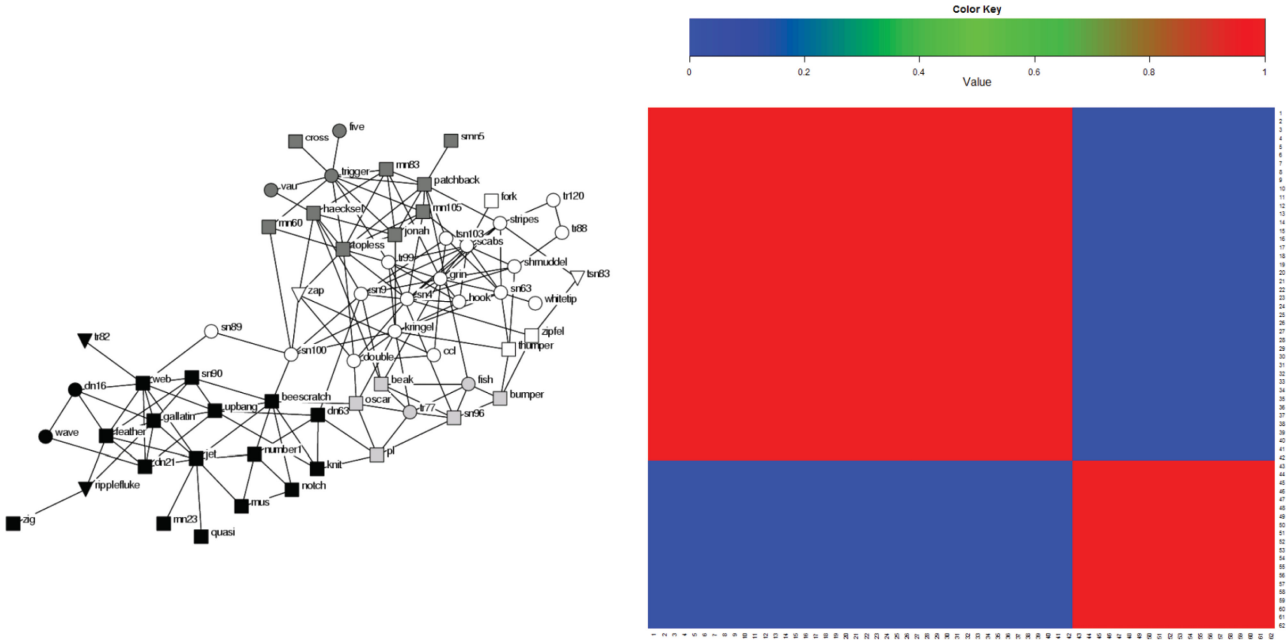

Figure 5. Reference configuration for the dolphin network. Left panel: Vertex color indicates community membership: black and nonblack vertices represent the principal division into two communities. Shades of gray represent sub-communities. Females are represented with circles, males with squares, and individuals with unknown gender with triangles. Right panel: Heatmap of the membership matrix B of the reference configuration z^0 defined as $B_{ij} = \mathbb{1}(z_i^0 = z_j^0)$.

Fig. 1 in Lusseau and Newman 2004). The reference clustering shows several sub-communities based on gender, age, and other demographic characteristics. There are 58 ties between males and males, 46 between females and females, and 44 between males and females, for a total of 159 ties altogether. We are interested in recovering the principal division into two communities as indicated by the black and the nonblack vertices just from the adjacency matrix itself.

Results from our method (MFM-SBM) is based on 10,000 MCMC iterations leaving out a burn-in of 4000, initialized at a randomly generated configuration with nine clusters. The elements of probability matrix Q are assigned independent Beta$(1, 1)$ priors. From Table 2, it is evident that our method (MFM-SBM), NBM, and BHM provide consistent estimate of the number of clusters (being same as the reference clustering), while the other three overestimated the number of clusters.

Table 2. Estimated number of clusters for dolphin data.

Method	MFM-SBM	NBM	BHM	LEM	HMM	MH-MCMC
Number of clusters	2	2	2	5	5	3

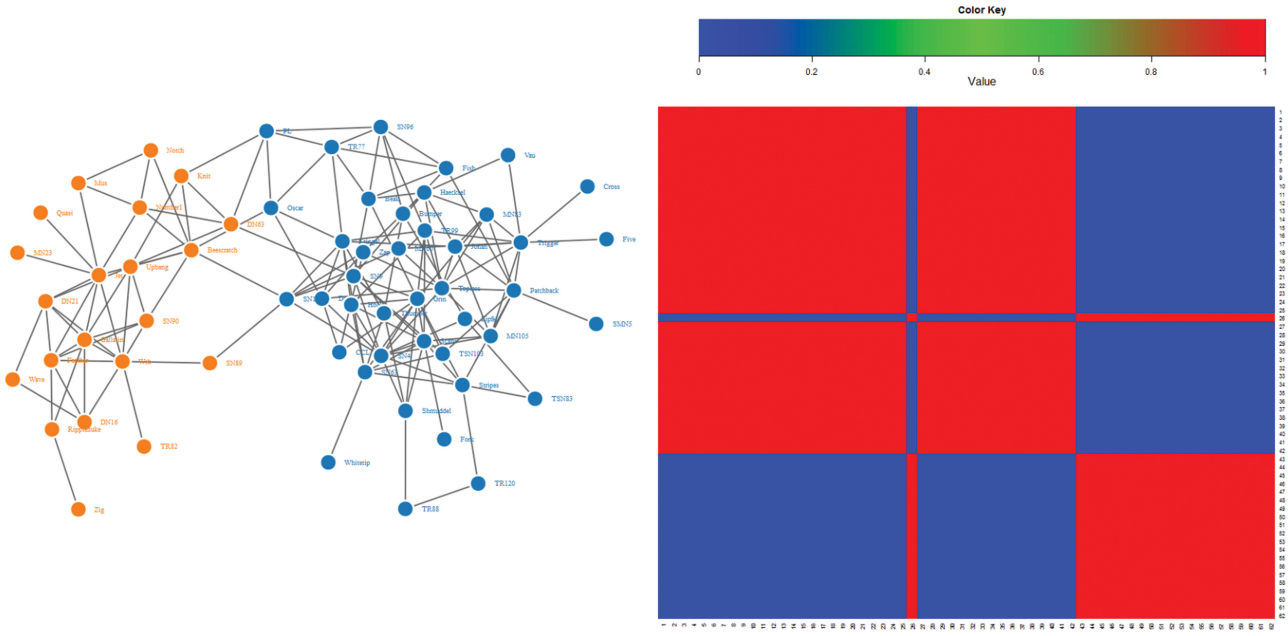

Figure 6. Estimated configuration for the dolphin network using MFM-SBM. Left panel: Vertex color indicates community membership. Right panel: Heatmap of the membership matrix $\hat{B}$ of the estimated configuration $\hat{z}$. Perfect concordance with the reference configuration except for the assignment of the 8th subject.

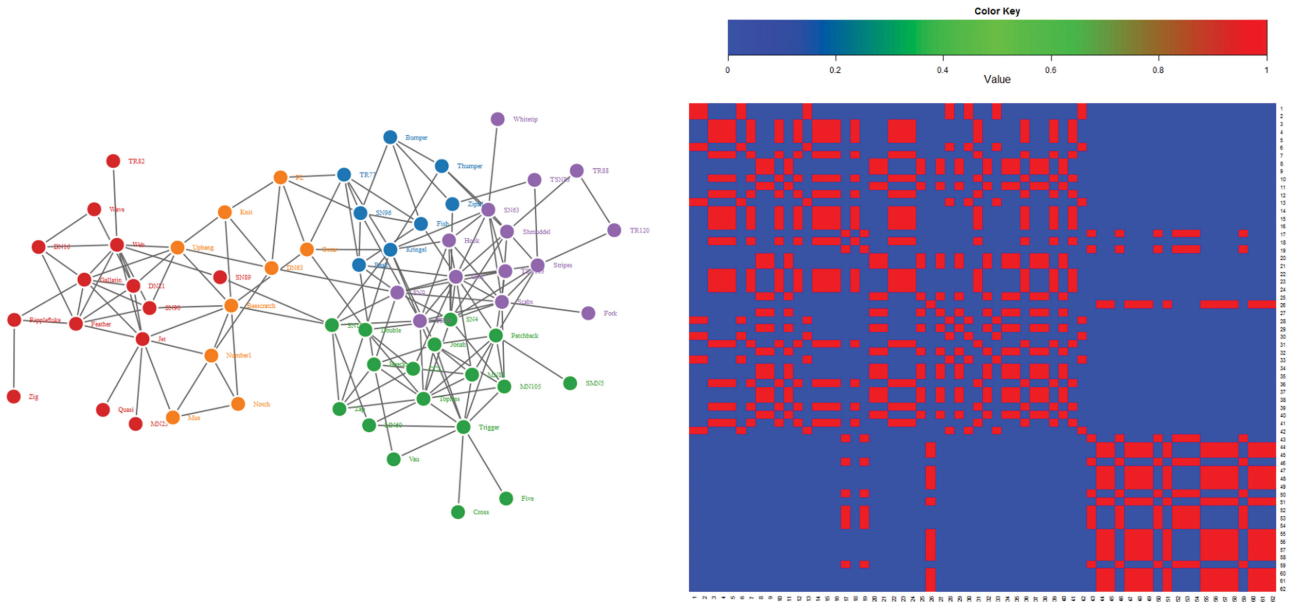

Figure 7. Estimated configuration for the dolphin network using LEM. Left panel: Vertex color indicates community membership. Right panel: Heatmap of the membership matrix $\hat{B}$ of the estimated configuration $\hat{z}$. The number of clusters is estimated to be 4. Aside from cluster splitting, the assignment of three subjects are different from that in reference configuration.

From Figure 6, we see that the estimated configuration from MFM-SBM is very similar to the reference clustering (the only difference is in the assignment of the 8th subject).

The heatmaps in Figures 7 and 8 show both LEM and HMM incur a few misclassified nodes. Figure 9 shows MH-MCMC overestimate the number of clusters, with the larger cluster corresponding to the reference configuration split into two smaller clusters indicating that the mixing of the MCMC has been affected by the transdimensional moves.

7. Discussion

We proposed a Bayesian approach for discovering the number of communities as well as the groups in a network, which has excellent performance in both simulation and real data examples. The contribution of the article is learning the number of communities and the configurations simultaneously in a coherent probabilistic framework. The approach is also proved to yield consistent detection of the number of communities, which is to the best of our knowledge the first such result in a Bayesian paradigm. As an intermediate result, we developed concentration inequalities for nonlinear functions of Bernoulli random variables (refer to Proposition E.1 in the supplemental document) which may be useful in analysis of related network models. The method can be extended easily to numerous modification of stochastic block models including the degree-corrected, mixed membership and the covariate-adjusted versions.

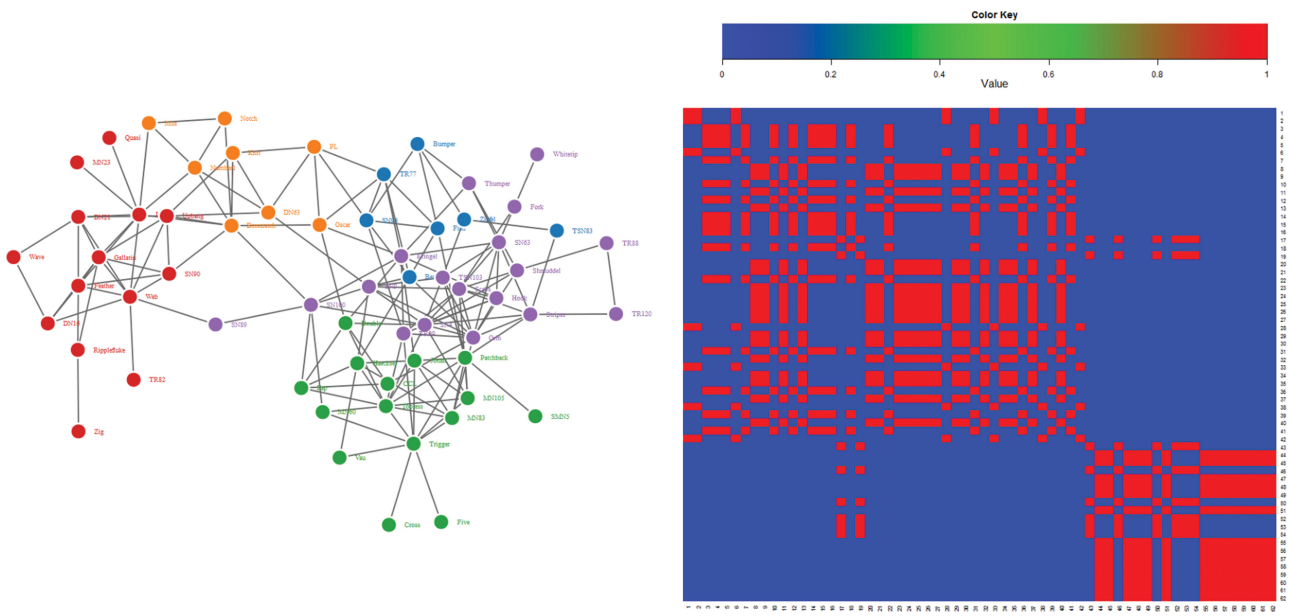

Figure 8. Estimated configuration for the dolphin network using HMM. Left panel: Vertex color indicates community membership. Right panel: Heatmap of the membership matrix $\hat{B}$ of the estimated configuration $\hat{z}$. The number of clusters is estimated to be 4 and the assignment of two subjects are different from that in reference configuration aside from cluster splitting.

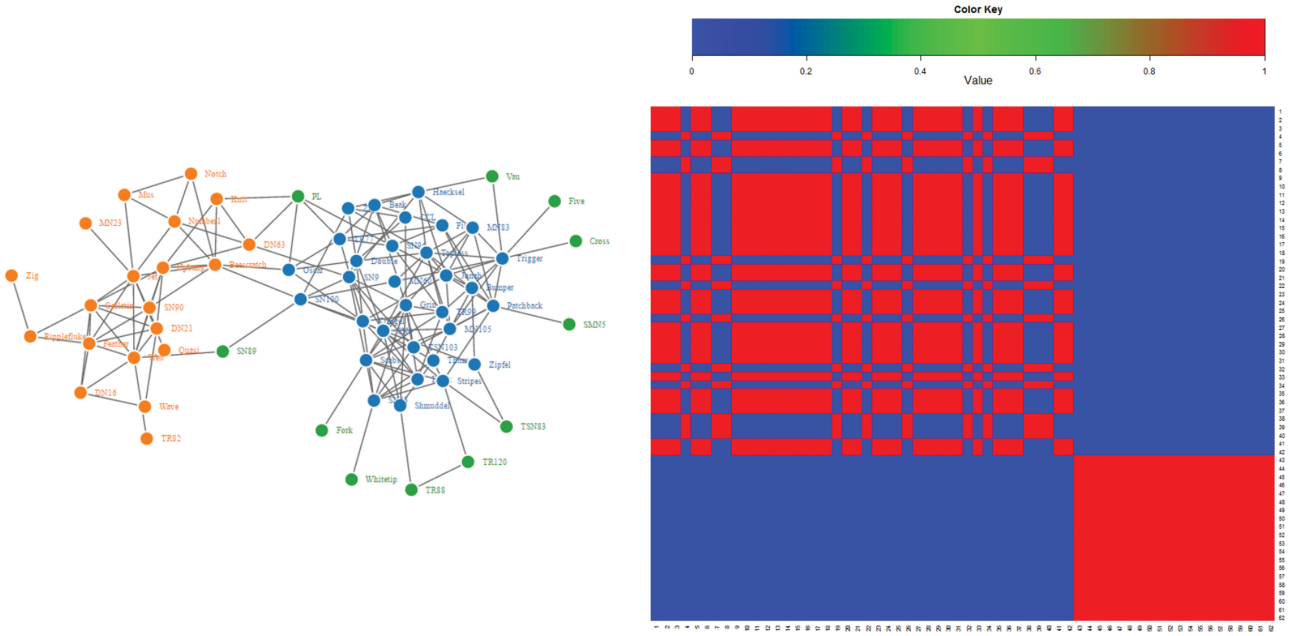

Figure 9. Estimated configuration for the dolphin network using MH-MCMC. Left panel: Vertex color indicates community membership. Right panel: Heatmap of the membership matrix $\hat{B}$ of the estimated configuration $\hat{z}$. The number of clusters is estimated to be 3.

Supplementary Materials

Additional simulations exploring sensitivity, convergence diagnostics, and robustness, and proofs of all technical results, are provided in the supplemental materials. The supplemental material additionally contains a second real data example.

Funding

Dr. Bhattacharya acknowledges NSF CAREER (DMS 1653404), NSF DMS 1613156, and National Cancer Institute's R01 CA158113, and Dr. Pati acknowledges NSF DMS 1613156 for supporting this research.

References

Abbe, E., and Sandon, C. (2015a), "Community Detection in the General Stochastic Block Model: Fundamental Limits and Efficient Algorithms for Recovery," in *Proceedings of 56th Annual IEEE Symposium on Foundations of Computer Science, Berkeley, CA*, pp. 670–688. [896,897]

——— (2015b), "Detection in the Stochastic Block Model with Multiple Clusters: Proof of the Achievability Conjectures, Acyclic BP, and the Information-Computation Gap,"*arXiv preprint arXiv:1512.09080*. [896]

——— (2015c), "Recovering Communities in the General Stochastic Block Model without knowing the Parameters," in *Advances in Neural Information Processing Systems*, Red Hook, NY: Curran Associates, Inc., pp. 676–684. [896]

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [893,894]

Aldous, D. J. (1985), *Exchangeability and Related Topics*, New York: Springer. [895]

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, 2097–2122. [893]

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, 21068–21073. [893,896]

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya urn Schemes," *The Annals of Statistics*, 1, 353–355. [895]

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008), "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008. [900]

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. W. (2015), "Bayesian Linear Regression with Sparse Priors," *The Annals of Statistics*, 43, 1986–2018. [896]

Dahl, D. B. (2009), "Modal Clustering in a Class of Product Partition Models," *Bayesian Analysis*, 4, 243–264. [901]

Daudin, J. J., Picard, F., and Robin, S. (2008), "A Mixture Model for Random Graphs," *Statistics and Computing*, 18, 173–183. [893]

Drton, M., and Plummer, M. (2017), "A Bayesian Information Criterion for Singular Models," *Journal of the Royal Statistical Society*, Series B, 79, 323–380. [899]

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017), "Achieving Optimal Misclassification Proportion in Stochastic Block Model," *The Journal of Machine Learning Research*, 18, 1980–2024. [896]

Goldenberg, A., Zheng, A., Fienberg, S., and Airoldi, E. (2010), "A Survey of Statistical Network Models," *Foundations and Trends® in Machine Learning*, 2, 129–233. [893]

Goldenberg, J., Libai, B., and Muller, E. (2001), "Talk of the Network: A Complex Systems look at the Underlying Process of Word-Of-Mouth," *Marketing Letters*, 12, 211–223. [893]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [894]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [893]

Johnson, V. E., and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660. [896]

Karrer, B., and Newman, M. E. J. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [893,894]

Kruijer, W., Rousseau, J., and Van Der Vaart, A. (2010), "Adaptive Bayesian Density Estimation with Location-Scale Mixtures," *Electronic Journal of Statistics*, 4, 1225–1257. [894]

Latouche, P., Birmele, E., and Ambroise, C. (2012), "Variational Bayesian Inference and Complexity Control for Stochastic Block Models," *Statistical Modelling*, 12, 93–115. [893]

Le, C. M., and Levina, E. (2015), "Estimating the Number of Communities in Networks by Spectral Methods," arXiv:1507.00827. [893,896]

Lusseau, D., and Newman, M. (2004), "Identifying the Role that Animals Play in their Social Networks," *Proceedings of the Royal Society of London B: Biological Sciences*, 271, S477–S481. [902]

Lusseau, D., Schneider, K., Boisseau, O., Haase, P., Slooten, E., and Dawson, S. (2003), "The Bottlenose Dolphin Community of Doubtful Sound features a Large Proportion of Long-Lasting Associations," *Behavioral Ecology and Sociobiology*, 54, 396–405. [901]

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. [894]

Miller, J. W. (2014), "Nonparametric and Variable-Dimension Bayesian Mixture Models: Analysis, Comparison, and New Methods," Ph.D. dissertation, Brown University, Providence, RI. [899]

Miller, J. W., and Harrison, M. T. (2017), "Mixture Models with a Prior on the Number of Components," *Journal of the American Statistical Association*, 113, 340–356. [894,895,896]

Narisetty, N. N., He, X., (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *The Annals of Statistics*, 42, 789–817. [896]

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265. [895]

Newman, M. E. (2004), "Detecting Community Structure in Networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, 38, 321–330. [893]

——— (2006), "Finding Community Structure in Networks using the Eigenvectors of Matrices," *Physical review E*, 74, 036104. [900]

Newman, M. E. J. (2012), "Communities, Modules and Large-Scale Structure in Networks," *Nature Physics*, 8, 25–31. [893]

Newman, M. E., and Girvan, M. (2004), "Finding and Evaluating Community Structure in Networks," *Physical Review E*, 69, 026113. [893]

Newman, M., and Reinert, G. (2016), "Estimating the Number of Communities in a Network," *Physical Review Letters*, 117, 078301. [900]

Nobile, A., and Fearnside, A. T. (2007), "Bayesian Finite Mixtures with an Unknown Number of Components: The Allocation Sampler," *Statistics and Computing*, 17, 147–162. [894,900]

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. [893,894]

Pitman, J. (1995), "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory and Related Fields*, 102, 145–158. [894,895]

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850. [899]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [893]

Rousseau, J., and Mengersen, K. (2011), "Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models," *Journal of the Royal Statistical Society*, Series B, 73, 689–710. [894,899]

Saldana, D. F., Yu, Y., and Feng, Y. (2015), "How Many Communities Are There?" *Journal of Computational and Graphical Statistics*, 26, 171–181. [894]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [895]

Shi, J., and Malik, J. (2000), "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 888–905. [893]

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018), "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-Dimensional Settings," *Statistica Sinica*, 28, 1053–1078. [896]

Snijders, T. A. B., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure," *Journal of classification*, 14, 75–100. [893,894]

Vershynin, R. (2012), "Introduction to the Non-Asymptotic Analysis of Random Matrices," in *Compressed Sensing: Theory and Applications*, eds. Y. C. Eldar and G. Kutyniok, Cambridge, UK: Cambridge University Press. [898]

Wang, Y. X., and Bickel, P. J. (2017), "Likelihood-Based Model Selection for Stochastic Block Models," *The Annals of Statistics*, 45, 500–528. [893]

White, S., and Smyth, P. (2005), "A Spectral Clustering Approach To Finding Communities in Graph," in *SDM* (Vol. 5), SIAM, pp. 76–84. [893]

Zanghi, H., Ambroise, C., and Miele, V. (2008), "Fast Online Graph Clustering via Erdős–Rényi mixture," *Pattern Recognition*, 41, 3592–3599. [894]

Zhang, A. Y., Zhou, H. H., (2016), "Minimax Rates of Community Detection in Stochastic Block Models," *The Annals of Statistics*, 44, 2252–2280. [896,897]

Zhang, S., Wang, R.-S., and Zhang, X.-S. (2007), "Identification of Overlapping Community Structure in Complex Networks using fuzzy c-Means Clustering," *Physica A: Statistical Mechanics and its Applications*, 374, 483–490. [893]

Zhao, Y., Levina, E., and Zhu, J. (2011), "Community Extraction for Social Networks," *Proceedings of the National Academy of Sciences*, 108, 7321–7326. [893]

——— (2012), "Consistency of Community Detection in Networks Under Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 40, 2266–2292. [893,896]