Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–28

Bayesian Community Detection for Networks with Covariates

Luyi Shen*, Arash Amini†, Nathaniel Josephs‡, and Lizhen Lin*

Abstract. The increasing prevalence of network data in a vast variety of fields and the need to extract useful information out of them have spurred fast developments in related models and algorithms. Among the various learning tasks with network data, community detection, the discovery of node clusters or "communities," has arguably received the most attention in the scientific community. In many real-world applications, the network data often come with additional information in the form of node or edge covariates that should ideally be leveraged for inference. In this paper, we add to a limited literature on community detection for networks with covariates by proposing a Bayesian stochastic block model with a covariate-dependent random partition prior. Under our prior, the covariates are explicitly expressed in specifying the prior distribution on the cluster membership. Our model has the flexibility of modeling uncertainties of all the parameter estimates including the community membership. Importantly, and unlike the majority of existing methods, our model has the ability to learn the number of the communities via posterior inference without having to assume it to be known. Our model can be applied to community detection in both dense and sparse networks, with both categorical and continuous covariates, and our MCMC algorithm is very efficient with good mixing properties. We demonstrate the superior performance of our model over existing models in a comprehensive simulation study and an application to two real datasets.

Keywords: community detection, networks with covariates, covariate-dependent random partition prior, Gibbs sampler.

1 Introduction

The ubiquity of network data in modern science and engineering and the need to extract meaningful information out of them has spurred rapid developments in the models, theory, and algorithms for the inference of networks (Erdős and Rényi, 1959; Bickel and Chen, 2009; Wolfe and Olhede, 2013; Kolaczyk, 2009; Lovász, 2012; Kolaczyk et al., 2020). Among the specific learning tasks with network data, community detection, which aims to detect communities or clusters among nodes, has arguably received the most attention in the scientific community. Various models and algorithms have been developed for community detection in networks including modularity-based methods (Newman, 2006), spectral clustering algorithms (Luxburg, 2007; Rohe et al., 2011),

*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, lshen4@nd.edu
†Department Statistics, UCLA, aaamini@ucla.edu
‡Department of Statistics, North Carolina State University, nathaniel.josephs@ncsu.edu
§Department of Mathematics, The University of Maryland, lizhen01@umd.edu

 https://doi.org/10.1214/24-BA1415

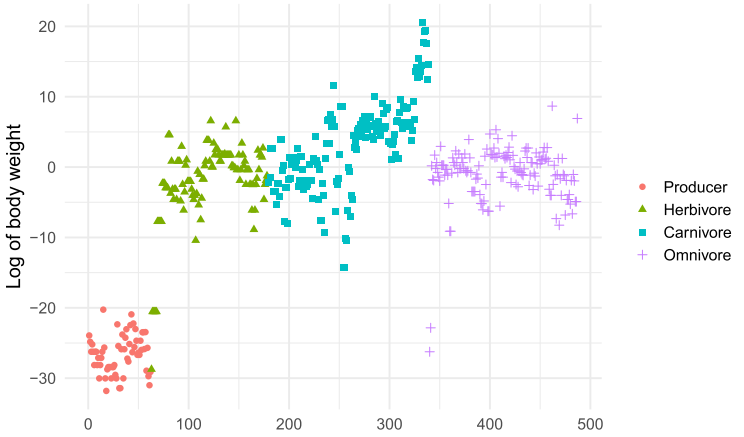

Figure 1: Log body mass for different species colored by their feeding type.

stochastic block models (Holland et al., 1983; Karrer and Newman, 2011; Ball et al., 2011), optimization-based approaches via semidefinite programming (Amini and Levina, 2018), and various Bayesian models (Mørup and Schmidt, 2012; Amini et al., 2019), among others.

Besides the edge information of an observed network, there are often additional covariates or nodal information available in many real-world networks. These additional covariates should be ideally utilized when performing community detection. For example, in a Facebook network, one can obtain from an individual's profile covariates including current city, workplace, hometown, education, and hobbies. Another example is the Weddell Sea trophic network, which describes the marine ecosystem of the Weddell Sea (Jacob et al., 2011). It is a predator-prey network that includes the average adult body mass for each of the species. If one were to only utilize the network information, it is hard to differentiate all the different feeding types. However, the body mass of each species shows a partial clustering by the group, as seen in Figure 1. Therefore, a better clustering should be achievable when both the network and covariates are incorporated.

Such network data have motivated an emerging line of work that aims to deal with community detection problems that leverage both the network and the exogenous covariates. A node-coupled SBM is proposed in Weng and Feng (2022) in which cluster information or the block matrix is uniquely encoded by the covariates. Another model from Zhang et al. (2019) specifies that the link probability between a pair of nodes is contributed additively by the block probability in an SBM and a similarity measure between the covariates of a pair of nodes. A similar class of block models is proposed in Sweet (2015) that also accommodates covariates in an additive way such that the link probability is influenced by both block membership and covariates. A covariate-assisted spectral clustering algorithm is proposed in Binkiewicz et al. (2017) and later modified for degree-corrected block models in Hu and Wang (2022). Categorical covariates on the actor level are included in the model in Tallberg (2004), and the block affiliation

probabilities are modeled conditional on the covariates via a multinomial probit model. Another prominent method in the frequentist literature is due to Zhang et al. (2016) in which a joint community detection criterion is proposed using both the adjacency matrix of the network and the node features, and their algorithm weights the edges according to feature similarities. Recently, the interplay between network information and covariates is investigated in an optimization framework for community detection in sparse networks in Yan and Sarkar (2021). From a Bayesian perspective, there are a few papers that are closely related to our work. We introduce them in Section 2.1 and provide a detailed discussion of their connections to our work (and to each other).

We add to this literature by proposing a Bayesian community detection procedure in which the effects of the covariates are incorporated via a covariate-dependent random partition prior on the node labels of an SBM. The covariates are explicitly expressed and incorporated in the prior probability of generating clusters. One of the distinctive features of our models compared with the ones already proposed in the literature is that ours has the ability to learn the number of communities via posterior inference without having to assume it to be known. The proposed model has the flexibility of assessing uncertainties for all the model parameters through an efficient MCMC algorithm for posterior inference. Note that there are several works in the literature that have employed the idea of a random partition prior or Bayesian nonparametric models for modeling network or relational data. We discuss these comparisons in Section 2.1 after introducing our model.

Our model can be applied in both dense and sparse regimes. In a sparse regime, as one of our simulation studies shows, our model outperforms other state-of-the-art methods such as that of Yan and Sarkar (2021), whose primary goal was to deal with sparse network condition with covariates. We also apply our methods to networks that have covariates with relatively high-dimensions. Our extensive simulations demonstrate our overall superior performance over existing methods in networks with continuous or categorical features, even when those methods are given the true number of communities.

The remainder of our paper is organized as follows. Section 2 is devoted to our model description and MCMC algorithms. In Section 3, we carry out several simulation studies in various settings to demonstrate the utility of our proposed model and algorithms. We also apply our model to two data examples in Section 4. We conclude in Section 5 with possibilities for future work.

2 Prior, model, and MCMC algorithm

Consider an observed network on n nodes represented by an $n \times n$ adjacency matrix $A = (A_{ij})$ with $A_{ij} = 1$ indicating the presence of a link between nodes i and j, and $A_{ij} = 0$ otherwise. Assume in addition that we have some covariate information $x_i \in \mathbb{R}^p$ for each node $i = 1, \ldots, n$. The covariate information associated with the node are often referred to as nodal information or node features of the network, and are frequently encountered in modern network data. We let $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$ denote all of the node covariates of a given network. Our goal is to perform network community detection by incorporating both the network structure and the nodal information. The

key challenge is how to jointly model these two sources of information. Below we propose a Bayesian model that incorporates the nodal information in the prior probability of cluster membership within an SBM.

Let $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathbb{N}^n$ be a node membership vector and $L(\boldsymbol{z}) = \max\{z_i : i \in [n]\}$ indicate the total number of clusters implied by $\boldsymbol{z}$. We do not assume $L(\boldsymbol{z})$ to be known *a priori*. Let $S_\ell(\boldsymbol{z}) = \{i \in [n] : \ z_i = \ell\}$ be the set of indices of nodes belonging to the ℓ^{th} cluster according to $\boldsymbol{z}$. For any subset $S \subseteq [n]$ and $\boldsymbol{x} = (x_1, \cdots, x_n)^T \in \ \mathbb{R}^{n \times p}$, let $g(S \,|\, \boldsymbol{x})$ be a nonnegative function that measures the homogeneity of the covariates $\{x_i, \ i \in S\}$. That is, $g(S \,|\, \boldsymbol{x})$ takes larger values when all of the x_i with $i \in S$ are more similar. One can think of S as a potential cluster of nodes and $g(S \,|\, \boldsymbol{x})$ as a measure of the quality of such cluster, with regards the nodal information

Inspired by Müller and Quintana (2010); Park and Dunson (2010); Müller et al. (2011), we consider the following covariate-dependent random partition model:

$$p(\boldsymbol{z} \,|\, \boldsymbol{x}) \propto \prod_{\ell=1}^{L(\boldsymbol{z})} g\big(S_\ell(\boldsymbol{z}) \,|\, \boldsymbol{x}\big) \cdot c\big(S_\ell(\boldsymbol{z})\big) \ . \tag{1}$$

The non-negative function $S \mapsto c(S)$ is known as the cohesion function of the product partition probability model. In a random partition model based on the Dirichlet process, with baseline probability measure G_0 and concentration parameter α, one has $c(S) \ = \ \alpha(|S| - 1)!$ (Ferguson, 1973; Sethuraman, 1994).

Borrowing from Müller et al. (2011), we define $g(S \,|\, \boldsymbol{x})$ based on an auxiliary probability model $q(\cdot \,|\, \cdot)$, where

$$g(S \,|\, \boldsymbol{x}) = \int \prod_{i \in S} q(x_i \,|\, \xi) \, \nu(\xi) \, d\xi \ . \tag{2}$$

Note that the covariates $\boldsymbol{x}$ are not random. The term $\prod_{i \in S} q(x_i \,|\, \xi)$ measures the effect or contribution of the covariates on the prior probability of cluster S. One can choose $q(x_i \,|\, \xi)$ and $\nu(\xi)$ as a conjugate pair to facilitate the analytic evaluation of $g(S \,|\, \boldsymbol{x})$. For the cohesion function, we adopt $c(S) = \alpha(|S| - 1)!$.

Combining equations (1) and (2), we have

$$p(\boldsymbol{z} \,|\, \boldsymbol{x}) \propto \prod_{\ell=1}^{L(\boldsymbol{z})} \left[\int \prod_{i \, \in \, S_\ell(\boldsymbol{z})} q(x_i \,|\, \xi_\ell) \, d\nu(\xi_\ell) \right] c\big(S_\ell(\boldsymbol{z})\big) \ . \tag{3}$$

In this model, ξ_ℓ can be considered the center of the nodal covariates in cluster ℓ, and $q(x_i \,|\, \xi_\ell)$ a measure of how far the covariates in cluster S_ℓ are from its center ξ_ℓ. The model then averages over all possible centers $\xi_\ell \sim \nu$.

The distribution in (3) can be written as the marginal of

$$p(\boldsymbol{z}, \boldsymbol{\xi} \,|\, \boldsymbol{x}) \propto \prod_{\ell=1}^{L(\boldsymbol{z})} \Big[c\big(S_\ell(\boldsymbol{z})\big) \prod_{i \, \in \, S_\ell(\boldsymbol{z})} q(x_i \mid \xi_\ell) \, \nu(\xi_\ell) \Big] \cdot \prod_{\ell=L(\boldsymbol{z})+1}^{\infty} \nu(\xi_\ell) \ , \tag{4}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots)$. One can use (4) to derive a Gibbs sampler for sampling the prior. To simplify the notation, we let

$$L = L(\boldsymbol{z}_{-i}) \quad \text{and} \quad S_\ell = S_\ell(\boldsymbol{z}_{-i}) \tag{5}$$

for $\ell \in [L]$ denote the number of clusters of $\boldsymbol{z}_{-i} = (z_j, j \neq i)$ and the clusters themselves. Let

$$\psi_k := \begin{cases} |S_k| & k \in [L] \\ \alpha & k = L+1 \end{cases} . \tag{6}$$

One can show that for $k \in [L+1]$,

$$p(z_i = k,\, \xi_{L+1} \,|\, \boldsymbol{z}_{-i},\, \boldsymbol{\xi}_{1:L}, \boldsymbol{x}) \;\propto\; \nu(\xi_{L+1}) \cdot \psi_k\, q(x_i \,|\, \xi_k) \;\; , \tag{7}$$

where $\boldsymbol{\xi}_{1:L} = (\xi_1, \dots, \xi_L)$. This is equivalent to first drawing $\xi_{L+1} \sim \nu(\cdot)$, and then drawing z_i as follows:

$$p(z_i = k \,|\, \boldsymbol{z}_{-i},\, \boldsymbol{\xi}_{1:L+1}, \boldsymbol{x}) \;\propto\; \psi_k\, q(x_i \,|\, \xi_k) \;\; . \tag{8}$$

Example. *For continuous features, we can take*

$$q(x \,|\, \xi) = N(x; \xi, s^2 I) \quad \textit{and} \quad \nu = N(0, \tau^2 I) \;\; , \tag{9}$$

where $N(x; \xi, s^2 I)$ denotes the density of a normal distribution with mean ξ and covariance matrix $s^2 I$, evaluated at x. Then, $k \mapsto q(x_i \,|\, \xi_k)$ in (8) will be proportional to $\exp(-\|x_i - \xi_k\|^2 / 2s^2)$, which shows that if ξ_ℓ is the closest to x_i among $\{\xi_k\}$, then the inclusion of the covariate information increases the chance of assigning z_i to cluster ℓ.

Example. *For categorical covariates, one can choose $q(x_i \,|\, \xi_k)$ to be a multinomial distribution and $\nu(\cdot)$ to be a Dirichlet distribution. Suppose there are R categorical features and the rth feature has a_r categories for $r = 1, \dots, R$. Then $x_i = (x_{i1}, \dots, x_{iR})$, where x_{ir} is the r-th feature of node i, and $x_{ir} \in \{1, \dots, a_r\}$. Each ξ_k collects parameters of R multinomial vectors, that is, $\xi_k = (\xi_{rk})_{r=1}^R$, where the coordinates $\xi_{rk} = (\xi_{rk}^1, \xi_{rk}^2, \cdots, \xi_{rk}^{a_r})$ are independent draws from $Dir(\gamma \mathbf{1}_{a_r})$. We have*

$$q(x_i \mid \xi_k) = \prod_{r=1}^{R} \prod_{c=1}^{a_r} (\xi_{rk}^c)^{1\{x_{ir}=c\}} \quad \textit{and} \quad \nu = \prod_{i=1}^{R} \mathsf{Dir}(\gamma \mathbf{1}_{a_r}) \;\; . \tag{10}$$

We usually take $\gamma = 1$.

An alternative approach to sample from the prior is to perform Gibbs sampling on the marginalized distribution (3). This leads to the following updates. For each $k \in [L+1]$,

$$p(z_i = k \,|\, \boldsymbol{z}_{-i},\, \boldsymbol{x}) \;\propto\; \psi_k\, \frac{g(S_k \cup \{i\} \,|\, \boldsymbol{x})}{g(S_k \,|\, \boldsymbol{x})} \;\; , \tag{11}$$

where $S_{L+1} = \emptyset$ and $g(\emptyset \,|\, \boldsymbol{x}) = 1$. This approach is, in particular, useful when $q(\cdot \,|\, \cdot)$ and $\nu(\cdot)$ are conjugate so that $g(S \,|\, \boldsymbol{x})$ is easy to compute.

With the priors thus defined, the network is assumed to follow a SBM, that is,

$$p(A \,|\, \boldsymbol{\eta}, \boldsymbol{z}) = \prod_{1 \le i < j \le n} \eta_{z_i, z_j}^{A_{ij}} (1 - \eta_{z_i, z_j})^{1 - A_{ij}} \quad , \tag{12}$$

where $\boldsymbol{\eta} = (\eta_{k,\ell})$ is the connectivity matrix of SBM, with $\eta_{k,\ell}$ representing the link probability between nodes in clusters k and ℓ.

It is possible to obtain closed forms for the full conditional distributions of the unknown model parameters $\boldsymbol{\eta}, \boldsymbol{z}$, and $\boldsymbol{\xi}$ with appropriate choices of $q(\cdot)$ and $\nu(\cdot)$, as demonstrated in Section 2.2. We note that since the prior random partition model (1) puts mass on all potential partitions of the n nodes, the posterior distribution $p(\boldsymbol{z} \,|\, A, \boldsymbol{x})$ also puts mass on all such partitions; however, the posterior will be concentrated around certain partition(s), hence a posteriori, there is a most likely value of $L(\boldsymbol{z})$, the number of communities in $\boldsymbol{z}$. That is how the model learns the number of communities.

2.1 Comparison with literature

There are several works in the literature similar to our model that also use a Bayesian nonparametric approach for tasks related to node clustering.

A pioneering work in the area is that of Kemp et al. (2006). Motivated by the complex system of relations underlying semantic knowledge, Kemp et al. (2006) propose the infinite relational model (IRM) for discovering and clustering underlying structure in relational data sets. In this framework, the observed data are assumed from n types (people, demographic features, answers to a personality test, etc.) and m relationships (person i likes person j, feature x causes answer y, etc). The motivating example given in Kemp et al. (2006) is that of clustering people, represented by set T^1, based on social predicates, represented by set T^2. The observed data is the tensor $T^1 \times T^1 \times T^2 \mapsto \{0, 1\}$ whose (i, j, p) entry determines whether persons i and j have social predicate type p. The idea is to simultaneously cluster T^1 and T^2 so that the tensor is roughly constant within the resulting clusters. In modern language, the model proposed by Kemp et al. (2006) is the so-called *tensor SBM* (Kim et al., 2017; Wang and Zeng, 2019; Lei et al., 2020) but with a CRP prior on the labels of each dimension to allow for infinite clusters a priori. IRM was later explicitly adopted in Mørup and Schmidt (2012) for community detection in network data, as opposed to relational data. IRM is quite flexible and can, for example, be used to incorporate an attribute or feature taking finite values, by taking T^2 above to be the levels of that attribute. However, this attribute should be interpreted as an edge feature, and moreover, IRM needs to have observations on the connectivity of persons (i, j) for all possible levels of this attribute. Adding each feature then requires increasing the dimension of the tensor by one, and demanding lots of observations which are not available in practice in network problems.

More recently, nonparametric Bayesian network models that accommodate nodal covariates have been considered, but mostly with the mixed membership SBM (MMSBM)

framework of Airoldi et al. (2008). For example, Kim et al. (2012) introduced the nonparametric metadata dependent relational (NMDR) model, that essentially couples the MMSBM likelihood with node-covariate-dependent prior on cluster labels. More specifically, they assume latent community vectors η_k that interact with node features ϕ_i to produce a score v_{ki} that determines how likely node i belongs to community k, a priori. They assume that v_{ki} are normal with mean $\langle \eta_k, \phi_i \rangle$ and then translate these real-valued affinities to probabilities π_{ki} via a logistic-stick breaking process (Ren et al., 2011). The $\pi_i = (\pi_{ki})$ then determine the edge probabilities via $\mathbb{E}[A_{ij} \,|\, \pi_i, \pi_j] = \pi_i^T W \pi_j$ where W is the connectivity matrix.

Along the same lines, Zhao et al. (2017) extends the edge partition model (EPM) of Zhou (2015) to incorporate binary node features. The EPM has similarities to MMSBM with novel uses of a Bernoulli-Poisson likelihood coupled with a nonparametric partition model. More specifically, the latent Poisson component $X = (X_{ij})$ still follows a MMSBM decomposition with $\mathbb{E}[X_{ij} \,|\, \phi_i, \phi_j] = \phi_i^T \Lambda \phi_j$ where ϕ_i are the soft community assignments and Λ the connectivity matrix. Similar to Kim et al. (2012), the nodal covariate information is incorporated in constructing a prior on $\phi_i = (\phi_{ik})$. The prior assumes ϕ_i to be drawn from a Gamma distribution with mean $\mathbb{E}[\phi_{ik}] = c_i b_k \prod_{\ell=1}^{L} h_{\ell k}^{f_{i\ell}}$ where $f_{i\ell}$ is the ℓth binary feature of node i. Note that by introducing $\eta_k := (\log h_{\ell k})$ and $f_i = (f_{i\ell})$, one can write $\mathbb{E}[\phi_{ik}] = c_i b_k \exp(\langle \eta_k, f_i \rangle)$, showing that essentially the same inner product interaction of feature and latent community vectors as in Kim et al. (2012) is used by Zhao et al. (2017).

As in Kim et al. (2012) and Zhao et al. (2017), our model also incorporates node features into the partition prior; however, we are modeling hard community assignments rather than soft assignment vectors, making the problem somewhat more challenging. More importantly, our approach allows for a more general dependence of the partition on the features via a kernel $q(x_i \,|\, \xi)$, compared to the simple inner product approach used in both Kim et al. (2012) and Zhao et al. (2017). Our approach is not limited to binary features and by incorporating more complexity into $q(x_i \,|\, \xi)$, we can potentially model more complex feature/community interactions in the prior.

The last closely related work that we discuss is that of Newman and Clauset (2016). They consider node features (metadata) that take values in a finite discrete set (say $\mathbb{X}$). Similar to our work, the node metadata is used to influence the prior on the community assignments. In our notation, they assume $p(z_i \,|\, x_i) = \gamma_{z_i, x_i}$ where $\gamma = (\gamma_{k,x}) \in [0,1]^{K \times |\mathbb{X}|}$ is a parameter matrix to be estimated form the data. Compared to the inner product model of Kim et al. (2012) and Zhao et al. (2017), this approach gives a more flexible model for the interaction of the communities and features. The drawback is that it is limited to discrete features and if $|\mathbb{X}|$ is large, there is potential for over-fitting without further regularization of the γ matrix. Newman and Clauset (2016) use an EM algorithm to estimate the parameters, and they assume the number of communities to be known.

2.2 Gibbs sampler

We now derive a Gibbs sampler to sample from the complete posterior distribution of $(\boldsymbol{z}, \boldsymbol{\xi}, \boldsymbol{\eta})$ given A and $\boldsymbol{x}$. The main challenge is deriving the updates for $\boldsymbol{z}$.

We sample from $\boldsymbol{z}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$ through their full conditional distributions, which are given bellow, until reaching convergence, and then obtain a sample of adequate size of the posterior distribution for inference.

Initialization

We initialize the labels by drawing from a Chinese Restaurant Process (CRP),

$$\boldsymbol{z} \sim \mathsf{CRP}(\alpha) \ .$$

A CRP can be seen as a special case of our prior without any covariates. This follows from (11) by setting $g(S \,|\, \boldsymbol{x}) = 1$. Once $\boldsymbol{z}$ is initialized, all the other parameters can be initialized by the Gibbs updates derived below.

Note that initializing the chain by sampling $\boldsymbol{z}$ from a CRP provides a random start without having to specify the number of the communities K. In many algorithms, spectral clustering is often used to initialize $\boldsymbol{z}$. For a Bayesian model, this is not a natural choice. Moreover, it requires the knowledge or an estimate of K.

Sampling $\boldsymbol{z}$

Let $b(x; a, b) = x^{a-1}(1-x)^{b-1}$ and for simplicity, define

$$\widetilde{c}_\ell(S) := c(S) \prod_{j \in S} q(x_j \,|\, \xi_\ell)\, \nu(\xi_\ell) \ . \tag{13}$$

Note that $\widetilde{c}_\ell(S)$ implicitly depends on ξ_ℓ. We have

$$\begin{aligned} p(A, \boldsymbol{z}, \boldsymbol{\xi}, \boldsymbol{\eta} \mid \boldsymbol{x}) &= p(A \mid \boldsymbol{\eta}, \boldsymbol{z}) \cdot p(\boldsymbol{z}, \boldsymbol{\xi} \mid \boldsymbol{x}) \cdot p(\boldsymbol{\eta}) \\ &\propto \prod_{1 \le i < j \le n} \eta_{z_i z_j}^{A_{ij}} (1 - \eta_{z_i z_j})^{1 - A_{ij}} \prod_{\ell=1}^{\infty} \widetilde{c}_\ell(S'_\ell) \prod_{1 \le m \le \ell \le \infty} b(\eta_{\ell m}; \beta, \beta) \ . \end{aligned}$$

Here, $S'_\ell = \{i \in [n] : \ z_i = \ell\}$ is the ℓth community of $\boldsymbol{z}$. We assume that $\boldsymbol{z}$ has L' communities $S'_1, S'_2, \ldots, S'_{L'}$ and let $S'_{L'+1} = S'_{L'+2} = \cdots = \emptyset$. The convention is that $c(\emptyset) = 1$ while $c(S) = \alpha\Gamma(|S|)$ when S is nonempty. Similarly, $\prod_{j \in \emptyset}(\cdots) = 1$. In the above, we assume that $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots)$ collects all possible ξ_ℓ and similarly for $\boldsymbol{\eta} = (\eta_{\ell m} : \ell, m \in \mathbb{N})$.

Fix i and let $S_\ell = \{j \in [n] \setminus i : z_j = \ell\}$ be the ℓth community of $\boldsymbol{z}_{-i}$. We assume that $\boldsymbol{z}_{-i}$ has L communities $S_1, S_2, \ldots, S_L$ and by convention, let $S_{L+1} = S_{L+2} = \cdots = \emptyset$. To get the communities of $\boldsymbol{z}$ from $\boldsymbol{z}_{-i}$, we either update S_k to $S'_k = S_k \cup \{i\}$ for some $k \in [L]$, or update S_{L+1} to $S'_{L+1} = S_{L+1} \cup \{i\} = \{i\}$, generating a new community.

With the convention we use, we can compactly write both cases as $S'_k = S_k \cup \{i\}$ for all $k \in [L+1]$.

For any $k \in [L+1]$, we obtain

$$p\big(z_i = k,\, \boldsymbol{\xi}_{L+1},\, \boldsymbol{\eta}_{L+1,[L]} \,|\, \boldsymbol{z}_{-i},\, A,\, \boldsymbol{\xi}_{[L]},\, \boldsymbol{\eta}_{[L],[L]},\, \boldsymbol{x}\big) \propto \prod_{j:j\neq i} \eta_{k,z_j}^{A_{ij}} (1-\eta_{k,z_j})^{1-A_{ij}} \cdot \frac{\widetilde{c}_k(S_k \cup \{i\})}{\widetilde{c}_k(S_k)} \prod_{\ell=1}^{L+1} \widetilde{c}_\ell(S_\ell) \prod_{1\le m\le \ell\le L+1} b(\eta_{\ell m};\beta,\beta) \ . \tag{14}$$

Letting

$$O_{i\ell} = \sum_{j:j\neq i} A_{ij} 1\{z_j = \ell\}, \quad n_\ell = \sum_{j:j\neq i} 1\{z_j = \ell\} \ , \tag{15}$$

we have $\prod_{j:j\neq i} \eta_{k,z_j}^{A_{ij}} (1-\eta_{k,z_j})^{1-A_{ij}} = \prod_{\ell=1}^{L} \eta_{k\ell}^{O_{i\ell}} (1-\eta_{k\ell})^{n_\ell - O_{i\ell}}$.

Noting that $\prod_{\ell=1}^{L} \widetilde{c}_\ell(S_\ell)$ is a constant in (14), and similarly for any term in

$$\prod_{1\le m\le \ell\le L+1} b(\pi_{\ell m};\beta,\beta)$$

that does not have an index equal to $L+1$, we obtain

$$p\big(z_i = k,\, \boldsymbol{\xi}_{L+1},\, \boldsymbol{\eta}_{L+1,[L]} \,|\, \boldsymbol{z}_{-i},\, A,\, \boldsymbol{\xi}_{[L]},\, \boldsymbol{\eta}_{[L],[L]},\, \boldsymbol{x}\big) \propto \prod_{\ell=1}^{L} \eta_{k\ell}^{O_{i\ell}} (1-\eta_{k\ell})^{n_\ell - O_{i\ell}} \cdot \frac{\widetilde{c}_k(S_k \cup \{i\})}{\widetilde{c}_k(S_k)} \widetilde{c}_{L+1}(S_{L+1}) \prod_{m=1}^{L} b(\eta_{L+1,m};\beta,\beta) \ . \tag{16}$$

The product over m runs up to L since only $\eta_{L+1,[L]}$ is a variable while $\eta_{L+1,L+1}$ is a constant. This is because z_i can take the new value $L+1$ but z_j with $j \neq i$ takes values in $[L]$, hence we do not need to sample $\eta_{L+1,L+1}$ at this stage.

Since $S_{L+1} = \emptyset$, we have

$$\begin{aligned} \widetilde{c}_{L+1}(S_{L+1}) &= \nu(\xi_{L+1}) \ , \\ \widetilde{c}_{L+1}(S_{L+1} \cup \{i\}) &= \alpha\Gamma(1)\, q(x_i \,|\, \xi_{L+1})\, \nu(\xi_{L+1}) \ . \end{aligned}$$

Hence for $k \in [L+1]$,

$$\frac{\widetilde{c}_k(S_k \cup \{i\})}{\widetilde{c}_k(S_k)} \widetilde{c}_{L+1}(S_{L+1}) = \nu(\xi_{L+1})\, \psi_k\, q(x_i \mid \xi_k) \ ,$$

where ψ_k is defined in (6). Thus, we can compactly write

$$p\big(z_i = k,\, \boldsymbol{\xi}_{L+1},\, \boldsymbol{\eta}_{L+1,[L]} \,|\, \boldsymbol{z}_{-i},\, A,\, \boldsymbol{\xi}_{[L]},\, \boldsymbol{\eta}_{[L],[L]},\, \boldsymbol{x}\big) \propto \nu(\xi_{L+1}) \prod_{m=1}^{L} b(\eta_{L+1,m};\beta,\beta) \cdot \psi_k q(x_i \,|\, \xi_k) \prod_{\ell=1}^{L} \eta_{k\ell}^{O_{i\ell}} (1-\eta_{k\ell})^{n_\ell - O_{i\ell}} \ . \tag{17}$$

This is equivalent to the following. First draw $\xi_{L+1} \sim \nu$ and $\eta_{L+1,m} \sim \text{Beta}(\beta, \beta)$ for $m \in [L]$, all independently. Then draw z_i from

$$p\big(z_i = k \,|\, \boldsymbol{z}_{-i},\, A,\, \boldsymbol{\xi}_{[L+1]}, \boldsymbol{\eta}_{[L+1],[L]}, \boldsymbol{x}\big) \;\propto \\ \psi_k \, q(x_i \mid \xi_k) \prod_{\ell=1}^{L} \eta_{k\ell}^{O_{i\ell}} (1-\eta_{k\ell})^{n_\ell - O_{i\ell}}, \quad k \in [L+1] \;. \tag{18}$$

For continuous features, we use (9) for $q(\cdot \,|\, \cdot)$ and ν, and for categorical variables we use (10) in the above updates.

Remark. *Note that update (18) is where a potentially new community (labeled $L+1$) is created. This happens if the following conditions are met: (a) when sampling z_i according to (18), we happen to pick $z_i = L+1$, (b) $L' = L$, that is, the current number of communities is L, and (c) currently z_i is not assigned to a singlton community. In such a case, the number of communities will increase from $L' = L$ to $L+1$. On the other hand, update (18) can also annihilate a community if the following holds: (a) When sampling z_i, we pick $z_i \in [L]$ and (b) $L' = L+1$ which means that z_i is currently assigned to a singleton community. In this case, the new number of communities will go from L' to $L'-1 = L$.*

Sampling $\boldsymbol{\xi}$

We have

$$p(\boldsymbol{\xi} \,|\, A, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\eta}) = p(\boldsymbol{\xi} \mid \boldsymbol{z}, \boldsymbol{x}) \propto \prod_{\ell=1}^{L'} H_{S'_\ell}(\xi_\ell) \;,$$

where H_S is the distribution with density

$$H_S(\xi) \propto \prod_{i \in S} q(x_i \,|\, \xi)\, \nu(\xi) \;. \tag{19}$$

That is, ξ_ℓ are independent draws from $H_{S'_\ell}$. We recall that $S'_\ell = \{i \in [n] :\; z_i = \ell\}$.

The details of sampling $\boldsymbol{\xi}$ are slightly different given different choices of $q(\cdot)$ and $\nu(\cdot)$ depending on whether continuous or categorical features are available. For continuous features with the Gaussian choice (9), $H_S(\xi) \propto \prod_{i \in S} N(x_i; \xi, s^2 I) \cdot N(\xi; 0, \tau^2 I)$ which gives

$$H_S = N\left(\frac{\tau^2 \sum_{i \in S} x_i}{|S|\tau^2 + s^2}, \frac{s^2 \tau^2}{|S|\tau^2 + s^2} I \right) \;.$$

For the categorical features with the choice (10), we have

$$H_S(\xi) \propto \prod_{i \in S} \prod_{r=1}^{R} \prod_{c=1}^{a_r} (\xi_r^c)^{1\{x_{ir}=c\}} \cdot \prod_{r=1}^{R} \prod_{c=1}^{a_r} (\xi_r^c)^{\gamma - 1} = \prod_{r=1}^{R} \prod_{c=1}^{a_r} (\xi_r^c)^{\alpha_r^c(S) - 1} \;,$$

where $\alpha_r^c(S) := \gamma + \sum_{i \in S} 1\{x_{ir} = c\}$. That is, H_S is the product of Dirichlet distributions,

$$H_S = \prod_{r=1}^{R} \mathsf{Dir}(\alpha_r^1(S), \ldots, \alpha_r^{a_r}(S)) \ .$$

Sampling $\boldsymbol{\eta}$

Let us define the index sets

$$\Gamma_{k\ell} = \begin{cases} \{(i,j) : 1 \le i < j \le n\} & k = \ell \\ \{(i,j) : 1 \le i \ne j \le n\} & k \ne \ell \end{cases} , \tag{20}$$

and block counts

$$M_{k\ell} = \sum_{(i,j) \in \Gamma_{k\ell}} A_{ij} 1\{z_i = k, z_j = \ell\}, \quad N_{k\ell} = \sum_{(i,j) \in \Gamma_{k\ell}} 1\{z_i = k, z_j = \ell\} \ . \tag{21}$$

Then, we have

$$p(\boldsymbol{\eta} \mid A, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\xi}) = p(\boldsymbol{\eta} \mid A, \boldsymbol{z}) \propto \prod_{k \le \ell} \eta_{k\ell}^{M_{k\ell}+\beta-1} (1 - \eta_{k\ell})^{N_{k\ell} - M_{k\ell} + \beta - 1} \ .$$

Thus, $\eta_{k\ell}$ are independent draws from $\mathsf{Beta}(M_{k\ell} + \beta, N_{k\ell} - M_{k\ell} + \beta)$.

Remark (Directed networks)**.** *Although our current MCMC algorithm is only for undirected networks, our prior can be used for Bayesian community detection in directed networks with covariates. One can simply replace the undirected SBM likelihood in* (12) *with a directed SBM, by replacing $i < j$ with $i \ne j$ and removing the symmetry constraint on $\boldsymbol{\eta}$. The resulting sampler will be almost identical, except for minor modifications to the counts* (15) *and* (21) *to account for the extra edge information.*

3 Simulation study

In this section, we carry out multiple simulation studies in which we compare our methods, which we refer to as BCDC (Bayesian community detection for networks with covariates), with i) the covariate-assisted spectral clustering (CASC) algorithm (Binkiewicz et al., 2017), which uses both the network and the covariates information in a spectral clustering algorithm, ii) the covariated-assisted clustering on ratios of singular vectors (CASCORE) algorithm (Hu and Wang, 2022), which modifies CASC for degree heterogeneity, iii) k-means algorithms (k-means) applied only to the covariates, iv) spectral-clustering (SC) of the adjacency matrix, and v) a Bayesian SBM (BSBM), which is essentially a special case of our model with $g(S \mid \boldsymbol{x}) = 1$, therefore utilizing only the network information. For CASC, the core idea is to first construct a new Laplacian matrix $L_x = L + \tau X X^T$, where L is the Laplacian matrix for the network and X denotes the n by p feature matrix, and then apply the standard spectral clustering algorithm

on L_x. Throughout, we select τ based on the automated procedure given in (Binkiewicz et al., 2017, Section 2.3).

We consider simulation designs with (a) continuous features, (b) discrete or categorical features, (c) a mix of continuous and discrete covariates, (d) high-dimensional features, and (e) homophily effects with networks simulated from stochastic block models with varying connectivity patterns and sparsity levels. The performance of the estimated communities is measured by normalized mutual information (NMI), a measure ranging from 0 (random guessing) to 1 (perfect agreement). NMI is a measure of similarity of two partitions, and is widely used in the community detection literature. It allows comparisons of two partitions with different number of clusters while accounting for the issue of label invariance.

To define NMI, consider two partitions (labelings) on a set of objects and let (X, Y) be the two labels of a randomly drawn object. The joint probability distribution of (X, Y) is the normalized confusion matrix between the two partitions. We can define the mutual information $I(X, Y)$ and joint and marginal entropies—$H(X, Y)$, $H(X)$ and $H(Y)$—based on the aforementioned joint distribution, using standard definitions. It is common to define NMI as $I(X, Y)/H(X, Y)$. However, there are other variants and to be consistent with prior work, in particular Yan and Sarkar (2021), we will use the variant implemented in the `R` package `NMI`, namely, $2I(X, Y)/(H(X) + H(Y))$, which is also referred to as *symmetric uncertainty* (Teukolsky et al. (1992, p. 634); Hall (1998)).

Overall, the simulation studies show that our method consistently outperforms the competitors and demonstrates the gain of our model by utilizing both the network and nodal information for detecting the community structures. The simulations were performed on a high-computing cluster. An R package for our samplers, as well as the code for these experiments, is available at the GitHub repository aaamini/bcdc (Shen et al., 2022). We ran our code on a high-performance cluster with an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz with 28 cores and 256 GB RAM.

3.1 Continuous covariates

We first consider simulated networks with continuous covariates, and in particular, the Gaussian setting (9). We generate networks from an SBM having connectivity matrix $\eta = (\eta_{k\ell}) \in [0, 1]^{K \times K}$ with

$$\eta_{k\ell} = \begin{cases} p & k = \ell \\ rp & k \neq \ell \end{cases} . \tag{22}$$

The parameter $r \in [0, 1]$ controls the magnitude of disparity between the within and between connectivites and is a measure of network information for the community structure. In our simulations, we set $p = 0.1$ and vary r. We consider $n = 150$ nodes with $K = 2$ communities of 100 and 50 nodes, respectively.

For each node, we generate $d = 2$ features, with one *signal* feature related to the community structure and one *noise* feature whose distribution is the same for all nodes.

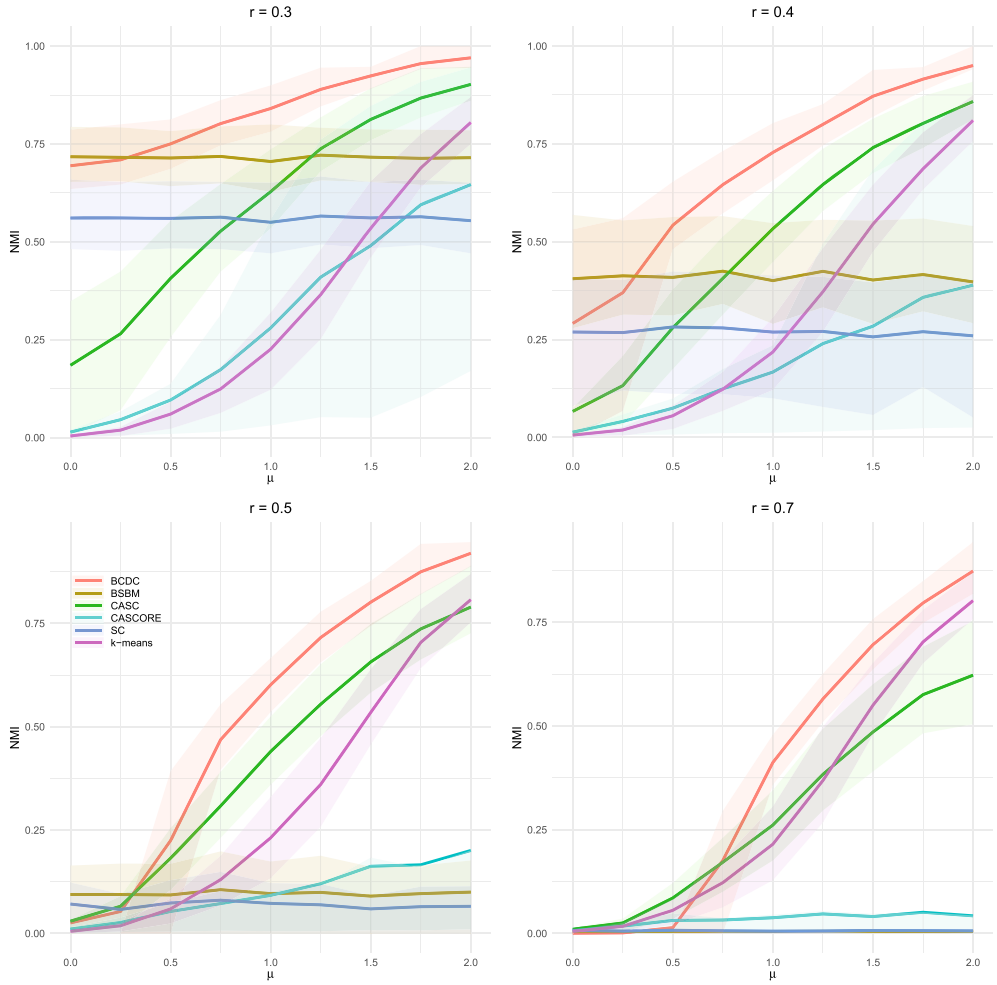

Figure 2: NMI results for five different methods on a 2-block SBM with continuous data. In all cases, $p = 0.1$, and we vary the network and covariate signal-to-noise ratios, r and μ, respectively.

Letting $x_i \in \mathbb{R}^2$ be the feature vector for node i and $e_1 = (1, 0)$, we take

$$x_i \,|\, z_i \sim N\big(\mu\sigma_{z_i} e_1, I_2\big) \ ,$$

where $\sigma_1 = +1$, $\sigma_2 = -1$ and $z_i \in \{1, 2\}$ is the community label of node i. Here $\mu \in [0, \infty)$ is proportional to the signal-to-noise ratio of the covariate information.

Figure 2 shows the mean NMI with a 50% quantile band from BCDC and competing methods, averaged over 500 replications, under different settings of r and μ. For BCDC, we have used parameters $\alpha = 10$, $\beta = 1$ and $\tau = s = 1$, and ran the sampler for 1000

iterations. Note that in our comparison, all of the competing methods were given an additional advantage by assuming the knowledge of the true number of communities. However, our method (red) consistently outperforms these other methods. An interesting notable case is that of high network information ($r = 0.3$) and pure noise covariate information ($\mu = 0$). In this case, BCDC performs as well as BSBM which only operates on network information, while CASC performs much worse being misled by pure noise covariates.

3.2 Categorical covariates

We next consider a simulation study for networks with categorical covariates. For each node i, we again generate $d = 2$ features with one *signal* feature related to the community structure and one *noise* feature whose distribution is the same for all nodes. We consider two designs:

(1) We consider networks with $n = 150$ nodes and $K = 3$ equally-sized communities. The signal features are taken to be the true community labels and the noise features are uniformly distributed on $\{1, 2, 3\}$.

(2) We consider networks with $n = 150$ nodes and $K = 2$ communities of 100 and 50 nodes. We create two 4-category features. Let $x_i = (x_{i1}, x_{i2}) \in \{1, 2, 3, 4\}^2$ be the feature vector for node i. We use the following generative model

$$\begin{aligned} \theta_1, \theta_2 &\sim \mathsf{Dir}(\mathbf{1}_4) \ , \\ x_{i1} \,|\, z_i \sim \theta_{z_i}, \quad x_{i2} \,|\, z_i &\sim \mathbf{1}_4/4 \ , \end{aligned}$$

where, for example, $x_{i1} \sim \theta_{z_i}$ means that x_{i1} is a categorical variable with probability vector θ_{z_i}.

In both cases, we use an SBM with connectivity (22), setting $p = 0.1$ and varying r from 0.1 to 0.8. For the parameters of our model, we again set $\alpha = 10$, and ran the chain for 1,500 iterations. As above, the results are given over 500 replicates.

Figure 3 shows NMI as a function of r (the network information measure) under the two covariate designs. Once again, all methods except BCDC were given the true number of communities. The NMI values obtained under the proposed model are generally higher than those of the other models with a slightly larger variance, which is likely due to the additional uncertainty in estimating the number of the communities.

3.3 Mix of continuous and discrete covariates

Here, we perform a simulation for larger networks with more communities and a mix of continuous and categorical variables. We let the number of nodes n vary from 300 to 1000 and set $K = n/50$ communities. The features are chosen so that neither perfectly separates the clusters, but both are informative. In particular, we take

$$x_{1i} \,|\, z_i \sim N\big(2(z_i \bmod 2) - 1, 1\big) \ ,$$

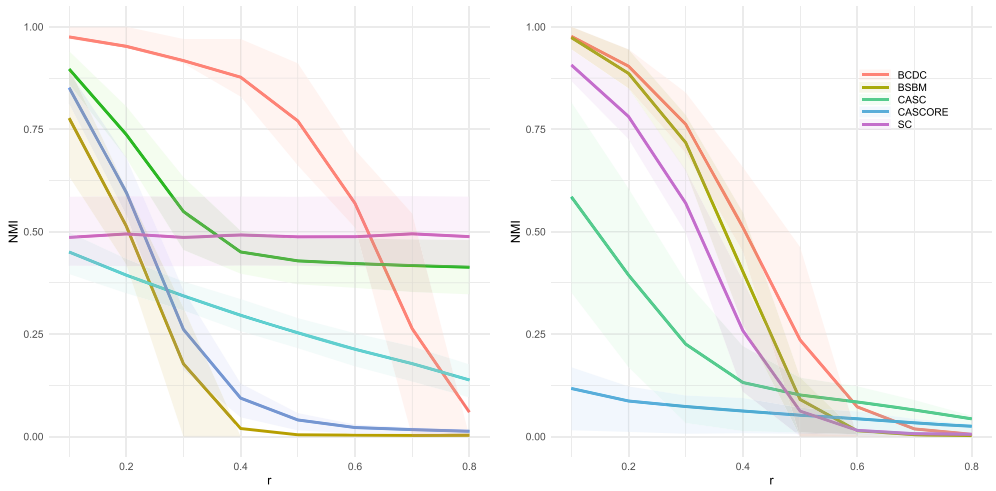

Figure 3: NMI results for five different methods on a 3-block (left) and 2-block (right) SBM with categorical data.

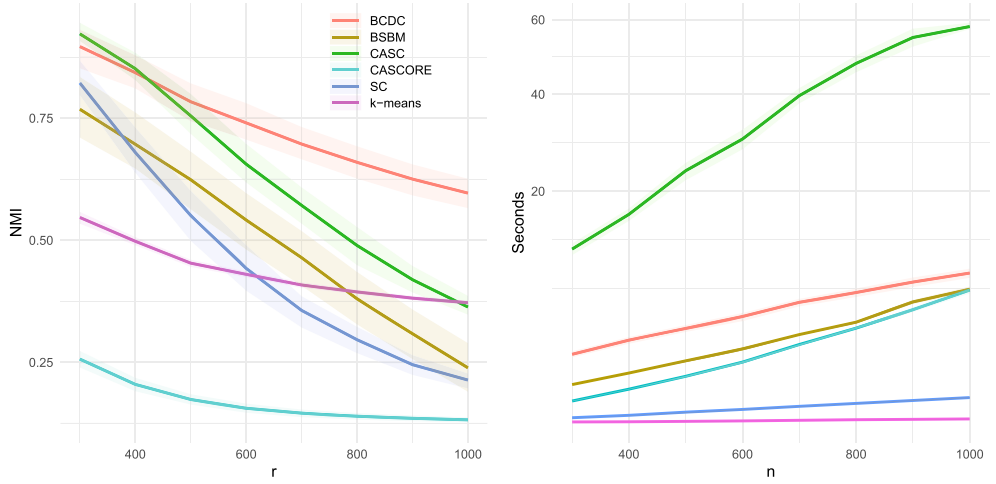

Figure 4: NMI (left) and run time (right) for five different methods when the network has a mix of continuous and discrete covariates.

i.e. $x_{1i} \sim N(1,1)$ for $z_i \in \{2,4,\ldots\}$ and $x_{1i} \sim N(-1,1)$ for $z_i \in \{1,3,\ldots\}$. We also take $x_{2i} = 1$ for $z_i \in \{1,2,\ldots K/2\}$ and $x_{2i} = 2$ for $z_i \in \{K/2+1,\ldots,K\}$. As before, we use an SBM with connectivity (22), setting $p = 0.3$ and $r = 0.35$.

The results are shown in Figure 4. We find that BCDC is competitive with the other methods when $n \leq 600$, but is superior for larger networks with $n > 600$. We also show in Figure 4 the run time for each method. We see that BCDC scales linearly with n and is faster than CASC for all of the networks.

3.4 Sparse networks and high-dimensional features

We next consider a setting from Yan and Sarkar (2021), who proposed a covariate-regularized procedure for community detection in sparse graphs. This allows us to explore whether our model works for sparse networks, as well as networks with high-dimensional features. We consider the exact simulation setting as in Yan and Sarkar (2021), in which the networks are generated from a 3-block SBM on 800 nodes with block-size ratios $3:4:5$. The true connectivity matrix is

$$B = 0.01 \begin{bmatrix} 1.6 & 1.2 & 0.16 \\ 1.2 & 1.6 & 0.02 \\ 0.16 & 0.02 & 1.2 \end{bmatrix} ,$$

leading to a very sparse network, with expected average degree ≈ 5.8. The covariates are generated from 100-dimensional Gaussian distributions $N(\mu_{z_i}, I_{100})$, with centers that are only non-zero on the first two dimensions:

$$\mu_1 = (0, 2, \mathbf{0}_{98}), \quad \mu_2 = (-1, -0.8, \mathbf{0}_{98}), \quad \mu_3 = (1, -0.8, \mathbf{0}_{98}) \ .$$

In this setting, it is difficult to distinguish clusters 1 and 2 using the network information alone, and clusters 2 and 3 based on nodal covariates alone.

This experiment was repeated 100 times for independently generated samples. In each replicate, we ran the chain for 1,000 iterations. The results are shown in Figure 5. Note that our method consistently achieves a better NMI than all of the other methods. Although we did not carry out a direct comparison with the method from Yan and Sarkar (2021) since their work focuses on regularizing high-dimensional features, our NMI results are stable and seem to be higher than those reported in Yan and Sarkar

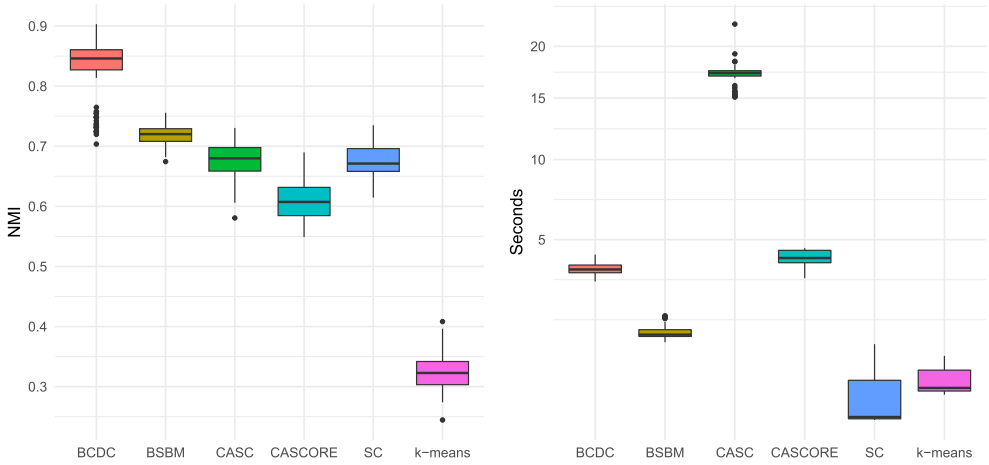

Figure 5: Boxplots of NMI (left) and run time (right) for five different methods when the network is sparse and the covariates are high-dimensional.

(2021). Importantly, we also see in Figure 5 that BCDC is only slightly slower than all of the methods that use only the network or only the covariates, but is considerably faster than CASC and CASCORE, which also use both the network and the nodal information. This illustrates the disadvantage of CASC for larger sparse networks and highlights the efficiency of our MCMC algorithm. That CASC slows down for larger networks can be attributed to the addition of the dense τXX^T to the sparse Laplacian L, resulting in an overall dense similarity matrix L_x.

3.5 Homophily

Finally, we consider a network model with homophily, which is the tendency for nodes to be connected when they share a nodal feature. For this, we sample a categorical covariate x_i with two levels, and let

$$\mathbb{P}(g_{ij} = 1 \,|\, z_i, z_j, x_i, x_j) = P_{z_i z_j} + \beta \mathbf{1}\{x_i = x_j\} \ ,$$

where $P_{z_i z_j}$ is an SBM with connectivity (22), setting $p = 0.3$ and $r = 0.7$. This creates $2K$ communities, and separates the effect of community structure from the effect of node-level covariates. We take $K = 3$ and vary β in $[-0.2, 0.2]$. Note that when $\beta > 0$, we say the nodes exhibit positive homophily. We run this for $n = 600$ and $n = 1200$.

The results are shown in Figure 6. We find that BCDC has superior performance to the other methods even when they are provided the true number ($2K$) of communities. The performance increases as the homophily effect increases in magnitude, which we should expect because the homophily effect is an informative covariate that is not included with BSBM.

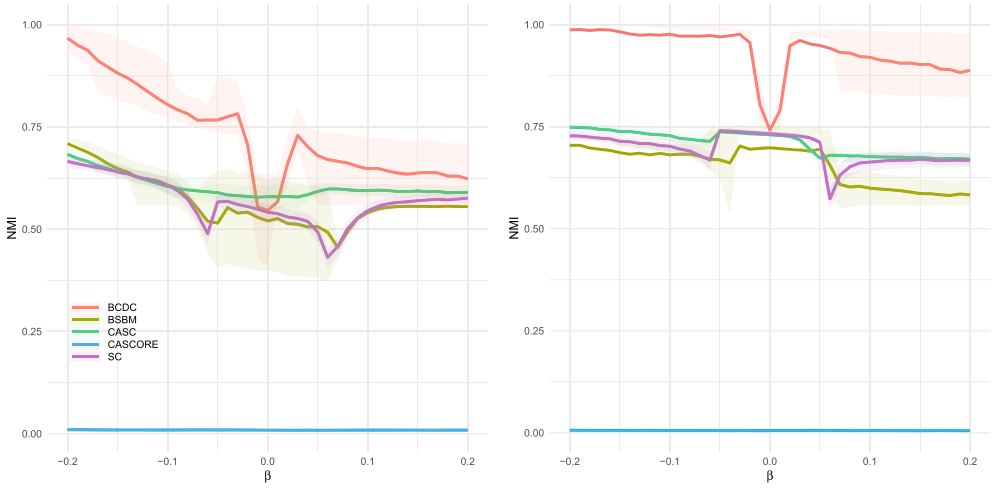

Figure 6: NMI results on a 3-block SBM with a homophily effect for $n = 600$ (left) and $n = 1200$ (right).

4 Real data analysis

In this section, we apply our model to the same two datasets from Yan and Sarkar (2021), a network representing Mexican political elites and a network representing the Weddell Sea ecosystem. We compare our the results from our models against several methods that use only the network, only the covariates, and both the network and covariates.

4.1 Performance measures

In addition to computing the NMI with the (alleged) ground truth labels, it is also helpful to compare the performance using some information criterion based on an SBM likelihood conditional on the labels. This is especially important because, unlike in the simulations, the "true" clusters are exogenously specified.

BIC The (conditional) Bayesian information criterion (BIC) is defined as the log-marginal likelihood multiplied by -2. That is,

$$\mathrm{BIC}(\boldsymbol{z}) = -2\log \int p(A \mid \boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{z}) p(\boldsymbol{\eta}) p(\boldsymbol{\pi})\, d\boldsymbol{\eta}\, d\boldsymbol{\pi} \tag{23}$$

$$\approx -2\log p(A \mid \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\pi}}, \boldsymbol{z}) + c(K)\log \binom{n}{2} \ , \tag{24}$$

where K is the number of communities in $\boldsymbol{z}$, $c(K) = \frac{1}{2}K(K+1) + (K-1)$ is the degrees of freedom in the parameters $(\boldsymbol{\eta}, \boldsymbol{\pi})$, $\boldsymbol{\pi}$ is the label prior, and $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\pi}})$ is the maximum likelihood estimator of those parameters, i.e., the maximizer of $(\boldsymbol{\eta}, \boldsymbol{\pi}) \mapsto p(A \,|\, \boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{z})$. We assume a uniform prior over $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$. Note that (24) is the well-known approximation to the BIC (Schwarz, 1978; Konishi and Kitagawa, 2008) and it shows the usefulness of $\mathrm{BIC}(\boldsymbol{z})$ as a measure of performance for real networks: Due to the presence of the complexity term $\approx c(K)\log(n^2)$, we get a good balance of the model fit and the number of communities. Label vectors $\boldsymbol{z}$ with smaller $\mathrm{BIC}(\boldsymbol{z})$ are thus more desirable from a block modeling standpoint, regardless of their relation to the ground truth.

We have, assuming uniform priors on $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$,

$$p(A \mid \boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{z}) p(\boldsymbol{\eta}) p(\boldsymbol{\pi}) = \prod_{k \le \ell} \eta_{k\ell}^{M_{k\ell}} (1-\eta_{k\ell})^{N_{k\ell} - M_{k\ell}} \prod_k \pi_k^{n_k(\boldsymbol{z})} \ , \tag{25}$$

where $n_k(\boldsymbol{z}) = \sum_i 1\{z_i = k\}$, and $M_{k\ell}$ and $N_{k\ell}$ are as in (21). Hence, the exact BIC in our setting is

$$\mathrm{BIC}(\boldsymbol{z}) = -2\Big[\sum_{k \le \ell} \log B(M_{k\ell}+1, N_{k\ell} - M_{k\ell} + 1) + \log \boldsymbol{B}(\boldsymbol{n}(z) + \mathbf{1}_K)\Big] \ ,$$

where $\boldsymbol{n}(\boldsymbol{z}) = (n_k(\boldsymbol{z}))$ and $\boldsymbol{B}(\cdot)$ is the multivariate Beta function.

WAIC In addition to BIC, we also consider WAIC (Watanabe, 2013, Section 7.1) for evaluating the methods. We mainly follow the notation and interpretation of Vehtari et al. (2017). To simplify the discussion, let us write $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\pi})$. Throughout, we condition on $\boldsymbol{z}$, so we drop the explicit mention of this conditioning. Conditioned on $\boldsymbol{z}$, the likelihood factorizes as independent (although not identically distributed) terms: $p(A \mid \boldsymbol{\theta}) = \prod_{i<j} p(A_{ij} \mid \boldsymbol{\theta})$. Let A be the data that we have observed and $\tilde{A}$ be some future data from the same unknown true distribution $\mathbb{Q}$. That is, A and $\tilde{A}$ are a pair of i.i.d. copies from $\mathbb{Q}$. One assumes that $\mathbb{Q}$ also factorizes over coordinates. WAIC is an approximation to the so-called ELPD, which we multiply by -1 compared to Vehtari et al. (2017),

$$-\sum_{i<j} \mathbb{E}_{\tilde{A}\sim\mathbb{Q}}[\log p^*(\tilde{A}_{ij} \mid A)] \tag{26}$$

were $p^*(\tilde{A}_{ij} \mid A)$ is the posterior predictive density of $\tilde{A}_{ij}$ given A under the model. Letting $\mathbb{E}_{\text{post}}$ denote the expectation under the posterior distribution of $\boldsymbol{\theta}$ given A,

$$p^*(\tilde{A}_{ij} \mid A) = \mathbb{E}_{\text{post}}[p(\tilde{A}_{ij} \mid \boldsymbol{\theta})].$$

WAIC is an approximation to (26) and is given by

$$\text{WAIC} = -\sum_{i<j} \log p^*(A_{ij} \mid A) + \sum_{i<j} \text{var}_{\text{post}}(\log p(A_{ij} \mid \boldsymbol{\theta}))$$

where the second term can be thought of as a measure of model complexity.

Remark. *Our WAIC is* -1 *times the WAIC of Vehtari et al. (2017) and* n *times that of Watanabe (2013).*

Note that,

$$\text{WAIC} = -\sum_{i<j} \Big\{ \log \mathbb{E}_{\text{post}}[p(A_{ij} \mid \boldsymbol{\theta})] + \text{var}_{\text{post}}(\log p(A_{ij} \mid \boldsymbol{\theta})) \Big\}$$

where both $\mathbb{E}_{\text{post}}$ and var_{post} are usually obtained by Monte Carlo approximation using a sample drawn from the posterior of $\boldsymbol{\theta}$.

Under the model we are considering here, however, WAIC can be derived in closed-from. The posterior of $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\pi})$ given A is proportional to (25), hence the posterior is $\eta_{k\ell} \sim \mathsf{Beta}(M_{k\ell} + 1, N_{k\ell} - M_{k\ell} + 1)$ and $\pi_k \sim \mathsf{Dir}(n_k(\boldsymbol{z}) + 1)$. Since $p(A_{ij} \mid \boldsymbol{\theta}) = \eta_{z_i z_j}^{A_{ij}}(1 - \eta_{z_i z_j})^{1-A_{ij}}$ only depends on $\boldsymbol{\eta}$, only the posterior of $\boldsymbol{\eta}$ is relevant. We have

$$\begin{aligned}
\mathbb{E}_{\text{post}}[p(A_{ij} \mid \boldsymbol{\theta})] &= \begin{cases} \mathbb{E}_{\text{post}}[\eta_{z_i z_j}] & A_{ij} = 1 \\ \mathbb{E}_{\text{post}}[1 - \eta_{z_i z_j}] & A_{ij} = 0 \end{cases} \\
&= \Big(\frac{M_{z_i z_j} + 1}{N_{z_i z_j} + 2}\Big)^{A_{ij}} \Big(\frac{N_{z_i z_j} - M_{z_i z_j} + 1}{N_{z_i z_j} + 2}\Big)^{1-A_{ij}}.
\end{aligned}$$

Next, we recall that if $X \sim \mathsf{Beta}(\alpha, \beta)$, then, $\mathrm{var}[\log X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$ and $\mathrm{var}[\log(1-X)] = \psi_1(\beta) - \psi_1(\alpha+\beta)$ where $\psi_1(\cdot)$ is the trigamma function. By considering the two possible values of A_{ij} as above, we have

$$\mathrm{var}_{\mathrm{post}}(\log p(A_{ij} \mid \boldsymbol{\theta})) = \begin{cases} \psi_1(M_{z_i z_j} + 1) - \psi_1(N_{z_i z_j} + 2), & A_{ij} = 1 \\ \psi_1(N_{z_i z_j} - M_{z_i z_j} + 1) - \psi_1(N_{z_i z_j} + 2), & A_{ij} = 0 \end{cases}.$$

Put together, we obtain our closed form for WAIC.

4.2 Mexican political elites

The first dataset we consider involves Mexican political elites (Gil-Mendieta and Schmidt, 1996). In this network, the $n = 35$ vertices represent Mexican presidents and their close collaborators, and the 117 edges represent significant political, kinship, friendship, or business ties among them. The ground truth is a classification of the politicians according to their professional background: military and civilians. The covariate we include is the number of years since 1990 that the actor first got a significant governmental position. Figure 7 reveals that this covariate has some discriminatory power in the cluster labels. This is due to the fact that after the Mexican revolution at the beginning of the twentieth century, the political elite was dominated by the military, and later the civilians gradually succeeded the power.

Table 1 contains the NMI results of our method compared with the same comparison methods in the simulation section. We see that our method achieves the best results, and we visualize our estimated clusters in the network compared with the true labels in Figure 8. Again, for the other methods, we assume the knowledge of the true number of clusters while ours learns the number of the clusters via posterior inference for NMI comparisons.

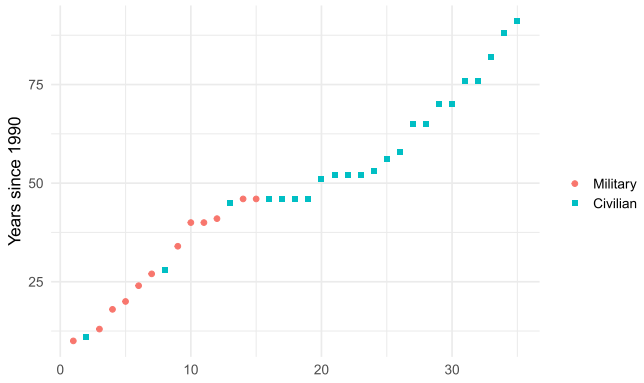

Figure 7: Node feature for the Mexican political network, which is the number of years since 1990 that the actor first got a significant governmental position.

Dataset	BCDC	CASC	CASCORE	k-MEANS	SC	BSBM
Mexican politicians	**0.43**	0.37	0.28	0.26	0.37	0.30
Weddell Sea	**0.44**	0.25	0.15	0.35	0.33	0.23

Table 1: NMI results on the two real datasets.

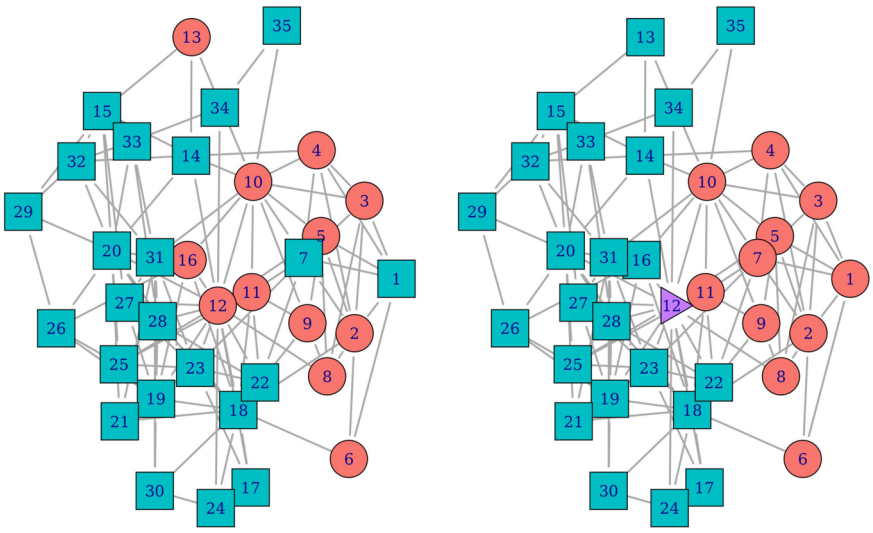

Figure 8: Mexican political network, colored by true (left) and estimated (right) clusters.

As already pointed out in Yan and Sarkar (2021), node 35 has exactly one connection to each of the military and civilian groups, but obtained a governmental position in the 90s, which greatly hinted at a civilian background. By using the covariate, our method accurately captures this label. On the other hand, node 9 seized power in 1940 when the government was almost equally represented by civilian and military politicians, which makes detecting his group difficult, but has more edges to the military group than the civilian group. In this case, our method correctly assigns the military label to it by considering the graph structure.

We also notice that node 1 has five connections to the military and only one connection to the civilian, and node 1 seized power in 1911. Similarly for node 7, which has five connections to the military and three connections to the civilian and seized power in 1928. For these nodes, both the network and the covariates strongly indicate a closer relationship to the military, which is what our method assigns despite the true label showing civilian. Finally, our method assigns node 12 to its own cluster. This is likely because this is the highest-degree node with 5 military connections and 12 civilian connections. However, in researching node 12, we discovered that Miguel Alemán

Valdés was the first civilian president after several military presidents, which suggests that there may have been a labeling error in the original publication of this dataset from Gil-Mendieta and Schmidt (1996). This could explain why our method has the best BIC and WAIC, even better than the "true" labels, in Tables 2 and 3.

Dataset	"TRUE"	BCDC	CASC	CASCORE	k-MEANS	SC	BSBM
Mexican politicians	636	**586**	587	608	626	587	587
Weddell Sea	138k	**21k**	124k	120k	144k	100k	71k

Table 2: BIC results on the two real datasets.

Dataset	"TRUE"	BCDC	CASC	CASCORE	k-MEANS	SC	BSBM
Mexican politicians	283	**248**	268	259	279	259	259
Weddell Sea	68k	**8k**	61k	59k	71k	49k	35k

Table 3: WAIC results on the two real datasets.

4.3 Weddell sea ecosystem

The second dataset we consider is a predator-prey, directed network representing the marine food web of the Weddell Sea off of the Antarctic Peninsula, which was collected by Jacob et al. (2011). Since ecosystems are complex, interconnected environments, network analyses have emerged as a popular technique for untangling these connections. The Weddell Sea network has 487 nodes that signify different marine species, and there is a link between nodes i and j if species i (predator) feeds on species j (prey). Following Yan and Sarkar (2021), we construct a binary, undirected network from this directed network in which $A_{ij} = 1$ if there are at least 5 common prey between species i and j, and $A_{ij} = 0$ otherwise. The network is shown in Figure 9.

In Jacob et al. (2011), the authors analyze the relationship between the body size of each species and its feeding type: primary producer, herbivorous/detrivorous, detrivorous, carnivorous, carnivorous/necrovorous, and omnivorous. Figure 1 shows these body sizes grouped by feeding type, where, again following Yan and Sarkar (2021), we group detrivorous, carnivorous, carnivorous/necrovorous as "Carnivore" to obtain four groups. The adjacency matrix in Figure 9 (left) is also sorted by these groups. The authors of Jacob et al. (2011) found body size to be positively correlated with trophic level, but noted that "predators on intermediate trophic levels do not necessarily feed on smaller or prey similar in size but depending on their foraging strategy have a wider

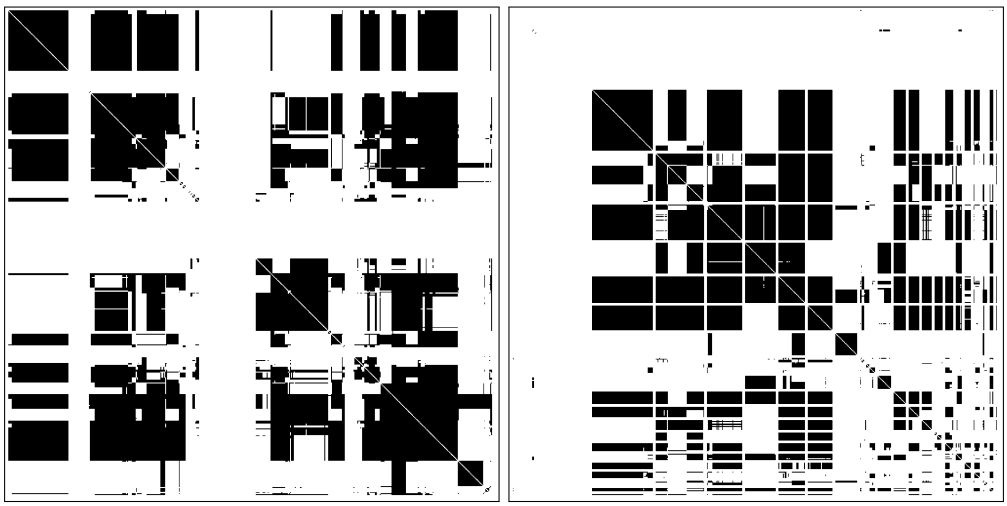

Figure 9: Visualization of the adjacency matrix for the Weddell Sea network clustered by the true feeding type (left) and estimated clusters (right). In each case, the rows and columns of the matrix are permutated so that nodes in the same cluster form contiguous blocks. Clusters on the right are ordered according to their size.

prey size range available." Therefore, body size is insufficient on its own to distinguish the groups, and it would be preferable to consider the interconnectedness of the food web when tasked with clustering the species.

Table 1 shows that BCDC provides the best clustering results compared to the other methods. Therefore, using all of the available information provides an improvement in clustering accuracy over the use of just the network structure or the nodal information. As before, BCDC is the only method that did not know that there are four "true" groups. Interestingly, BCDC estimates many more clusters – 21 in total – which may explain its higher NMI, since we see qualitatively in Figure 9 (right) a more refined block structure. This is quantified and corroborated through BIC and WAIC in Tables 2 and 3, which again shows our method outperforms even the "true" clusters. All of this suggests there may be distinct sub-blocks within the Herbivore, Carnivore, and Omnivore classes.

5 Discussion

In this work, we proposed a Bayesian model for community detection in networks with covariates in which both the network and node features of the network are jointly utilized for estimating community structure. In particular, the contribution of nodal information is explicitly modeled in the prior distribution for the community labels via a covariate-dependent random partition prior. We proposed efficient MCMC algorithms for sampling the posterior distributions of all the parameters including the community

labels and the number of the communities. Numerical studies demonstrated the overall superior performance of our model over many of the existing methods.

Compared to an almost exclusive literature of frequentist methods, our work is among the first in proposing a Bayesian approach for tackling the problem, which confers some notable advantages in terms of uncertainty quantification, as well as estimating all the model parameters. Notably, unlike the other methods in the literature, our model estimates the number of communities via posterior inference without any knowledge or prior information on the true number. Future work will be devoted to developing Bayesian models for community detection in degree-corrected SBMs and dynamic network models.

We can also easily extend our model to a partially-observed SBM in the spirit of Zhou (2015). Specifically, we can modify (12) to

$$P(A \mid \boldsymbol{\eta}, \boldsymbol{z}) = \prod_{1 \le i < j \le n} \left[\eta_{z_i, z_j}^{A_{ij}} (1 - \eta_{z_i, z_j})^{1 - A_{ij}} \right]^{m_{ij}} ,$$

where $\boldsymbol{m} = (m_{ij}) \in \{0, 1\}^{n \times n}$ is a (symmetric) observation mask, with $m_{ij} = 1$ for the observed edges. This only effects sampling $\boldsymbol{z}$ and β through the modified counts

$$O_{i\ell} = \sum_{j : j \neq i} m_{ij} A_{ij} 1\{z_j = \ell\}, \quad n_{i\ell} = \sum_{j : j \neq i} m_{ij} 1\{z_j = \ell\} ,$$

replacing (15), and

$$M_{k\ell} = \sum_{(i,j) \in \Gamma_{k\ell}} m_{ij} A_{ij} 1\{z_i = k, z_j = \ell\}, \quad N_{k\ell} = \sum_{(i,j) \in \Gamma_{k\ell}} m_{ij} 1\{z_i = k, z_j = \ell\} ,$$

replacing (21). This allows our model to also predict missing edges.

Finally, it may be of interest to test whether there is an association between the node covariates and inferred community structure. One approach to this is with Bayes factors comparing models with and without covariates, using, for example, the approach in Legramanti et al. (2020) for testing partition structures in SBMs. While this is a principled Bayesian approach, it only tests whether the set of covariates provides a more parsimonious clustering than without the covariates rather than identifying which covariates are significant. One idea for testing individual covariate significance is to test for a difference between the posterior distributions of the cluster centers $\boldsymbol{\xi}$ implied by $\boldsymbol{z}$. Note that this corresponds to testing whether the priors ν on auxiliary probability distributions q have overlapping variances, but we leave a rigorous treatment of this idea to future work.

Acknowledgments

We are very grateful to the Editor, the Associate Editor and two reviewers for their valuable comments.

Funding

LS and LL were supported by NSF grants DMS 2113642 and DMS 1654579. AA would like to acknowledge the support of NSF grant DMS-1945667. NJ was partially supported by NIH/NICHD grant 1DP2HD091799-01.

References

Airoldi, E. M., Blei, D., Fienberg, S., and Xing, E. (2008). "Mixed membership stochastic blockmodels." *Advances in neural information processing systems*, 21. 7

Amini, A. A. and Levina, E. (2018). "On semidefinite relaxations for the block model." *The Annals of Statistics*, 46(1): 149 – 179. URL https://doi.org/10.1214/17-AOS1545 MR3766949. doi: https://doi.org/10.1214/17-AOS1545. 2

Amini, A. A., Paez, M. S., and Lin, L. (2019). "Hierarchical Stochastic Block Model for Community Detection in Multiplex Networks." *arXiv e-prints*, arXiv:1904.05330. MR4692550. doi: https://doi.org/10.1214/22-ba1355. 2

Ball, B., Karrer, B., and Newman, M. (2011). "Efficient and principled method for detecting communities in networks." *Physical Review E*, 84(3): 036103. 2

Bickel, P. J. and Chen, A. (2009). "A nonparametric view of network models and Newman Girvan and other modularities." *Proceedings of the National Academy of Sciences of the Unites States of America*, 106(50): 21068–21073. 1

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017). "Covariate-assisted spectral clustering." *Biometrika*, 104(2): 361–377. MR3698259. doi: https://doi.org/10.1093/biomet/asx008. 2, 11, 12

Erdős, P. and Rényi, A. (1959). "On Random Graphs I." *Publicationes Mathematicae (Debrecen)*, 6: 290–297. MR0120167. doi: https://doi.org/10.5486/pmd.1959.6.3-4.12. 1

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The annals of statistics*, 209–230. MR0350949. 4

Gil-Mendieta, J. and Schmidt, S. (1996). "The political network in Mexico." *Social Networks*, 18(4): 355–381. 20, 22

Hall, M. A. (1998). "Correlation-based feature subset selection for machine learning." *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*. 12

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). "Stochastic blockmodels: First steps." *Social networks*, 5(2): 109–137. MR0718088. doi: https://doi.org/10.1016/0378-8733(83)90021-7. 2

Hu, Y. and Wang, W. (2022). "Covariate-Assisted Community Detection on Sparse Networks." *arXiv*. URL https://arxiv.org/abs/2208.00257 2, 11

Jacob, U., Thierry, A., Brose, U., Arntz, W. E., Berg, S., Brey, T., Fetzer, I., Jonsson,

T., Mintenbeck, K., Möllmann, C., et al. (2011). "The role of body size in complex food webs: A cold case." *Advances in ecological research*, 45: 181–223. 2, 22

Karrer, B. and Newman, M. E. J. (2011). "Stochastic blockmodels and community structure in networks." *Phys. Rev. E*, 83: 016107. URL http://link.aps.org/doi/10.1103/PhysRevE.83.016107 MR2788206. doi: https://doi.org/10.1103/PhysRevE.83.016107. 2

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). "Learning Systems of Concepts with an Infinite Relational Model." In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, 381–388. AAAI Press. 6

Kim, C., Bandeira, A. S., and Goemans, M. X. (2017). "Community detection in hypergraphs, spiked tensor models, and sum-of-squares." In *2017 International Conference on Sampling Theory and Applications (SampTA)*, 124–128. IEEE. 6

Kim, D. I., Hughes, M. C., and Sudderth, E. B. (2012). "The Nonparametric Metadata Dependent Relational Model." In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress. URL http://icml.cc/2012/papers/771.pdf 7

Kolaczyk, E. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Verlag. MR2724362. doi: https://doi.org/10.1007/978-0-387-88146-1. 1

Kolaczyk, E. D., Lin, L., Rosenberg, S., Walters, J., and Xu, J. (2020). "Averages of unlabeled networks: Geometric characterization and asymptotic behavior." *The Annals of Statistics*, 48(1): 514 – 538. URL https://doi.org/10.1214/19-AOS1820 MR4065172. doi: https://doi.org/10.1214/19-AOS1820. 1

Konishi, S. and Kitagawa, G. (2008). "Information criteria and statistical modeling." MR2367855. doi: https://doi.org/10.1007/978-0-387-71887-3. 18

Legramanti, S., Rigon, T., and Durante, D. (2020). "Bayesian testing for exogenous partition structures in stochastic block models." *Sankhya A*, 1–19. MR4402748. doi: https://doi.org/10.1007/s13171-020-00231-2. 24

Lei, J., Chen, K., and Lynch, B. (2020). "Consistent community detection in multi-layer network data." *Biometrika*, 107(1): 61–73. MR4064140. doi: https://doi.org/10.1093/biomet/asz068. 6

Lovász, L. (2012). *Large Networks and Graph Limits*, volume 60. American Mathematical Society Providence. MR3012035. doi: https://doi.org/10.1090/coll/060. 1

Luxburg, U. V. (2007). "A tutorial on spectral clustering." *Statistics and Computing*, 17(4): 395–416. MR2409803. doi: https://doi.org/10.1007/s11222-007-9033-z. 1

Mørup, M. and Schmidt, M. N. (2012). "Bayesian Community Detection." *Neural Comput.*, 24(9): 2434–2456. URL http://dx.doi.org/10.1162/NECO_a_00314 MR2986776. doi: https://doi.org/10.1162/NECO_a_00314. 2, 6

Müller, P. and Quintana, F. (2010). "Random partition models with regression on covariates." *Journal of statistical planning and inference*, 140(10): 2801–2808. MR2651966. doi: https://doi.org/10.1016/j.jspi.2010.03.002. 4

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A product partition model with regression on covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 4

Newman, M. E. and Clauset, A. (2016). "Structure and inference in annotated networks." *Nature communications*, 7(1): 1–11. 7

Newman, M. E. J. (2006). "Modularity and community structure in networks." *Proceedings of the National Academy of Sciences*, 103(23): 8577–8582. URL http://www.pnas.org/content/103/23/8577.abstract MR2676073. doi: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001. 1

Park, J.-H. and Dunson, D. B. (2010). "Bayesian generalized product partition model." *Statistica Sinica*, 1203–1226. MR2730180. 4

Ren, L., Du, L., Carin, L., and Dunson, D. (2011). "Logistic Stick-Breaking Process." *J. Mach. Learn. Res.*, 12(null): 203–239. MR2773552. 7

Rohe, K., Chatterjee, S., and Yu, B. (2011). "Spectral clustering and the high-dimensional stochastic block model." *Annals of Statistics*, 39: 1878–1915. MR2893856. doi: https://doi.org/10.1214/11-AOS887. 1

Schwarz, G. (1978). "Estimating the dimension of a model." *The annals of statistics*, 461–464. MR0468014. 18

Sethuraman, J. (1994). "A CONSTRUCTIVE DEFINITION OF DIRICHLET PRIORS." *Statistica Sinica*, 4(2): 639–650. MR1309433. 4

Shen, L., Amini, A. A., Josephs, N., and Lin, L. (2022). "BCDC model for community detection with node covariates." https://github.com/aaamini/bcdc. 12

Sweet, T. M. (2015). "Incorporating Covariates Into Stochastic Blockmodels." *Journal of Educational and Behavioral Statistics*, 40(6): 635–664. URL https://ideas.repec.org/a/sae/jedbes/v40y2015i6p635-664.html 2

Tallberg, C. (2004). "A Bayesian Approach to Modeling Stochastic Blockstructures with Covariates." *The Journal of Mathematical Sociology*, 29(1): 1–23. URL https://doi.org/10.1080/00222500590889703 MR2753229. 2

Teukolsky, S. A., Flannery, B. P., Press, W., and Vetterling, W. (1992). "Numerical recipes in C." *SMR*, 693(1): 59–70. MR1201159. 12

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and computing*, 27: 1413–1432. MR3647105. doi: https://doi.org/10.1007/s11222-016-9696-4. 19

Wang, M. and Zeng, Y. (2019). "Multiway clustering via tensor block models." *Advances in neural information processing systems*, 32. 6

Watanabe, S. (2013). "A widely applicable Bayesian information criterion." *The Journal of Machine Learning Research*, 14(1): 867–897. MR3049492. 19

Weng, H. and Feng, Y. (2022). "Community detection with nodal information: Likelihood and its variational approximation." *Stat*, 11. MR4394988. doi: https://doi.org/10.1002/sta4.428. 2

Wolfe, P. J. and Olhede, S. C. (2013). "Nonparametric graphon estimation." *ArXiv e-prints*. MR2908387. 1

Yan, B. and Sarkar, P. (2021). "Covariate Regularized Community Detection in Sparse Graphs." *Journal of the American Statistical Association*, 116(534): 734–745. MR4270020. doi: https://doi.org/10.1080/01621459.2019.1706541. 3, 12, 16, 18, 21, 22

Zhang, Y., Chen, K., Sampson, A., Hwang, K., and Luna, B. (2019). "covariate Adjusted Stochastic Block Model." *Journal of Computational and Graphical Statistics*, 28(2): 362–373. MR3974886. doi: https://doi.org/10.1080/10618600.2018.1530117. 2

Zhang, Y., Levina, E., Zhu, J., et al. (2016). "Community detection in networks with node features." *Electronic Journal of Statistics*, 10(2): 3153–3178. MR3571965. doi: https://doi.org/10.1214/16-EJS1206. 3

Zhao, H., Du, L., and Buntine, W. (2017). "Leveraging Node Attributes for Incomplete Relational Data." In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 4072–4081. PMLR. URL https://proceedings.mlr.press/v70/zhao17a.html 7

Zhou, M. (2015). "Infinite edge partition models for overlapping community detection and link prediction." In *Artificial intelligence and statistics*, 1135–1143. PMLR. 7, 24