

A Bayesian Nonparametric Stochastic Block Model for Directed Acyclic Graphs

Clement Lee¹ and Marco Battiston²

¹School of Mathematics, Statistics and Physics, Newcastle University, UK

²Department of Mathematics and Statistics, Lancaster University, UK

August 19, 2025

Abstract

Random graphs have been widely used in statistics, for example in network analysis and graphical models. In some applications, the data may contain an inherent hierarchical ordering among its vertices, which prevents directed edges between pairs of vertices that do not respect this order. For example, in bibliometrics, older papers cannot cite newer ones. In such situations, the resulting graph forms a Directed Acyclic Graph. In this article, we extend the Stochastic Block Model (SBM) to account for the presence of such ordering in the data, ignoring which can lead to biased estimates of the number of blocks. The proposed approach includes in the model likelihood a *topological ordering*, which is treated as an unknown parameter and endowed with a prior distribution. We describe how to formalise the model and perform posterior inference for a Bayesian nonparametric version of the SBM in which both the hierarchical ordering and the number of latent blocks are learnt from the data. Finally, an illustration with real-world datasets from bibliometrics is presented. Additional supplementary materials are available online.

Key Words: Topological Ordering, Directed Acyclic Graphs, Stochastic Block Model, Markov chain Monte Carlo, Pitman-Yor process.

1 Introduction

Random graphs have been widely used in statistics and computer science over the last years to model network, interaction and bibliometric data, just to mention a few among many fields of applications. For example, in social networks, friendships or message exchanges among users can be modelled as random graphs, having users as vertices, and edges between two users if they are friends or have exchanged messages. As another example, citation datasets in bibliometrics analysis can be modelled as random graphs, having articles as vertices, and directed edges from citing articles to cited articles.

In some applications, the dataset and the corresponding graph can contain an inherent hierarchical order among its vertices. This means that directed edges can be present only from vertices that appear earlier in this order to ones that appear later. As an example, when modelling citation networks among journal articles, e.g. Ji and Jin (2016), articles can cite only articles that were published earlier. As another example, in a network having the employees of a big company as vertices and directed edges from the supervisor to the supervisee, a directed path from one vertex to another can be present only if the former is higher than the latter one in the company hierarchy. Similarly, the same structure appears in a dataset having directed edges from PhD supervisors to their students. In all these applications, the observed graph is a *Directed Acyclic Graph* (DAG), i.e. it cannot display any directed cycles (any directed path having the same initial and final vertex). Moreover, it respects a latent *topological ordering* among its vertices, i.e. a linear ordering \prec of the vertex set \mathcal{V} such that every pair $(p, q) \in \mathcal{V} \times \mathcal{V}$ can belong to the edge set \mathcal{E} only if $p \prec q$. In general, the topological ordering is not unique, i.e. for a DAG there can be multiple orderings that satisfy the definition above.

When analysing a given network dataset that is a DAG, if the latent topological ordering is not properly accounted for, the statistical models can assign positive probability to ‘impossible’ edges, i.e. directed edges in which the first vertex appears later than the second one in the topological ordering. If this is the case for many edges, the resulting estimates from the model can deteriorate dramatically. The same phenomenon arises in other statistical contexts, for example in presence of structural zeros when modelling tabular data (Manrique-Vallier and Reiter, 2014), in which assigning, whatever small, positive probability to ‘impossible’ cells can deteriorate estimates and bias results.

In this work, we propose an approach to extending the popular *Stochastic Block Model* (SBM) (Holland et al., 1983) for random graphs in general to graphs that are specifically DAGs. The SBM, being among the most popular random graph models for vertex clustering and community detection, was initially introduced by Holland et al. (1983) and Wang and Wong (1987), then formalised as latent models by Snijders and Nowicki (1997) and Nowicki and Snijders (2001), and has further gained popularity with the incorporation of degree correction by Karrer and Newman (2011). In online Appendices A.1 and A.2, we provide a brief review of this model, its generalisations and approaches to posterior inference. For recent comprehensive reviews, the reader is also referred to Abbe (2018) and Lee and Wilkinson (2019), which focus on theoretical results and modelling approaches, respectively. When extending the SBM to model a given DAG, the underlying topological ordering is unknown and not necessarily unique. Furthermore, ignoring the ordering can lead to dramatic underestimates or overestimates of the number of blocks, as the simulations in Section 3.1 demonstrate. Therefore, it will be included in the likelihood of the model as a parameter, endowed with a prior and inferred a posteriori, together with the community memberships and other parameters of the model.

The background literature regarding the use of the SBM for DAGs is quite limited and mainly consists of an unpublished technical report by Lee and Wilkinson (2018), which only presents a preliminary data analysis of a DAG using a SBM. Not only was degree correction

not incorporated, but there was also no formal consideration of the issue of the number of blocks. Instead, the model was merely fitted with different numbers of blocks separately with no model selection, thus not justifying the interpretations of their results.

The issue of the number of communities or blocks is common with SBMs and clustering in general, and has been reviewed in Table 1 of Lee and Wilkinson (2019). One common approach to this issue is the use of Bayesian nonparametrics, which assigns a prior for the allocation vector and allows the posterior sampler to infer the number of blocks from the data. Kemp et al. (2006) and Schmidt and Mørup (2013) initially considered the *infinite relational model*, which is a variant of the plain SBM, using a Dirichlet process prior, while Geng et al. (2019) extends this model by the use of a finite regime Pitman-Yor (PY) process (Pitman and Yor, 1997), which is reviewed in the online Appendix A.3. A further generalisation is in Legramanti et al. (2022), in which the Gibbs-type prior formulation is used instead of the PY prior, for an application to criminal networks.

In the proposed model, we also use a Bayesian nonparametric version of the SBM, but by adopting the PY process in a different way. This process allows two possible regimes: one in which the total number of blocks is finite (and unknown) and another one in which it increases with the sample size. While other works such as Geng et al. (2019) have used Bayesian nonparametric formulation for the block membership, they usually focused only on the first of the two regimes. As these two regimes imply different asymptotic behaviours of the number of blocks, the model might be misspecified if only one regime is assumed before fitting to the data. To circumvent this, the choice of regime becomes another part of the proposed model, with a model selection step based on Gibbs variable selection (Carlin and Chib, 1995) embedded in the Markov chain Monte Carlo (MCMC) sampler.

The proposed approach is illustrated using both simulated data and real datasets from bibliometrics. In the simulation study, where the proposed model is compared to a SBM for directed graphs, which does not account for the latent order in the likelihood, the latter brings about the aforementioned issue of assigning positive probability to many ‘impossible’

edges, as well as substantial biases in the estimates of the overall number of blocks. In the applications to two datasets of citation networks, it is shown that the model can infer the community structure and latent hierarchical order among vertices.

The rest of this article is organised as follows. The SBM for DAGs is presented in Section 2. Section 3 presents both simulated (Section 3.1) and real data (Sections 3.2 and 3.3) illustrations. A final discussion in Section 4 concludes the article. The online appendices contain reviews of the SBM and PY process, a detailed description of the MCMC sampler to perform posterior inference for the SBM for DAG, additional information on the likelihood derivations, and plots from the real-data illustrations.

2 Model

To introduce the model and notation used throughout the rest of the paper, let us consider a directed network to be represented as a directed multi-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the vertex set and \mathcal{E} is the edge set. The size of \mathcal{V} , denoted by $n := |\mathcal{V}|$, is the number of vertices of \mathcal{G} . The $n \times n$ adjacency matrix of the graph is denoted by $\mathbf{Y} := (Y_{pq})_{1 \leq p, q \leq n}$. If there are y directed edges from vertex p to vertex q , i.e. y copies of (p, q) in \mathcal{E} , then $Y_{pq} = y$, otherwise $Y_{pq} = 0$. We also assume no self-loops, i.e. $Y_{pp} = 0$ for all $p = 1, 2, \dots, n$. Next, we define $\mathbf{Z}_n := (Z_1, \dots, Z_n)$ to be the allocation vector of length n , where Z_p is a label associated to vertex p . Two vertices, p and q , belong to the same group if and only if $Z_p = Z_q$. Essentially, \mathbf{Z}_n represents the group memberships of the vertices. We assume there are $K_n > 1$ unique labels displayed in \mathbf{Z}_n , denoted $(Z_1^*, \dots, Z_{K_n}^*)$ hence K_n groups of vertices. When whether Z_i^* is the label associated with vertex p is of concern, for notational convenience, hereafter we write $\mathbb{I}(Z_p = i)$ in place of $\mathbb{I}(Z_p = Z_i^*)$, where $\mathbb{I}(A)$ denotes the indicator function of the event A , i.e. $\mathbb{I}(A) = 1$ if A is true, 0 otherwise. We also define $\mathbf{C} := (C_{ij})_{1 \leq i, j \leq K_n}$ to be the block matrix, where C_{ij} represents the general connectivity between blocks i and j . Similar to above, given \mathbf{C} and $Z_p = Z_k^*$, $C_{Z_p j}$ will denote element C_{kj} .

The main idea of the classic SBM is that the number of edges Y_{pq} from vertex p to vertex q is independent of that of any other dyad, *conditional on their group memberships* Z_p and Z_q (Holland et al., 1983, Nowicki and Snijders, 2001). Moreover, we will consider the *degree-corrected* version of the SBM (Karrer and Newman, 2011), which takes into account the degree heterogeneity of the vertices within the same group, by introducing the vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ of vertex-specific parameters, where $\xi_p > 0$ is the degree correction factor for vertex p . Namely, the number of edges between vertex p and q is modelled as

$$Y_{pq} | \mathbf{C}, \mathbf{Z}_n, \boldsymbol{\xi} \stackrel{\text{ind}}{\sim} \text{Poisson}(\xi_p \xi_q C_{Z_p Z_q}). \quad (1)$$

To introduce the proposed model, we restrict \mathcal{G} to be a DAG for the rest of this section. To utilise a unique feature of DAGs, as discussed in Section 1, we define $\boldsymbol{\sigma} := (\sigma_1, \dots, \sigma_n)$ as the random vector that represents the ordering of \mathcal{G} , with the collection of all permutations of $\{1, 2, \dots, n\}$ as the sample space. A value of $\boldsymbol{\sigma}$ is deemed topological with respect to \mathcal{G} if it satisfies the definition of topological ordering, i.e. if vertex p precedes vertex q in the ordering, then there cannot be edges from q to p . We define several quantities implied from $\boldsymbol{\sigma}$. First, $\boldsymbol{\phi} := (\phi_1, \dots, \phi_n)$ is the “inverse” of $\boldsymbol{\sigma}$, which means that if vertex p is the r -th vertex in the topological ordering, we have $\sigma_r = p$ and $\phi_p = r$. Essentially, $\boldsymbol{\phi}$ contains the position of each vertex in $\boldsymbol{\sigma}$. Second, we define $\mathbf{Z}_n^\sigma := (Z_1^\sigma, Z_2^\sigma, \dots, Z_n^\sigma)$ as the *reordered* allocation vector, where $Z_p^\sigma = Z_{\sigma_p}$ and Z_{σ_p} comes from \mathbf{Z}_n . In the same way, we define $\boldsymbol{\xi}^\sigma := (\xi_1^\sigma, \xi_2^\sigma, \dots, \xi_n^\sigma)$ as the reordered version of $\boldsymbol{\xi}$. Lastly, \mathbf{Y}^σ is the adjacency matrix reordered by $\boldsymbol{\sigma}$ for the columns and rows of \mathbf{Y} simultaneously, such that $Y_{pq}^\sigma = Y_{\sigma_p \sigma_q}$.

The central component of the SBM is the distribution assumption about each dyad of \mathcal{G} , essentially each element of \mathbf{Y} . Due to the acyclic nature of \mathcal{G} , the combinations of the possible values of (Y_{pq}, Y_{qp}) are restricted to be $(0, 0)$, $(0, y)$ or $(y, 0)$, for some positive integer $y \in \mathbb{N}$. Equivalently, as \mathbf{Y}^σ is upper triangular for a $\boldsymbol{\sigma}$ that is topological, we require that,

for all dyads (p, q) where $1 \leq p < q \leq n$,

$$Y_{qp}^\sigma = 0, \quad Y_{pq}^\sigma | \mathbf{C}, \mathbf{Z}_n, \boldsymbol{\sigma}, \boldsymbol{\xi} \stackrel{\text{ind}}{\sim} \text{Poisson} \left(\xi_p^\sigma \xi_q^\sigma C_{Z_p^\sigma Z_q^\sigma} \right). \quad (2)$$

2.1 Likelihood

With the model established in Equation 2, we can proceed to compute the likelihood. As \mathbf{Z}_n is unknown prior to fitting the model, it will be treated as a vector of latent variables and assigned a prior, of which the parametrisation will be detailed in Section 2.2. However, given this prior choice, several quantities can be defined to facilitate the likelihood derivations. First, there will be K_n distinct values $(Z_1^*, \dots, Z_{K_n}^*)$ appearing in \mathbf{Z}_n with respective frequencies $\mathbf{N} := (N_1, \dots, N_{K_n})$. The i -th element of \mathbf{N} , i.e. $N_i = \sum_{p=1}^n \mathbb{I}(Z_p = Z_i^*) = \sum_{p=1}^n \mathbb{I}(Z_p = i)$ is then the number of vertices in group i . Second, we derive from \mathbf{Z}_n and \mathbf{Y} (or \mathbf{Y}^σ) two $K_n \times K_n$ matrices $\mathbf{E} := (E_{ij})_{1 \leq i, j \leq K_n}$ and $\mathbf{M} := (M_{ij})_{1 \leq i, j \leq K_n}$. The matrix \mathbf{E} is the edge matrix between the groups, where

$$E_{ij} = \sum_{p=1}^n \sum_{q=1}^n Y_{pq} \mathbb{I}(Z_p = i, Z_q = j) = \sum_{p=1}^{n-1} \sum_{q=p+1}^n Y_{pq}^\sigma \mathbb{I}(Z_p^\sigma = i, Z_q^\sigma = j), \quad (3)$$

while in the matrix \mathbf{M} , M_{ij} is the number of dyads (p, q) between groups i and j such that vertex p is topologically in front of vertex q . Mathematically,

$$M_{ij} = \sum_{p=1}^{n-1} \sum_{q=p+1}^n \xi_p^\sigma \xi_q^\sigma \mathbb{I}(Z_{\sigma_p} = i, Z_{\sigma_q} = j) = \sum_{p=1}^{n-1} \sum_{q=p+1}^n \xi_p^\sigma \xi_q^\sigma \mathbb{I}(Z_p^\sigma = i, Z_q^\sigma = j). \quad (4)$$

With all required quantities defined, we will derive the likelihood with observed \mathbf{Y} given \mathbf{Z}_n and $\boldsymbol{\sigma}$. We first check that $\boldsymbol{\sigma}$ is topological, or equivalently \mathbf{Y}^σ is upper triangular, i.e. $Y_{qp}^\sigma = 0$ for all dyads (p, q) where $1 \leq p < q \leq n$, otherwise the likelihood is 0. Once $\boldsymbol{\sigma}$ is

checked to be topological, using Equation (2), the observed data likelihood is

$$\mathbb{P}(\mathbf{Y}|\mathbf{C}, \mathbf{Z}_n, \boldsymbol{\sigma}, \boldsymbol{\xi}) = \mathbb{I}(\mathbf{Y}^\sigma \text{ upper tri.}) \times \bar{Y} \times \prod_{p=1}^n \prod_{q=1}^n (\xi_p \xi_q)^{Y_{pq}} \times \prod_{i=1}^{K_n} \prod_{j=1}^{K_n} e^{-C_{ij} M_{ij}} C_{ij}^{E_{ij}}, \quad (5)$$

where $\bar{Y} = \prod_{p=1}^{n-1} \prod_{q=p+1}^n (Y_{pq}^\sigma!)^{-1}$, and E_{ij} and M_{ij} are given by Equations 3 and 4, respectively. Detailed derivations are in the online Appendix B. Likelihood (5) is influenced by \mathbf{Z}_n and $\boldsymbol{\sigma}$ through the two matrices \mathbf{E} and \mathbf{M} .

2.2 Priors and posterior density

We shall assign independent priors one by one to \mathbf{C} , \mathbf{Z}_n , $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}$, in order to carry out inference within the Bayesian framework. In the subsequent calculations, some additional parameters of the priors used will be included in the notation.

For \mathbf{C} , we assume each C_{ij} is *a priori* independent and identically distributed according to the Gamma(a, b) distribution, where a and b are the positive shape and rate parameters, respectively. This enables \mathbf{C} to be integrated out, to obtain

$$\begin{aligned} \mathbb{P}(\mathbf{Y}|\mathbf{Z}_n, \boldsymbol{\sigma}, \boldsymbol{\xi}, a, b) &= \int \mathbb{P}(\mathbf{Y}|\mathbf{C}, \mathbf{Z}_n, \boldsymbol{\sigma}, \boldsymbol{\xi}) \mathbb{P}(\mathbf{C}|a, b) d\mathbf{C} \\ &= \mathbb{I}(\mathbf{Y}^\sigma \text{ upper tri.}) \times \bar{Y} \times \prod_{p=1}^n \prod_{q=1}^n (\xi_p \xi_q)^{Y_{pq}} \times \left(\frac{b^a}{\Gamma(a)} \right)^{K_n^2} \prod_{i=1}^{K_n} \prod_{j=1}^{K_n} \frac{\Gamma(E_{ij} + a)}{(M_{ij} + b)^{E_{ij} + a}}. \end{aligned} \quad (6)$$

Independent and relatively uninformative gamma prior distributions are assigned to the parameters a and b , as well as the components of $\boldsymbol{\xi}$.

For $\boldsymbol{\sigma}$, we assign a uniform prior to all permutations of $\{1, 2, \dots, n\}$, i.e. $\pi(\boldsymbol{\sigma}) = (n!)^{-1}$. There is no issue with an ordering that is not topological having a positive prior probability, as such an ordering will result in $\mathbb{I}(\mathbf{Y}^\sigma \text{ upper tri.})$ and the likelihood (6) being equal to 0.

For \mathbf{Z}_n , we assume that its components are the first n elements of an exchangeable sequence $(Z_p)_{p \in \mathbb{N}}$ driven by a *Pitman-Yor (PY) process* (Pitman and Yor, 1997), i.e. $Z_p | P \stackrel{iid}{\sim}$

P for all $p \in \mathbb{N}$ and $P \sim \text{PY}(\alpha, \theta, P_0)$. The PY process, reviewed in the online Appendix A.3, is a distribution for an unknown probability distribution P and is parametrised by three hyperparameters (α, θ, P_0) , where P_0 , called base distribution, is a distribution on the sample space, and α and θ are two scalars satisfying either: 1) $0 \leq \alpha < 1$ and $\theta \geq -\alpha$; 2) $\alpha < 0$ and $\theta = k|\alpha|$ for $k \in \mathbb{N}$. The Dirichlet process corresponds to the special case $\alpha = 0$.

2.3 Bayesian model selection

The PY process prior for \mathbf{Z}_n helps select the number of blocks and distinguish between the *infinite* and *finite* regimes. Specifically, in the *infinite regime*, corresponding to $0 \leq \alpha < 1$ and $\theta \geq -\alpha$, P has an infinite number of support points, and the sequence $(Z_p)_{p \in \mathbb{N}}$ will display an infinite number of distinct values. In the *finite regime*, corresponding to $\alpha < 0$ and $\theta = k|\alpha|$ for $k \in \mathbb{N}$, P has k support points, which is also the total number of distinct values in $(Z_p)_{p \in \mathbb{N}}$. When considering the finite regime, i.e. when $\alpha < 0$, we apply the reparametrisation $(\alpha, \theta) \rightarrow (\gamma, k)$, with $\gamma := |\alpha| > 0$ and $k := \theta/|\alpha| \in \mathbb{N}$, and assign a prior to $(\gamma, k) \in \mathbb{R}_+ \times \mathbb{N}$.

As detailed in the online Appendix A.3, it is possible to compute the marginal likelihood (when P is marginalised out) of a sample \mathbf{Z}_n driven by the PY process, which we denoted by $\mathbb{P}(\mathbf{Z}_n | \boldsymbol{\eta}_r)$, where $\boldsymbol{\eta}_r$ is a parameter vector of length 2, dependent on the choice of regime $r \in \{0, 1\}$, either infinite $\boldsymbol{\eta}_0 = (\alpha, \theta)$ or finite $\boldsymbol{\eta}_1 = (\gamma, k)$. This in turn requires the specification of the prior of $\boldsymbol{\eta}_r$ under both regimes. Under the infinite regime, $r = 0$ and $\boldsymbol{\eta}_r = \boldsymbol{\eta}_0 = (\alpha, \theta)$, and we assume that $\alpha \sim \text{Uniform}[0, 1]$ and $\theta + \alpha$ follows a Gamma distribution. This choice of dependent priors for α and θ is done in order to include all possible values in parameter space. Under the finite regime, $r = 1$ and $\boldsymbol{\eta}_r = \boldsymbol{\eta}_1 = (\gamma, k)$, and we assume that γ and k are independent *a priori*, γ follows a Gamma distribution, and k follows a truncated negative binomial distribution, with parameters a_k and b_k , and density

$$\mathbb{P}(k = k' | a_k, b_k) = (1 - b_k^{a_k})^{-1} \times \frac{\Gamma(k' + a_k)}{\Gamma(a_k)k'!} b_k^{a_k} (1 - b_k)^{k'}, \quad k' = 1, 2, \dots$$

where $\Gamma(\cdot)$ is the gamma function. The factor $(1 - b_k^{a_k})^{-1}$ is due to the truncation of 0 from the original support of the negative binomial distribution. The negative binomial is over-dispersed, allowing possibly high variance, hence high prior uncertainty of the value of k .

The regime-dependent parameters and their priors are introduced this way because, ultimately, we want to enable model selection of the regime, which in turn requires the prior of r , denoted by $\mathbb{P}(r)$. The boundary cases $\mathbb{P}(r = 0) = 1$ and $\mathbb{P}(r = 1) = 1$ represent staying within the infinite and finite regimes, respectively, while model selection takes place when $0 < \mathbb{P}(r = 0) < 1$, allowing the data to select the more suitable among the two regimes. The model selection step, implemented within the MCMC sampler, follows the algorithm of Carlin and Chib (1995). Moreover, the prior choice for r allows unbalanced weights between models, as this may improve mixing of the model selection step, as shown in Friel and Pettitt (2008), in the context of reversible jump MCMC. At the end of Section 3.1, in Table 1, we show results from a simulated study in which the posterior sampler in general recovers the correct regime between the finite and infinite ones. A detailed description of the model selection step can be found in the online Appendix C.2.

As all priors required have been specified, the joint posterior of \mathbf{Z}_n , $\boldsymbol{\sigma}$, $\boldsymbol{\eta}_r$, a and b (and r), up to a proportionality constant, is

$$\mathbb{P}(\mathbf{Z}_n, \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\eta}_r, a, b, r | \mathbf{Y}) \propto \mathbb{P}(\mathbf{Y} | \mathbf{Z}_n, \boldsymbol{\sigma}, \boldsymbol{\xi}, a, b) \mathbb{P}(\mathbf{Z}_n | \boldsymbol{\eta}_r) \mathbb{P}(\boldsymbol{\sigma}) \mathbb{P}(\boldsymbol{\xi}) \mathbb{P}(\boldsymbol{\eta}_r | r) \mathbb{P}(a) \mathbb{P}(b) \mathbb{P}(r). \quad (7)$$

We sample from this joint posterior using MCMC, of which the detailed description is available in the online Appendix C.1.

2.4 Posterior point estimate of \mathbf{Z}_n

To provide a posterior point estimate of \mathbf{Z}_n , we follow the clustering approach introduced in Meilă (2007) and further discussed in Wade and Ghahramani (2018). Specifically, the point

estimate, denoted by $\hat{\mathbf{Z}}_n$, is obtained using a decision theoretic approach, by minimizing with respect to the posterior distribution a loss function on the space of allocation vectors,

$$\hat{\mathbf{Z}}_n = \underset{\tilde{\mathbf{Z}}_n}{\operatorname{argmin}} \mathbb{E} \left[L(\mathbf{Z}_n, \tilde{\mathbf{Z}}_n) | \mathbf{Y} \right] = \underset{\tilde{\mathbf{Z}}_n}{\operatorname{argmin}} \sum_{\mathbf{Z}_n} L(\mathbf{Z}_n, \tilde{\mathbf{Z}}_n) \mathbb{P}(\mathbf{Z}_n | \mathbf{Y}). \quad (8)$$

For the loss function $L(\mathbf{Z}_n, \tilde{\mathbf{Z}}_n)$, Meilă (2007) chose the *Variation of Information* (VI), defined as

$$\text{VI}(\mathbf{Z}_n, \tilde{\mathbf{Z}}_n) = \sum_{i=1}^{K_n} \frac{n_{i+}}{n} \log \left(\frac{n_{i+}}{n} \right) + \sum_{j=1}^{\tilde{K}_n} \frac{n_{+j}}{n} \log \left(\frac{n_{+j}}{n} \right) - 2 \sum_{i=1}^{K_n} \sum_{j=1}^{\tilde{K}_n} \frac{n_{ij}}{n} \log \left(\frac{n_{ij}}{n} \right),$$

where $n_{ij} = \sum_{p=1}^n \mathbb{I}(Z_p = i, \tilde{Z}_p = j)$, $n_{i+} = \sum_{j=1}^{\tilde{K}_n} n_{ij}$, and $n_{+j} = \sum_{i=1}^{K_n} n_{ij}$. The loss function $L(\mathbf{Z}_n, \tilde{\mathbf{Z}}_n)$ can be seen as a distance between \mathbf{Z}_n and $\tilde{\mathbf{Z}}_n$, which can be computed even if $K_n \neq \tilde{K}_n$, i.e. the numbers of groups implied by \mathbf{Z}_n and $\tilde{\mathbf{Z}}_n$ are different.

2.5 Identifiability of \mathbf{Z}_n and σ

Our inference approach does not raise any concern of identifiability issues. For \mathbf{Z}_n , the problem of label switching in Bayesian MCMC estimation of finite mixture models and related models is indeed less common in nonparametric models, given that the prior of the block weights is not symmetric (Jasra et al., 2005, Papaspiliopoulos and Roberts, 2008). This issue is further avoided through obtaining the posterior point estimate of \mathbf{Z}_n defined in Equation 8, which minimizes the posterior expectation of a well-chosen loss function. For σ , or equivalently ϕ , the likelihood is not invariant to its permutations as it is an ordering. The absence of identifiability issues for σ is further evidenced in our applications in Section 3, where the posterior estimates closely approximate the known true topological ordering in the simulated examples (Figure 3), and the posterior distributions of ϕ do not look particularly flat for a real dataset (Figure 7).

3 Application

3.1 Simulated data

This section presents an illustrative simulation study in which the proposed DAG-SBM (Equation (2)) is compared with the SBM for a directed graph (Equation (1), referred to as the directed SBM). The purpose of this study is to show how fitting the directed SBM without considering the inherent hierarchical ordering among vertices can result in assigning positive posterior probability to many “impossible” edges, and substantial bias in the estimate of the overall number of blocks in the model.

Upon sampling the parameters from their priors specified in Section 2.2, adjacency matrices are sampled from the DAG-SBM for different sample sizes $n = \{250, 500, 1000\}$ and combinations of the PY parameters, $(\alpha, \theta) = \{(-1, 10), (0, 1), (0.1, 1), (0.6, 1)\}$, thus including both finite and infinite regimes. Both the DAG-SBM and directed SBM are fitted to the resulting adjacency matrices, also under both regimes.

Figure 1 plots the actual adjacency matrix \mathbf{Y}^σ with $n = 250$ (not shown for $n = 500$ and $n = 1000$), ordered according to the true topological ordering, alongside with the posterior point estimates of \mathbf{Y}^σ , for each simulated dataset fit by the directed SBM. The true adjacency matrices in the first column are necessarily upper triangular, as they are reordered according to the true σ . However, from the second and third columns, it is evident that many edges in the lower triangular sections have positive posterior probability, while it should be equal to zero. This is especially true for vertices having either high in-degree or out-degree counts. This is due to the fact that the directed SBM is row and column exchangeable, and vertices with high in-degrees tend to also have high out-degree counts. However, when a hierarchical ordering is present, the model is not exchangeable anymore and the out- and in-degrees of each vertex are negatively correlated, with vertices at the beginning of σ having higher out-degree than in-degree counts, and the opposite for vertices at the end of σ .

Figure 2 displays the posterior histograms of the number of blocks K_n estimated using

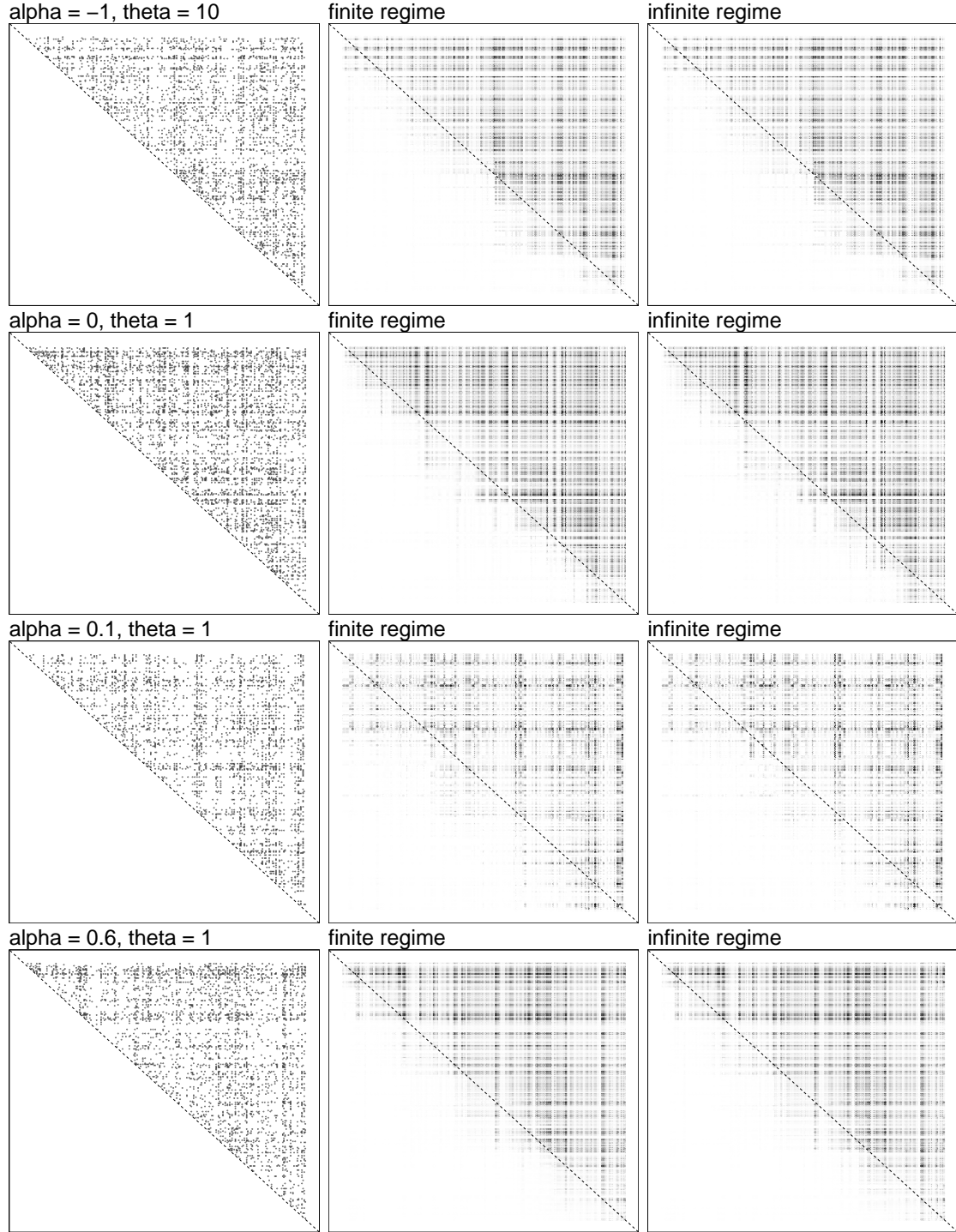


Figure 1: Actual adjacency matrix (left) for data simulated from the DAG-SBM (Equation (2)) with $n = 250$ and different values of (α, θ) , and estimated probabilities by the directed SBM (Equation (1)) under the finite (middle column) and infinite (right) regime, ordered by true σ . The middle and right plots correspond to the results in finite and infinite regimes in the corresponding bottom left panels of Figure 2. The dashed line is the major diagonal.

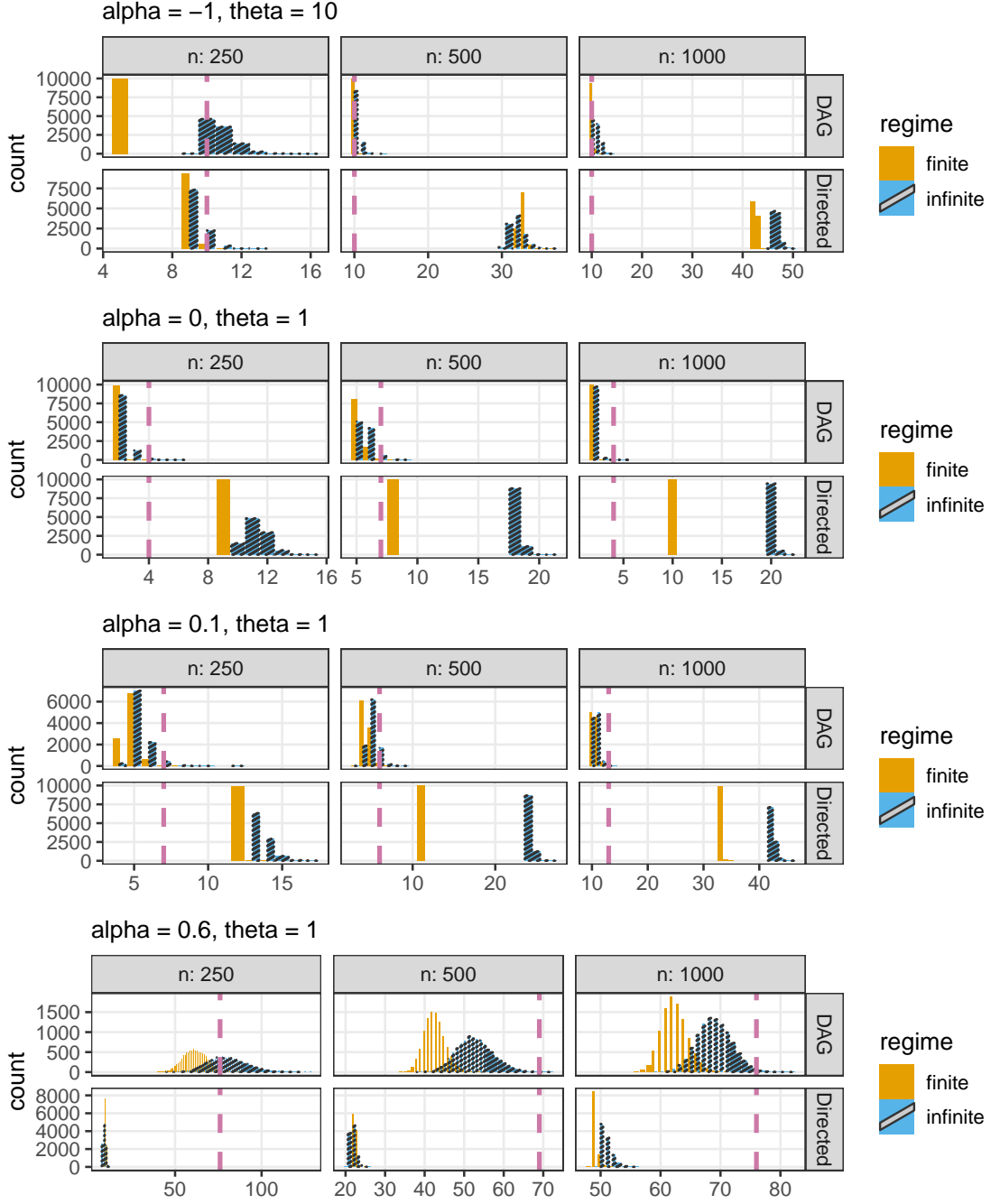


Figure 2: Posterior histogram of K_n for data simulated from the DAG-SBM (Equation (2)), with different values of (α, θ, n) , fitted by both the DAG-SBM (top row) and directed SBM (Equation (1), bottom row), under the finite (solid) and infinite (stripe) regimes. The vertical dashed line is the true value of K_n .

α	θ	Regime	$n = 250$	$n = 500$	$n = 1000$
-1	10	Finite	8.30	612.58	2.72
0	1	Infinite: Logarithm growth	2.56	0.85	19.12
0.1	1	Infinite: Polynomial growth	1.11	0.40	0.89
0.6	1	Infinite: Polynomial growth	7.93×10^{-4}	7.62×10^{-7}	1.92×10^{-11}

Table 1: Estimated Bayes factor for the finite regime to the infinite regime when applying the model selection in the MCMC sampler for the DAG-SBM to the simulated data, for different values of (α, θ, n) . A value above (below) 1 favours the finite (infinite) regime.

the DAG-SBM and the directed SBM, for the same sets of simulated data. On one hand, the directed SBM, which does not include the topological ordering in the likelihood, has the tendency of heavily underestimating or overestimating K_n . The increase in the number of blocks also negatively affects the computational costs of the algorithm, compared to the DAG-SBM. A comparison of the computational times between the two models is provided in the online Appendix D.1. On the other hand, the DAG-SBM provides reasonable estimates of K_n across different sample sizes and for both regimes of the PY parameters. Furthermore, for the DAG-SBM, the model selection step in the MCMC sampler successfully recovers the true regime of the PY parameters. Indeed, Table 1 displays the estimated Bayes factor for the model selection step between the finite and infinite regime, for a range of different values of the hyperparameters of the PY process, including both logarithm (corresponding to the Dirichlet Process, $\alpha = 0$) and polynomial growth ($0 < \alpha < 1$) for K_n . The estimated Bayes factor suggests that the posterior sampler is capable of recovering the correct regime from the data, particularly when α is far from 0, and being generally indifferent between the regimes when α is close to 0, and K_n grows logarithmically, hence very slowly.

Finally, we look at the posterior of σ , or equivalently the positions of the vertices in the topological ordering, ϕ . The mixing of the MCMC is good, and the trace plots are available for some components of ϕ in the online Appendix D. The posterior density of each component of ϕ is plotted as a row in Figure 3, with the rows in the true topological ordering. The concentration of posterior mass on the main diagonal suggests that this latent ordering can be well recovered without identifiability issues.

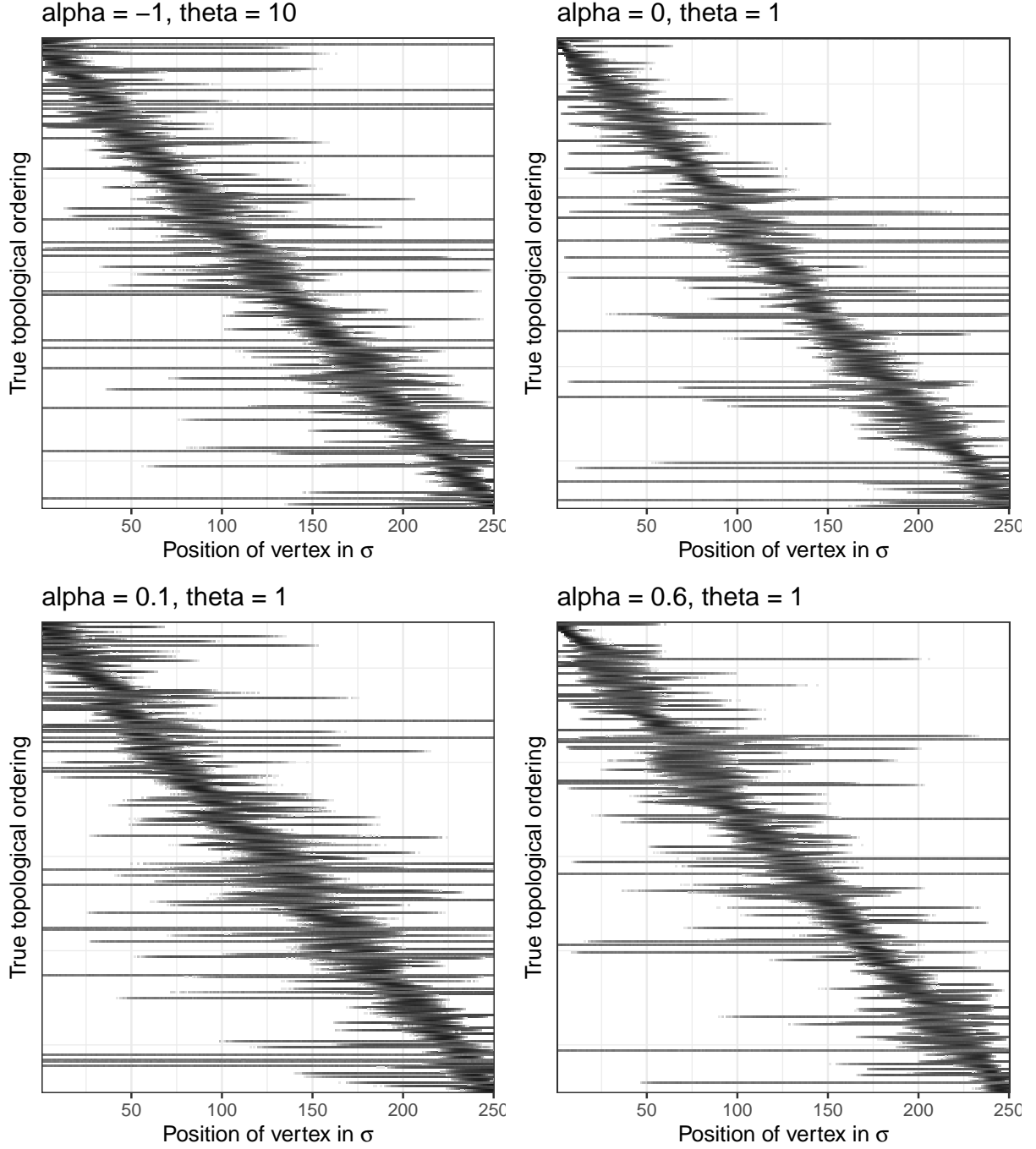


Figure 3: The posterior density of the positions of the vertices in σ for $n = 250$ and different values of (α, θ) for data simulated from the DAG-SBM (Equation (2)), fitted by the DAG-SBM under the true regime according to α . Each row represents the posterior density of a component of ϕ .

3.2 SNA citation network

Next, we look at the citation network analysed by Lee and Wilkinson (2018). It contains 1118 citations (edges) between 135 articles (vertices) which are related to social network analysis (SNA). We shall call this data the SNA citation network hereafter.

The MCMC sampler for the DAG-SBM outlined in the online Appendix C was applied, with 20000 iterations obtained after a burn-in period of 100000 and thinning of 2000, i.e. keeping 1 sample every 2000 iterations, with $a \sim \text{Gamma}(1, 0.01)$, $b \sim \text{Gamma}(1, 0.01)$. Such settings were used three times, once assuming the infinite regime ($r = 0$) with $\theta + \alpha \sim \text{Gamma}(1, 0.01)$, once assuming the finite regime ($r = 1$) with $\gamma \sim \text{Gamma}(1, 0.01)$ and $k \sim \text{truncated negative binomial}(1, 0.01)$, and once for model selection with $\mathbb{P}(r = 1) = 0.2$, i.e. $\mathbb{P}(r = 0) = 0.8$. All three runs were performed on a Linux machine with Intel Core i7-7700 Processor (3.6GHz), and took 0.0089, 0.0077, and 0.0083 seconds per iteration, respectively.

Figures 4 to 8 show some key inference results, except for the trace plots of the parameters, which are in the online Appendix D.2. Of more importance is the posterior density (or mass function in the case of K_n and k) in Figures 4 and 5. On one hand, in the panels for k and γ , the infinite regime is naturally missing, while the results for the finite regime and model selection coincide as expected. Similarly, in the panels for θ and α , the finite regime is naturally missing, while the results for the infinite regime and model selection coincide as expected. On the other hand, the panels for K_n , a and b illustrate that the posterior density from model selection is essentially a weighted average of that of the two regimes.

The departure of the posterior densities of α and γ from 0 suggests that either the infinite regime or the finite regime is preferred to their shared boundary, i.e. the Dirichlet process when $\alpha = \gamma = 0$. Between the two regimes, with $\mathbb{P}(r = 1) = 0.2$ resulting in $\mathbb{P}(r = 1|\mathbf{Y}) = 0.5474$, the Bayes factor $B_{10} = \frac{\mathbb{P}(r = 1|\mathbf{Y})}{\mathbb{P}(r = 0|\mathbf{Y})} \bigg/ \frac{\mathbb{P}(r = 1)}{\mathbb{P}(r = 0)} = 4.8368$, suggesting a slight preference for the finite regime, for this citation network.

The high number of thinning is mainly due to the Markov chain for the finite regime occasionally getting stuck at $K_n = 4$, which is in turn due to the mixing of k and γ , and

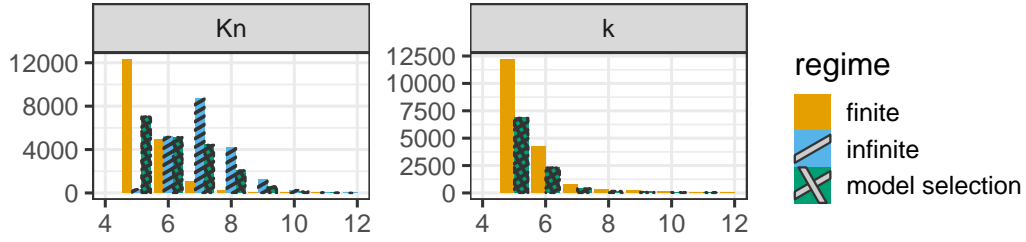


Figure 4: Posterior histogram of K_n and k , for the finite (solid) and infinite (stripe) regimes of the DAG-SBM, and model selection (crosshatch), for the SNA citation network. The right panel is for finite regime and model selection only as k does not exist in the infinite regime. For readability, values of k above 12 are not shown.

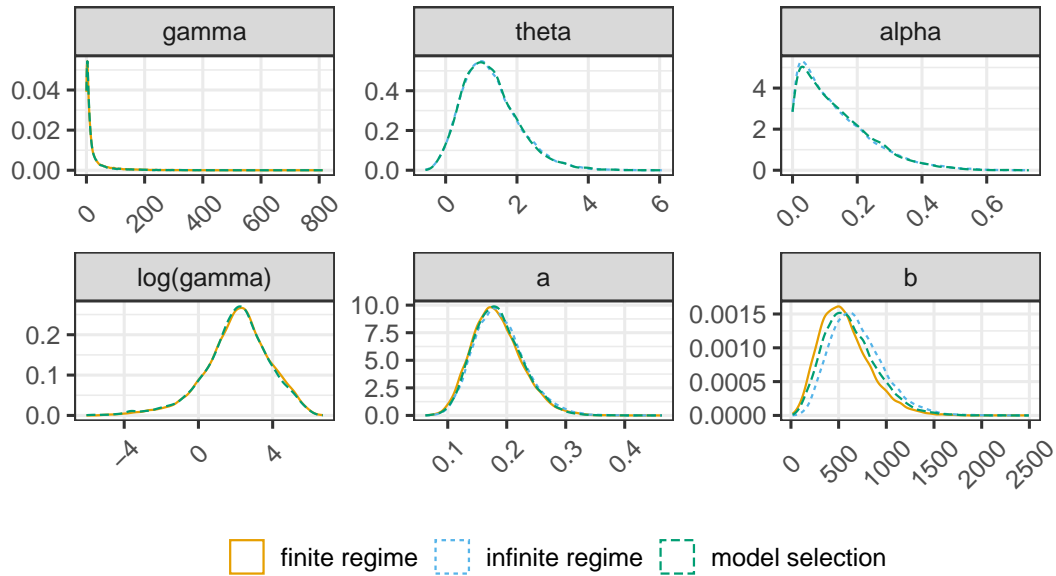


Figure 5: Posterior density of the parameters for the infinite (dotted) and finite (solid) regime of the DAG-SBM, and model selection (dashed), for the SNA citation network.

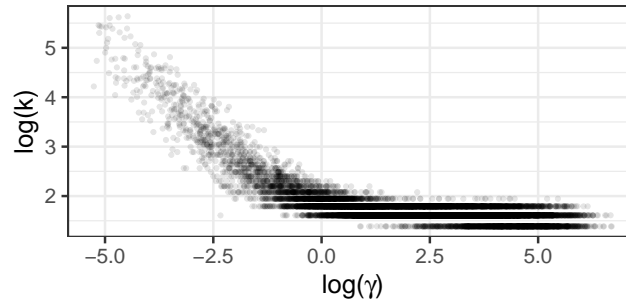


Figure 6: Joint posterior of $\log \gamma$ and $\log k$ for the finite regime of the DAG-SBM, for the SNA citation network.

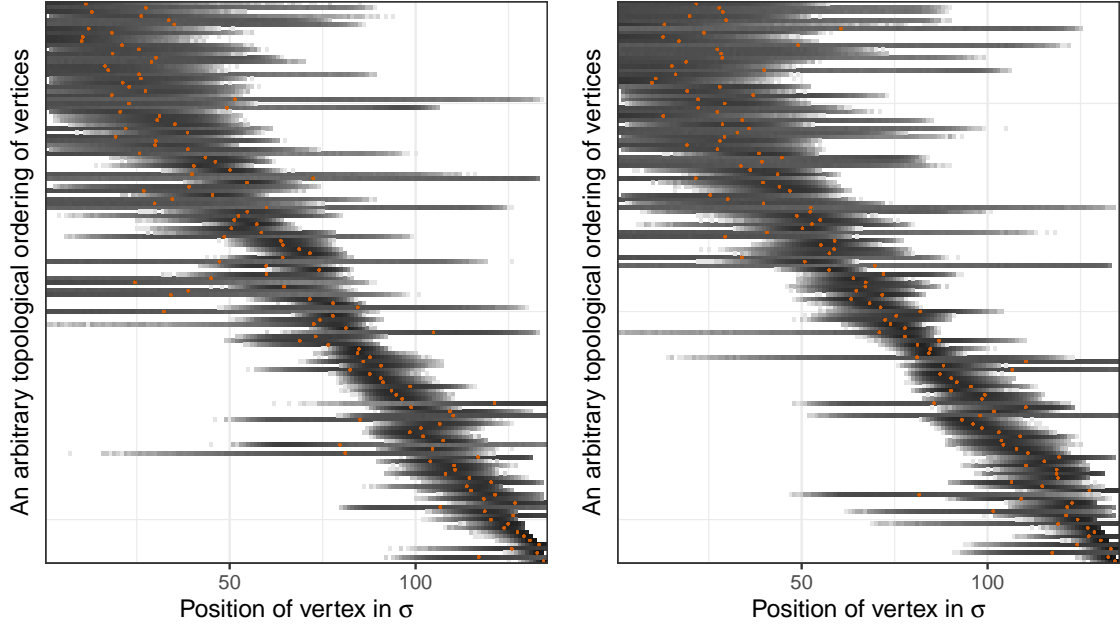


Figure 7: The posterior density of the positions of the vertices in σ for the infinite (left) and finite (right) regime of the DAG-SBM, for the SNA citation network. Each row represents the posterior density of a component of ϕ . The coloured dots are the mean positions.

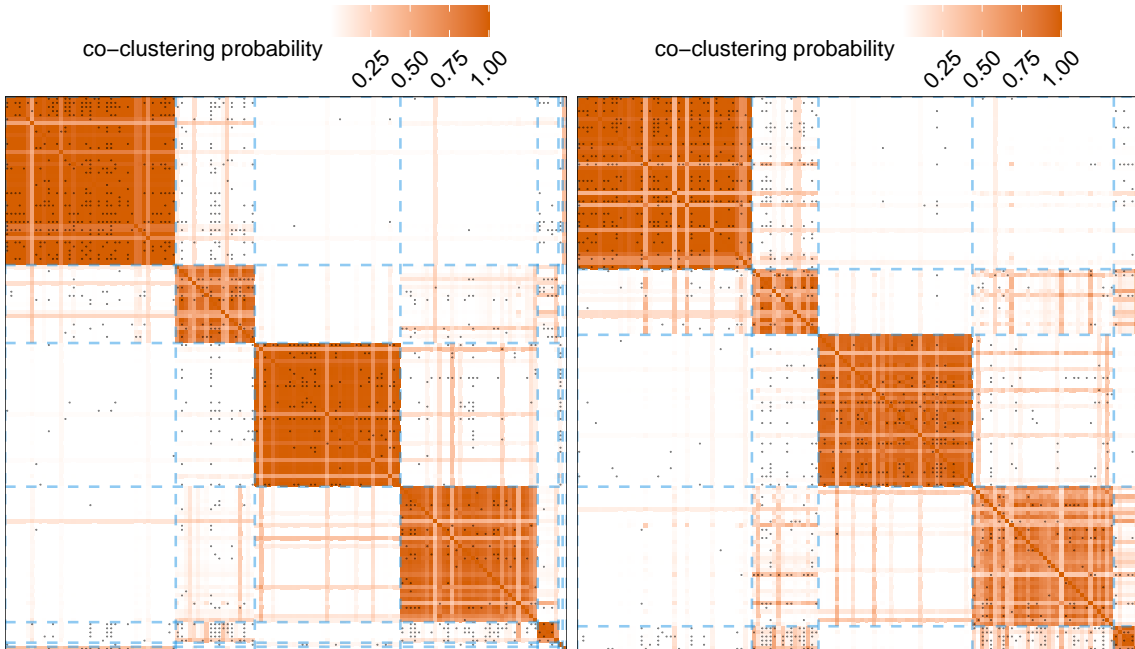


Figure 8: The co-clustering matrix (spectrum) and the adjacency matrix (dots) for the infinite (left) and finite (right) regime of the DAG-SBM, for the SNA citation network. The vertices are clustered (dashed lines) according to the point estimate $\hat{\mathbf{Z}}_n$.

the skewness of their joint posterior, which is plotted on log scale in Figure 6. There is a dense cloud of points towards the bottom right of the plot, indicating high posterior density around small values of k (and high values of γ), thus limiting K_n to grow. However, such issue of getting stuck at small values of k and K_n does not exist for either the infinite regime or the application to the larger dataset in the next subsection.

Next, we look at the posterior of σ , or equivalently the positions of the vertices in the topological ordering, ϕ . The mixing of the MCMC is good, and the trace plots for some components of ϕ are shown in the online Appendix D.2. The posterior density of each component of ϕ is plotted as a row in Figure 7, with the rows themselves in an arbitrary topological ordering. As the citations, i.e. the edges in the DAG in general go from more recent works to older ones, the top (bottom) rows, which correspond to topologically earlier (later) vertices, are generally the more recent (older) articles in the network. The coloured part of each row represents the support of the posterior of the corresponding article, and in general is narrower towards the bottom. An interpretation is that generally chronologically older works are likely to cite between themselves and get cited by later ones, thus limiting their positions in σ . Lastly, using Equation 8 and the SALSO algorithm (Dahl et al., 2022) to minimise the loss function, the point estimate $\hat{\mathbf{Z}}_n$ is obtained, with the co-clustering and adjacency matrices plotted in Figure 8.

3.3 Statistics citation network

We apply the model to a second citation network. The original data analysed by Ji and Jin (2016) contains 5722 citations between 3248 articles in the top statistics journals, from 2003 to the first half of 2012. We only consider the largest connected component, and remove one of the two edges in every pair of cyclic citations (there were only 9 pairs of these). Upon the data cleaning, we arrive at a citation network that is a DAG and contains 5563 citations between 2248 articles. We refer to this data as the statistics citation network hereafter.

The MCMC sampler was applied with 1.5×10^4 iterations obtained after a burn-in period

of 5.5×10^4 and thinning of 10, with the same priors for a , b , α , θ , k and γ as those in Section 3.2. For model selection, even with $\mathbb{P}(r = 0) = 0.99$, the whole chain stays in the finite regime, i.e. $\mathbb{P}(r = 1|\mathbf{Y}) = 1$. Therefore, we shall report the results under the finite regime only, as it is heavily preferred to the infinite counterpart for this network. The trace plots and posterior densities are provided in the online Appendix D.3. The posterior of K_n ranges from 92 to 124, which seems reasonable with around 30 articles per cluster.

Similar to Figure 8, the adjacency matrix for the statistics citation network is plotted in Figure 9, with the vertices clustered according to the point estimate. The co-clustering matrix is not being overlaid here to preserve the figure readability. The concentration of the black dots along the major block diagonal suggests that most groups are closely knitted. On the other hand, there are some concentrated blocks which are off-diagonal and asymmetric, indicating high number of one-way citations from one group to another.

We report several groups of top articles according to median topological ordering for

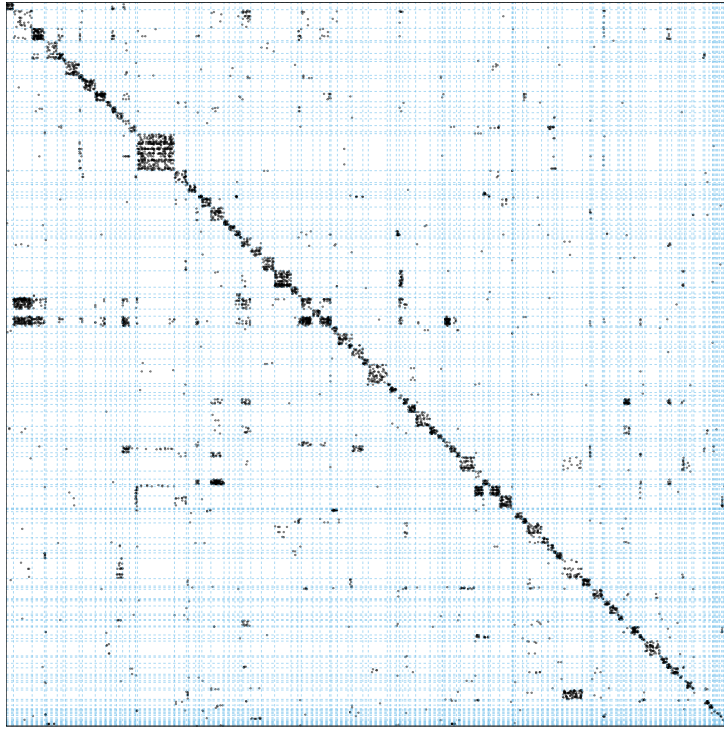


Figure 9: The adjacency matrix (black dots) for the finite regime, for the statistics citation network. The vertices are clustered (dashed lines) according to the point estimate $\hat{\mathbf{Z}}_n$.

some large groups in Table 2. Specifically, for a selection of blocks, we report the top 5 articles in topological ordering within the block. From the article titles, groups seem to roughly correspond to various topics in statistics. For example, group 18 seems to cover false discovery rates, group 45 design and randomisation, group 2 wavelets, group 71 missing/perturbed data, group 25 functional linear regression, and group 65 spatial processes. However, note that the reported group memberships are just point estimates obtained from the MCMC draws, and therefore there might be some possibly large posterior uncertainty

Table 2: Titles of top articles according to topological ordering for selected groups.

Group	Article
18 “False Discovery”	Genovese and Wasserman (AoS, 2004)
18 “False Discovery”	Storey (AoS, 2003)
18 “False Discovery”	Johnstone (AoS, 2008)
18 “False Discovery”	Black (JRSSB, 2004)
18 “False Discovery”	Cox and Wong (JRSSB, 2004)
45 “Design and Randomization”	Lockwood, Schervish, Gurian and Small (JASA, 2004)
45 “Design and Randomization”	Montanari and Ranalli (JASA, 2005)
45 “Design and Randomization”	Rubin (JASA, 2005)
45 “Design and Randomization”	Barnard, Frangakis, Hill and Rubin (JASA, 2003)
45 “Design and Randomization”	Greevy, Silber, Cnaan and Rosenbaum (JASA, 2004)
2 “Wavelet techniques”	Averkamp and Houdré (AoS, 2005)
2 “Wavelet techniques”	Morris, Vannucci, Brown and Carroll (JASA, 2003)
2 “Wavelet techniques”	Yu and Jones (JASA, 2004)
2 “Wavelet techniques”	Foster and Stine (JASA, 2004)
2 “Wavelet techniques”	Signorini and Jone (JASA, 2004)
71 “Missing Data”	Rao, Mingo, Speicher and Edelman (AoS, 2008)
71 “Missing Data”	Li, Aragon, Shedden and Agnan (JASA, 2003)
71 “Missing Data”	Stute, Xue and Zhu (JASA, 2007)
71 “Missing Data”	Liang, Wang, Robins and Carroll (JASA, 2004)
71 “Missing Data”	Chen, Leung and Qin (JASA, 2003)
25 “Functional Regression”	Crambes, Kneip and Sarda (AoS, 2009)
25 “Functional Regression”	Hall and Horowitz (AoS, 2007)
25 “Functional Regression”	Cai and Hall (AoS, 2006)
25 “Functional Regression”	Wu and Liu (JASA, 2007)
25 “Functional Regression”	Mammen and Nielsen (Biometrika, 2003)
65 “Spatial Processes”	Wang and Carey (JASA, 2004)
65 “Spatial Processes”	Schmidt and O’Hagan (JRSSB, 2003)
65 “Spatial Processes”	Schlather, Ribeiro Jr and Diggle (JRSSB, 2004)
65 “Spatial Processes”	McElroy and Politis (AoS, 2007)
65 “Spatial Processes”	Gelfand, Kim, Sirmans and Banerjee (JASA, 2003)

in the topological ordering for some of these articles.

Lastly, we report in Table 3 some top articles according to the median topological ordering, this time without the group memberships. In general, these articles have high number of citations. However, we also notice that some of these articles might not have a very high number of citations. This is due to the fact that these articles are cited by some other influential articles towards the end of the topological ordering.

4 Discussion

In this article, we proposed a Bayesian nonparametric SBM for DAGs. By conditioning on a latent topological ordering, the likelihood of the data (which is composed of directed edges) becomes that of an upper diagonal adjacency matrix. The topological ordering is treated as an unknown parameter, endowed with a prior and inferred a posteriori within the MCMC

Table 3: Titles of top 20 articles according to topological ordering.

Article
Johnstone (AoS, 2008)
Genovese and Wasserman (AoS, 2005)
Zhang, Siegmund, Ji and Li (Biometrika, 2010)
Averkamp and Houdré (AoS, 2005)
Ishwaran and Rao (JASA, 2003)
Zeng (AoS, 2004)
Beskos, Papaspiliopoulos and Roberts (AoS, 2009)
Schlather and Tawn (Biometrika, 2003)
Mammen and Nielsen (Biometrika, 2003)
Wang and Carey (Biometrika, 2003)
Zhang (JASA, 2003)
Huang (AoS, 2004)
Hu and Rosenberger (JASA, 2003)
Ramsay, Hooker, Campbell and Cao (JRSSB, 2007)
Storey (AoS, 2003)
Lee and Pun (JASA, 2006)
Zhu, Miao and Peng (JASA, 2006)
Barnard, Frangakis, Hill and Rubin (JASA, 2003)
Foster and Stine (JASA, 2004)
Lue (Biometrika, 2004)

sampler, using a modified leap-and-shift proposal, previously used for the ranking in the Mallows model by Vitelli et al. (2018). The use of the PY process prior for the allocation vector \mathbf{Z}_n allows the model to infer the number of groups K_n from the data. Moreover, a model selection step for the two regimes of the PY process (one in which the total number of blocks is finite and unknown, and another one in which it increases with the sample size) is incorporated in the MCMC sampler.

The model can be generalised in different ways. For example, it can be extended by introducing covariate information, such as the keywords of each document or its publication year. This could be achieved by modelling the degree correction factors with a covariate dependent distribution. Also, in terms of parametrisation, the two regimes of the PY process could be unified so that γ and α become one parameter that can take a value between $-\infty$ and 1, with its posterior density directly implying which regime is preferred. The main obstacle to overcome here would be the sampling from the non-standard joint parameter space of θ and α across the two regimes. Another issue to be resolved is the inference of k , which is naturally highly correlated with K_n , under the finite regime. This is apparent in the parameter trace plots (in online Appendix D.3) for the statistics citation network, while for the SNA citation network the model selection improves the mixing of k in the MCMC. Ideally k is integrated out, but the computations required mean that this is feasible only under certain special cases. Such issue with k remains to be resolved.

There are potential extensions regarding inference procedure and results. Similar to how $\hat{\mathbf{Z}}_n$ is computed for \mathbf{Z}_n , a point estimate could be provided for $\boldsymbol{\sigma}$, but the distance function for the ordering has to be carefully considered. Relatedly, a Mallow’s model prior (Vitelli et al., 2018) could be used for $\boldsymbol{\sigma}$, as opposed to the uniform prior used here, to potentially provide more information to facilitate the inference. Lastly, the derivation of an efficient Variational Bayes algorithm (Blei et al., 2017) would possibly allow the applicability of the proposed model to much larger datasets.

Funding Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Disclosure Statement

The authors report there are no competing interests to declare.

Supplementary Materials

A single archive, available online, contains all of the following supplementary materials:

Appendices: Background materials of SBM and PY process, likelihood derivations, MCMC algorithms, and additional plots (**appendices.pdf**, pdf file).

R package for simulations & MCMC sampler: R-package “dagsbm” containing code to run the MCMC sampler for the DAG-SBM, and to simulate graphs from the model. The package also contains the dataset of SNA citation network presented in the article. (**dagsbm_0.0.8.tar.gz**, GNU zipped tar file)

Code: Script for running the MCMC sampler using the R package. (**sim_dc.R**, R script)

Results: Pre-generated results from running `sim_dc.R`, for the README and reproducibility checks. (**results/*sim_dc_alpha=0.00_theta=1.00_n=250*.rds**, Rds file)

README: Instructions for reproducing the results in the paper using the R package and code. (**README.Rmd**, R Markdown script)

References

- Abbe, E. (2018), ‘Community detection and stochastic block models: recent developments’, *Journal of Machine Learning Research* **18**, 1–86.
- Blei, D., Kucukelbir, A. and McAuliffe, J. (2017), ‘Variational inference: A review for statisticians’, *Journal of the American Statistical Association* **112**(518), 859–877.
- Carlin, B. P. and Chib, S. (1995), ‘Bayesian model choice via Markov chain Monte Carlo methods’, *Journal of Royal Statistical Society: Series B* **157**(3), 473–484.
- Dahl, D. B., Johnson, D. J. and Müller, P. (2022), ‘Search algorithms and loss functions for Bayesian clustering’, *Journal of Computational and Graphical Statistics* **31**(4), 1189–1201.
- Friel, N. and Pettitt, A. N. (2008), ‘Marginal likelihood estimation via power posteriors’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(3), 589–607.
- Geng, J., Bhattacharya, A. and Pati, D. (2019), ‘Probabilistic community detection with unknown number of communities’, *Journal of the American Statistical Association* **114**(526), 893–905.
- Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983), ‘Stochastic blockmodels: First steps’, *Social Networks* **5**(2), 109–137.
- Jasra, A., Holmes, C. C. and Stephens, D. A. (2005), ‘Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling’, *Statistical Science* **20**(1), 50–67.
- Ji, P. and Jin, J. (2016), ‘Coauthorship and citation networks for statisticians’, *Annals of Applied Statistics* **10**(4), 1779–1812.
- Karrer, B. and Newman, M. E. J. (2011), ‘Stochastic blockmodels and community structure in networks’, *Physical Review E* **83**, 016107.

- 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’, pp. 381–388.
- Lee, C. and Wilkinson, D. J. (2018), ‘A social network analysis of articles on social network analysis’, *ArXiv e-prints* . arXiv:1810.09781.
- Lee, C. and Wilkinson, D. J. (2019), ‘A review of stochastic block models and extensions for graph clustering’, *Applied Network Science* **4**(122).
- Legramanti, S., Rigon, T., Durante, D. and Dunson, D. B. (2022), ‘Extended stochastic block models with application to criminal networks’, *The Annals of Applied Statistics* **16**, 2369–2395.
- Manrique-Vallier, D. and Reiter, J. P. (2014), ‘Bayesian estimation of discrete multivariate latent structure models with structural zeros’, *Journal of Computational and Graphical Statistics* **23**(4), 1061–1079.
- Meilă (2007), ‘Comparing clusterings - an information based distance’, *Journal of Multivariate Analysis* **98**, 561–566.
- Nowicki, K. and Snijders, T. A. B. (2001), ‘Estimation and prediction for stochastic block-structures’, *Journal of the American Statistical Association* **96**(455), 1077–1087.
- Papaspiliopoulos, O. and Roberts, G. O. (2008), ‘Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models’, *Biometrika* **95**(1), 169–186.
- Pitman, J. and Yor, M. (1997), ‘The two parameter Poisson-Dirichlet distribution derived from a stable subordinator’, *Annals of Probability* **25**, 855–900.
- Schmidt, M. N. and Mørup, M. (2013), ‘Nonparametric Bayesian modeling of complex networks: An introduction’, *IEEE Signal Processing Magazine* **30**, 110–128.

- Snijders, T. A. B. and Nowicki, K. (1997), ‘Estimation and prediction for stochastic block-models for graphs with latent block structure’, *Journal of Classification* **14**(1), 75–100.
- Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A. and Arjas, E. (2018), ‘Probabilistic preference learning with the Mallows rank model’, *Journal of Machine Learning Research* **18**(158), 1–49.
- Wade, S. and Ghahramani, Z. (2018), ‘Bayesian cluster analysis: Point estimation and credible balls’, *Bayesian Analysis* **13**, 559–626.
- Wang, Y. J. and Wong, G. Y. (1987), ‘Stochastic blockmodels for directed graphs’, *Journal of the American Statistical Association* **82**, 8–19.