Statistics and Computing
https://doi.org/10.1007/s11222-020-09946-6

Bayesian estimation of the latent dimension and communities in stochastic blockmodels

Francesco Sanna Passino[1] · **Nicholas A. Heard**[1]

Received: 12 July 2019 / Accepted: 10 May 2020
© The Author(s) 2020

Abstract
Spectral embedding of adjacency or Laplacian matrices of undirected graphs is a common technique for representing a network in a lower dimensional latent space, with optimal theoretical guarantees. The embedding can be used to estimate the community structure of the network, with strong consistency results in the stochastic blockmodel framework. One of the main practical limitations of standard algorithms for community detection from spectral embeddings is that the number of communities and the latent dimension of the embedding must be specified in advance. In this article, a novel Bayesian model for simultaneous and automatic selection of the appropriate dimension of the latent space and the number of blocks is proposed. Extensions to directed and bipartite graphs are discussed. The model is tested on simulated and real world network data, showing promising performance for recovering latent community structure.

Keywords Community detection · Gaussian mixture modelling · Random dot product graph · Spectral embedding · Stochastic blockmodel

1 Introduction

A network can be represented as a graph $\mathbb{G} = (V, E)$, where V is a set of nodes and $E \subseteq V \times V$ is a set of edges indicating the pairs of nodes which have interacted. The graph can be characterised by the adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, where $n = |V|$ and for $1 \leq i, j \leq n$, $A_{ij} = \mathbb{1}_E\{(i, j)\}$, such that $A_{ij} = 1$ if a link between the nodes i and j exists, and $A_{ij} = 0$ otherwise. The graph is said to be undirected if $(i, j) \in E \iff (j, i) \in E$ and $\mathbf{A}$ is constrained to be symmetric; otherwise, the graph is said to be directed. It will be assumed that a node cannot link to itself, implying $\mathbf{A}$ is a hollow matrix.

Latent space models (Hoff et al. 2002) represent a flexible approach to statistical analysis of networks: each node i is assigned a latent position $\boldsymbol{x}_i$ in a d-dimensional latent space $\mathbb{X}$, and edges between pairs of nodes are typically generated independently, with the probability of observing a link between nodes i and j obtained through a *kernel* function $\psi : \mathbb{X} \times \mathbb{X} \to [0, 1]$ of the respective latent positions: $\mathbb{P}(A_{ij} = 1) = \psi(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Different ideas and techniques for embedding observed graphs into low dimensional spaces are explored in the literature (for a survey, see, for example, Cai et al. 2018). *Random dot product graphs* (RDPGs) (Young and Scheinerman 2007) are a popular class of latent position models, where $\mathbb{X} \subseteq \mathbb{R}^d$, and the function $\psi(\cdot)$ is an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{X} \times \mathbb{X}$. RDPGs are analytically tractable and have therefore been extensively studied; a survey of the existing statistical inference techniques is presented in Athreya et al. (2017).

The *stochastic blockmodel* (SBM) (Holland et al. 1983) is the classical statistical model for clustering graphs (Snijders and Nowicki 1997; Nowicki and Snijders 2001). Assuming K communities, each node is assigned a community membership $z_i \in \{1, \ldots, K\}$ with probabilities $\boldsymbol{\theta}$, from the $K - 1$ probability simplex. The probability of a link only depends on the community allocations z_i and z_j of the two nodes. Given a symmetric matrix $\mathbf{B} \in [0, 1]^{K \times K}$ of inter-community probabilities, then independently $\mathbb{P}(A_{ij} = 1) = B_{z_i z_j}$. Stochastic blockmodels have appealing statistical properties, and can well approximate any independent-edge network model if

The authors gratefully acknowledge funding from the EPSRC and the Heilbronn Institute for Mathematical Research.

✉ Francesco Sanna Passino
francesco.sanna-passino16@imperial.ac.uk

Nicholas A. Heard
n.heard@imperial.ac.uk

[1] Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom

Published online: 27 May 2020

the number of communities is sufficiently large (Bickel and Chen 2009; Wolfe and Olhede 2013). Stochastic blockmodels can also be easily represented as random dot product graphs: each community is assigned a latent position, which is common to all the nodes belonging to the cluster, and $\mathbf{B}$ is obtained from the inner products of those positions. Hence, in this framework, $d = \text{rank}(\mathbf{B}) \leq K$.

Spectral clustering (von Luxburg 2007) provides a consistent statistical estimation procedure for the latent positions and communities in SBMs (Rohe et al. 2011; Lei and Rinaldo 2015) and more generally in random dot product graphs (Tang et al. 2013; Sussman et al. 2014). Rubin-Delanchy et al. (2017) directly links adjacency and Laplacian spectral embedding to the generalised random dot product graph (GRDPG), an extension of the RDPG, and advocates for Gaussian mixture modelling (GMM) of the rows of the embedding. Alternatives to spectral clustering include variational methods (Celisse et al. 2012) and pseudo-likelihood approaches (Amini et al. 2013). SBMs have been extended to the directed case (Wang and Wong 1987; Rohe et al. 2016), and appropriate embeddings for co-clustering, in most cases based on the singular value decomposition (SVD), have been derived in the literature (Dhillon 2001; Malliaros and Vazirgiannis 2013; Zheng and Skillicorn 2015).

One of the practical issues of spectral embedding, and in general most graph embedding algorithms, is that it requires a suitable prespecified latent dimensionality d (usually $d \ll |V|$) as input, and, subsequently, a suitable number of clusters K, conditionally, crucially, on the previous choice of d. For a practical example of this procedure on a real world network, see Priebe et al. (2019). In general, in spectral clustering, similarly to what practitioners do in principal component analysis (PCA), the investigator examines the scree-plot of the eigenvalues and chooses the dimension based on the location of *elbows* in the plot (Jolliffe 2002), or uses the *eigengap* heuristic (see, for example, von Luxburg 2007). Automatic methods for thresholding have also been suggested (Zhu and Ghodsi 2006; Chatterjee 2015). A relevant body of literature is also devoted to methods for the selection of the number of communities in stochastic blockmodels (Zhao et al. 2011; Bickel and Sarkar 2016; Newman and Reinert 2016; Franco Saldaña et al. 2017; Chen and Lei 2018). Often, practitioners simply set $d = K$, for some d, assuming that $\mathbf{B}$ has full rank in the stochastic blockmodel framework. Under the full rank assumption, one may estimate $d = K$ as the number of eigenvalues of $\mathbf{A}$ which are larger than $\sqrt{n}$ (Chatterjee 2015; Lei 2016). In this work, the problem of selecting d is approached from the perspective of variable selection in model based clustering, which is widely studied in the literature (Fowlkes et al. 1988; Law et al. 2004; Tadesse et al. 2005; Raftery and Dean 2006; Maugis et al. 2009; Witten and Tibshirani 2010). Similarly, the problem of correctly selecting the number of clusters is also common in K-means or GMMs, since it is usually required to specify a number of components in the mixture. Usually the parameter is chosen by minimising information criteria (for example, AIC or BIC). A widely used selection criterion is the Integrated Classification Likelihood (ICL) of Biernacki et al. (2000). Exact versions of the ICL based on the adoption of prior conjugated distributions of the model parameters have been obtained in the literature (for example, Côme and Latouche 2015; Wyse et al. 2017). Numerous other techniques have also been proposed for GMMs with unknown number of components (Mengersen et al. 1996; Richardson and Green 1997; Stephens 2000; Nobile 2004; Dellaportas and Papageorgiou 2006; Miller and Harrison 2018).

Clearly, the sequential approach in estimating d and K is suboptimal, and it is desirable to jointly estimate the two parameters, a problem which is not explored in the literature. This article addresses the problem in a Bayesian framework, proposing a novel methodology to automatically select d and K, simultaneously. Techniques for selection of K in GMMs will be incorporated within the spectral embedding framework, allowing for K and d, the number of communities and latent dimension of the latent positions, to be random and learned from the data. A structured Bayesian model for simultaneously inferring the dimension of the latent space, the number of communities, and the community allocations is proposed. The model is based on asymptotic results (Athreya et al. 2016; Rubin-Delanchy et al. 2017; Tang and Priebe 2018) on the leading components of spectral embeddings, obtained for d fixed and known. The asymptotic theory is combined with realistic assumptions about the remaining components of the embedding, empirically tested and justified on simulated data. Furthermore, extensions to the directed and bipartite case will be discussed. The proposed model has multiple advantages: the latent dimension d and number of communities K are modelled separately, and the Bayesian framework allows for automatic selection of the two parameters. The model also allows estimation of d even when $d < K$, and gives insights on the goodness-of-fit of the stochastic blockmodel on observed network data, based on the embedding. The method is tested on simulated data and applied to real world computer and transportation networks. It should be noted that Yang et al. (2019) have simultaneously and independently proposed a similar inferential procedure within a frequentist framework.

The article is organised as follows: Sect. 2 introduces adjacency spectral and Laplacian embeddings and the GRDPG. The novel Bayesian model for selection of the appropriate dimension of the latent space is discussed in Sect. 3, followed by careful illustration of posterior inference procedures in Sect. 4. Section 5 discusses the effects of the curse of dimensionality on the model, and suggests a remedy. Extensions of the model are presented in Sect. 6, and results and applications are finally discussed in Sect. 7.

2 GRDPG and spectral embeddings

In this work, the stochastic block model will be interpreted as a specific case of a generalised random dot product graph (GRDPG) (Rubin-Delanchy et al. 2017). For $d > 0$, let d_+, d_- be non-negative integers such that $d_+ + d_- = d$. Let $\mathbb{X} \subseteq \mathbb{R}^d$ such that $\forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}, 0 \leq \boldsymbol{x}^\top \mathbf{I}(d_+, d_-)\boldsymbol{x}' \leq 1$, where

$$\mathbf{I}(p, q) = \text{diag}(1, \ldots, 1, -1, \ldots, -1)$$

with p ones and q minus ones. Let $\mathcal{F}$ be a probability measure on $\mathbb{X}$, $\mathbf{A} \in \{0, 1\}^{n\times n}$ be a symmetric matrix and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{X}^n$. Then $(\mathbf{A}, \mathbf{X}) \sim \text{GRDPG}_{d_+,d_-}(\mathcal{F})$ if $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{iid}{\sim} \mathcal{F}$ and for $i < j$, independently,

$$\mathbb{P}(A_{ij} = 1) = \boldsymbol{x}_i^\top \mathbf{I}(d_+, d_-)\boldsymbol{x}_j.$$

To represent the K-community stochastic blockmodel with inter-community probabilities $\mathbf{B}$ as a GRDPG, $\mathcal{F}$ can be chosen to have mass concentrated on atoms $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^d$ such that $\boldsymbol{\mu}_i^\top \mathbf{I}(d_+, d_-)\boldsymbol{\mu}_j = B_{ij}\ \forall\, i, j \in \{1, \ldots, K\}$. For consistent estimation of the latent positions in a SBM, interpreted as a GRDPG, Rubin-Delanchy et al. (2017) suggest to fit a Gaussian mixture model with K components to the d-dimensional adjacency or Laplacian spectral embedding.

Adjacency spectral embedding (ASE) and Laplacian spectral embedding (LSE) are two common techniques to embed the adjacency matrix of an undirected graph into a latent space of dimension d. Suppose $\mathbf{A} \in \{0, 1\}^{n\times n}$ is a symmetric adjacency matrix of an undirected graph with n nodes. Then:

Definition 1 (ASE – Adjacency spectral embedding) For $d \in \{1, \ldots, n\}$, consider the spectral decomposition $\mathbf{A} = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Gamma}}^\top + \hat{\mathbf{\Gamma}}_\perp\hat{\mathbf{\Lambda}}_\perp\hat{\mathbf{\Gamma}}_\perp^\top$, where $\hat{\mathbf{\Lambda}}$ is a $d \times d$ diagonal matrix containing the top d eigenvalues in magnitude, in decreasing order, $\hat{\mathbf{\Gamma}}$ is a $n \times d$ matrix containing the corresponding orthonormal eigenvectors, and the matrices $\hat{\mathbf{\Lambda}}_\perp$ and $\hat{\mathbf{\Gamma}}_\perp$ contain the remaining $n - d$ eigenvalues and eigenvectors. The adjacency spectral embedding $\hat{\mathbf{X}} = [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_n]^\top$ of $\mathbf{A}$ in $\mathbb{R}^d$ is $\hat{\mathbf{X}} = \hat{\mathbf{\Gamma}}|\hat{\mathbf{\Lambda}}|^{1/2} \in \mathbb{R}^{n\times d}$, where the operator $|\cdot|$ applied to a matrix returns the absolute value of its entries.

Definition 2 (LSE – Laplacian spectral embedding) For $d \in \{1, \ldots, n\}$, consider the Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, $\mathbf{D} = \text{diag}(\sum_{j=1}^n A_{ij})$, and its spectral decomposition $\mathbf{L} = \tilde{\mathbf{\Gamma}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Gamma}}^\top + \tilde{\mathbf{\Gamma}}_\perp\tilde{\mathbf{\Lambda}}_\perp\tilde{\mathbf{\Gamma}}_\perp^\top$. The Laplacian spectral embedding $\tilde{\mathbf{X}} = [\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n]^\top$ of $\mathbf{A}$ in $\mathbb{R}^d$ is $\tilde{\mathbf{X}} = \tilde{\mathbf{\Gamma}}|\tilde{\mathbf{\Lambda}}|^{1/2}$.

The modified Laplacian $\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ (Rohe et al. 2011) is preferred to $\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ since the eigenvalues of the former lie in $(-1, 1)$, providing a convenient interpretation for disassortative networks (Rubin-Delanchy et al. 2016).

Intuitively, the estimation procedure proposed by Rubin-Delanchy et al. (2017) holds because, taking a graph with m nodes, and restricting the attention to the first n nodes, with $n < m$, the following central limit theorem holds: $\mathbf{Q}_m\hat{\boldsymbol{x}}_i \longrightarrow \mathbb{N}\{\boldsymbol{\mu}_{z_i}, m^{-1}\mathbf{\Sigma}(\boldsymbol{\mu}_{z_i})\}$ in distribution for $m \to \infty$, $i = 1, \ldots, n$, where $\mathbf{Q}_m$ is a matrix from the indefinite orthogonal group $\mathbb{O}(d_+, d_-)$ and $\mathbf{\Sigma}(\boldsymbol{\mu}_{z_i})$ can be analytically computed (Rubin-Delanchy et al. 2017). The result holds for d fixed and known, but in this work it is of interest to treat d as a random, unknown parameter. If a m-dimensional embedding is considered, with $m > d$, then asymptotic theory implies an approximate normal distribution with non-zero means and a full covariance within each cluster for the top-d components of the embedding; but, *to the best of our knowledge*, no theoretical results have been obtained for the remaining $m-d$ columns. It is therefore necessary to propose a model for the remaining part of the embedding, which will be carefully described in Sect. 3, and empirically justified and assessed in Sect. 7.1.

3 A Bayesian model for SBM embeddings

For simplicity, the embeddings will be generically denoted as $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times m}$, $\boldsymbol{x}_i \in \mathbb{R}^m$ for some m, $d \leq m \leq n$. In this article, m is always assumed to be fixed and obtained from a preprocessing step. Choosing an appropriate value of m is arguably much easier than choosing the correct d, and, in the proposed model, the correct d can be recovered for any choice of m, as long as $d \leq m$. Let $\mathbf{X}_{:d}$ denote the first d columns of $\mathbf{X}$, and $\mathbf{X}_{d:}$ the $m - d$ remaining columns. The notation $\boldsymbol{x}_{i,:d}$ denotes the first d elements $(x_1, \ldots, x_d)$ of the vector $\boldsymbol{x}_i$, and similarly $\boldsymbol{x}_{i,d:}$ denotes the last $m - d$ elements $(x_{d+1}, \ldots, x_m)$. Suppose a latent dimension d, K communities, and latent community assignments $\boldsymbol{z} = (z_1, \ldots, z_n)$. The latent positions of nodes within each community are assumed to be generated from an m-dimensional community specific Gaussian distribution:

$$\boldsymbol{x}_i | d, z_i, \boldsymbol{\mu}_{z_i}, \mathbf{\Sigma}_{z_i}, \sigma^2_{z_i} \sim \mathbb{N}_m\left(\begin{bmatrix}\boldsymbol{\mu}_{z_i}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{z_i} & \mathbf{0}\\ \mathbf{0} & \sigma^2_{z_i}\mathbf{I}_{m-d}\end{bmatrix}\right). \quad (1)$$

Following the results in Sect. 2, the initial components $\boldsymbol{x}_{i,:d}$ are assumed to have unconstrained mean vector $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and positive definite $d \times d$ covariance matrix $\mathbf{\Sigma}_k$. In contrast, for $\boldsymbol{x}_{i,d:}$, two constraints are imposed: the mean is an $(m - d)$-dimensional vector of zeros, and the covariance is a diagonal matrix $\sigma_k^2\mathbf{I}_{m-d}$ with positive entries $\boldsymbol{\sigma}_k^2 = (\sigma^2_{k,d+1}, \ldots, \sigma^2_{k,m})$. The validation of the model assumptions will be discussed in details in Sect. 7.1. Assuming group-specific covariances $\boldsymbol{\sigma}_k^2$ for $\mathbf{X}_{d:}$ adds extra complexity to the

model, and implies that d cannot take the simple interpretation of being the number of dimensions relevant for clustering (see, for example, Raftery and Dean 2006). However, empirical evidence, discussed in Sects. 5 and 7.1, suggests that the last $m - d$ components of the embedding contain a cluster structure which reflects the communities in the top-d dimensions. Following the discussion in the previous sections, the parameter d actually represents the dimension of the latent positions that generate the network, equivalent to the rank of the unobserved matrices $\mathbb{E}(\mathbf{A})$ and $\mathbf{B} \in [0, 1]^{K \times K}$.

In order to complete the model specification, conjugate priors can be placed on the parameters as follows:

$$\begin{aligned}
(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)|d &\overset{iid}{\sim} \text{NIW}_d(\mathbf{0}, \kappa_0, \nu_0 + d - 1, \boldsymbol{\Delta}_d), \\
\sigma^2_{k,j} &\overset{iid}{\sim} \text{Inv-}\chi^2(\lambda_0, \sigma^2_0), \;\; j = d+1, \ldots, m, \\
d|z &\sim \text{Uniform}\{1, \ldots, K_\varnothing\}, \\
z_i|\boldsymbol{\theta} &\overset{iid}{\sim} \text{Multinoulli}(\boldsymbol{\theta}), \;\; i = 1, \ldots, n, \\
\boldsymbol{\theta}|K &\sim \text{Dirichlet}\,(\alpha/K, \ldots, \alpha/K)\,, \\
K &\sim \text{Geometric}(\omega),
\end{aligned} \tag{2}$$

where, if $n_k = \sum_{i=1}^n \mathbb{1}_k\{z_i\}$ is the size of community k, $K_\varnothing = \sum_{k=1}^K \mathbb{1}_{\mathbb{N}_+}\{n_k\}$ is the number of non-empty communities. In (1), NIW_d denotes the d-dimensional normal inverse Wishart distribution, and Inv-χ^2 is a scaled inverse chi-square distribution with λ_0 degrees of freedom and scaling parameter σ^2_0.

Yang et al. (2019) have simultaneously and independently proposed a model similar to (1) in a frequentist framework, reaching similar assumptions. The conjecture on the distribution of $\mathbf{X}_{d:}$ is essentially the same, except for the choice of the diagonal elements of the cluster-specific covariance matrix: Yang et al. (2019) use a common variance parameter σ^2_k for the last $m - d$ columns of the embedding, whereas a $(m - d)$-dimensional vector of variances $\boldsymbol{\sigma}^2_k$ is used in this paper. Additionally, as a second difference from Yang et al. (2019), the full model proposed here will also incorporate a second-level community cluster structure on these vectors of variances, which will be introduced in Sect. 5.

Note that the condition $d \leq K$ is explicitly enforced in (1). More specifically, $d \leq K_\varnothing$, which avoids an artificial matching between d and K using empty clusters, which are given non-zero probability mass under the Dirichlet-Multinoulli prior on $(\boldsymbol{\theta}, z)$. One can also model d and K separately in an analogous way, changing the prior $p(d)$ to, for example,

$$d \sim \text{Geometric}(\delta), \tag{3}$$

independently of K and z; this will later be referred to as the unconstrained model. The alternative prior (3) is particularly useful in practical applications and provides a useful interpretation of d: when $d \leq K$, then $d = \text{rank}(\mathbf{B})$, but when $d > K$, this implies that the observed data might deviate from the stochastic blockmodel assumption, and provides a useful diagnostic for model validation and goodness-of-fit.

The likelihood associated with the spectral embedding $\mathbf{X} \in \mathbb{R}^{n \times m}$ obtained from a stochastic blockmodel can be expressed as:

$$L(\mathbf{X}) = \prod_{i=1}^n \left\{ \phi(\boldsymbol{x}_{i,:d}; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \prod_{j=d+1}^m \phi(x_{i,j}; 0, \sigma^2_{z_i,j}) \right\},$$

where $\phi(\cdot)$ denotes the (possibly multivariate) Gaussian density function. Hence, the posterior distribution, up to a normalising constant, has form

$$\begin{aligned}
&p(\{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\boldsymbol{\sigma}^2_k\}, z, \boldsymbol{\theta}, K, d|\mathbf{X}) \propto L(\mathbf{X}) \times \\
&\prod_{k=1}^K \left\{ p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|d) \prod_{j=d+1}^m p(\sigma^2_{k,j}|d) \right\} \prod_{i=1}^n p(z_i|\boldsymbol{\theta}) p(K) p(d).
\end{aligned}$$

The $\text{NIW}_d(\mathbf{0}, \kappa_0, \nu_0 + d - 1, \boldsymbol{\Delta}_d)$ prior on the pair $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is conjugate and yields a conditional posterior $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)|\mathbf{X}, z, d \sim \text{NIW}_d(\boldsymbol{m}^{(k)}_{:d}, \kappa_{n_k}, \nu_{n_k} + d - 1, \mathbf{D}^{(k)}_{:d})$. By standard methods for inference in a multivariate Gaussian mixture model with NIW prior, the covariance matrix $\boldsymbol{\Sigma}_k$ can be explicitly integrated out from the posterior to obtain the marginal $p(\boldsymbol{\mu}_k|\mathbf{X}, z, d) = t_{\nu_{n_k}}\{\boldsymbol{\mu}_k|\boldsymbol{m}^{(k)}_{:d}, \mathbf{D}^{(k)}_{:d}/(\kappa_{n_k}\nu_{n_k})\}$, the density of the multivariate Student t distribution with ν_{n_k} degrees of freedom, where $\nu_{n_k} = \nu_0 + n_k$, $\kappa_{n_k} = \kappa_0 + n_k$,

$$\begin{aligned}
\boldsymbol{m}^{(k)}_{:d} &= \textstyle\sum_{i:z_i=k} \boldsymbol{x}_{i,:d} \big/ \kappa_{n_k}, \\
\mathbf{D}^{(k)}_{:d} &= \boldsymbol{\Delta}_d + \textstyle\sum_{i:z_i=k} \boldsymbol{x}_{i,:d}\boldsymbol{x}^\top_{i,:d} - \kappa_{n_k}\boldsymbol{m}^{(k)}_{:d}\boldsymbol{m}^{(k)\top}_{:d}.
\end{aligned} \tag{4}$$

Henceforth, $\boldsymbol{\mu}_k$ can easily be resampled in a simple Gibbs sampling step, conditional on the actual value of d and on the community allocations z. In this work, the location vectors $\boldsymbol{\mu}_k$ are collapsed out too, but the distribution is instructive to present other distributional results below, and could be also used if the objective of the analysis is to also recover the explicit form of the latent positions.

In a multivariate Gaussian model with conjugate NIW prior, it is also possible to analytically express the marginal likelihood of the observed data, conditioning on a community specific Gaussian, on the assignments z and on the dimension of the latent space d:

$$\begin{aligned}
p\left(\mathbf{X}^{(k)}_{:d} \middle| d, z\right) = \pi^{-n_k d/2} &\frac{\kappa_0^{d/2}|\boldsymbol{\Delta}_d|^{(\nu_0+d-1)/2}}{\kappa_{n_k}^{d/2}|\mathbf{D}^{(k)}_{:d}|^{(\nu_{n_k}+d-1)/2}} \\
&\times \prod_{i=1}^d \frac{\Gamma\{(\nu_{n_k} + d - i)/2\}}{\Gamma\{(\nu_0 + d - i)/2\}},
\end{aligned} \tag{5}$$

where $\mathbf{X}^{(k)}_{:d}$ is the subset of rows of $\mathbf{X}_{:d}$ such that $z_i = k$.

Given the Inv-$\chi^2(\lambda_0, \sigma_0^2)$ prior, the posterior for $\sigma^2_{j,k}$ is Inv-$\chi^2(\lambda_{n_k}, s_j^{(k)})$, where $\lambda_{n_k} = \lambda_0 + n_k$, and $s_j^{(k)} = \left\{\lambda_0\sigma_0^2 + \sum_{i:z_i=k} x^2_{i,j}\right\}/\lambda_{n_k}$. Similar calculations give the full marginal likelihood for the remaining portion of the embedding $\mathbf{X}^{(k)}_{d:}$:

$$p\left(\mathbf{X}^{(k)}_{d:}\middle| d, z\right) = \pi^{-n_k(m-d)/2}\left\{\frac{\Gamma(\lambda_{n_k}/2)}{\Gamma(\lambda_0/2)}\right\}^{m-d} \times \prod_{j=d+1}^{m} \frac{(\lambda_0\sigma_0^2)^{\lambda_0/2}}{\left(\lambda_{n_k} s_j^{(k)}\right)^{\lambda_{n_k}/2}}. \tag{6}$$

Also, note that the probabilities $\boldsymbol{\theta}$ associated to the community assignment can be easily integrated out, resulting in the following marginal likelihood, conditional on K:

$$p(z|K) = \frac{\Gamma(\alpha)\prod_{k=1}^{K}\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)^K\Gamma(n+\alpha)}. \tag{7}$$

The distributional results presented in (4), (5), and (7) (for a proof, see, for example, Murphy 2007) are the building blocks for the MCMC sampler which is used to make Bayesian inference on the model parameters of interest.

4 Inference via MCMC

Since the full posterior is not analytically tractable, inference is performed using MCMC sampling with trans-dimensional moves (Green 1995). The main objective of the analysis is to cluster the nodes, and therefore the locations $\boldsymbol{\mu}_k$, the variance parameters $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\sigma}_k^2$ and the community probabilities $\boldsymbol{\theta}$ are considered as nuisance parameters and integrated out. Essentially, in this type of collapsed Gibbs sampler (Liu 1994), four moves are available (Richardson and Green 1997), described in the subsequent four subsections. A similar sampler is used by Wyse and Friel (2012) to estimate the number of clusters in stochastic blockmodels.

4.1 Change in the community allocations

A fully collapsed Gibbs update for each community assignment is available:

$$p(z_i = k|\boldsymbol{z}_{-i}, \mathbf{X}, d, K) \propto p(z_i = k|\boldsymbol{z}_{-i}, d, K) \times p(\boldsymbol{x}_i|\{\boldsymbol{x}_j\}_{j\neq i:z_j=k}, d). \tag{8}$$

In the special case where $d = K_\varnothing$ and $n_{z_i} = 1$, the full conditional distribution for z_i assigns probability one to retaining the same value since the model does not permit $d > K_\varnothing$. Otherwise, from (7):

$$p(z_i = k|\boldsymbol{z}_{-i}, d, K) \propto \frac{n_k^{-i} + \alpha/K}{n - 1 + \alpha}. \tag{9}$$

where $n_k^{-i} = n_k - \mathbb{1}_k(z_i)$. Similarly, the remaining term in (8), $p(\boldsymbol{x}_i|\{\boldsymbol{x}_j\}_{j\neq i:z_j=k}, d)$, can be obtained as the ratio of marginal likelihoods

$$p(\boldsymbol{x}_i|\{\boldsymbol{x}_j\}_{j\neq i:z_j=k}, d) = \frac{p(\boldsymbol{x}_i, \{\boldsymbol{x}_j\}_{j\neq i:z_j=k}|d)}{p(\{\boldsymbol{x}_j\}_{j\neq i:z_j=k}|d)}. \tag{10}$$

The ratio (10) can be decomposed as the product of two ratios of marginal likelihoods. Using (5), the first ratio is equivalent to t distribution (Murphy 2007):

$$p(\boldsymbol{x}_{i,:d}|\{\boldsymbol{x}_{j,:d}\}_{j\neq i:z_j=k}, d) = t_{\nu_{n_k^{-i}}}\left(\boldsymbol{x}_{i,:d}\middle|\boldsymbol{m}^{(k)}_{:d}, \frac{\kappa_{n_k^{-i}}+1}{\kappa_{n_k^{-i}}\nu_{n_k^{-i}}}\mathbf{D}^{(k),-i}_{:d}\right), \tag{11}$$

where the additional superscript $-i$ denotes a cluster quantity that is computed excluding the allocation z_i of the i-th node, and $t_\nu(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density function of a (possibly multivariate) Student's t distribution with ν degrees of freedom, location $\boldsymbol{\mu}$, and shape matrix $\boldsymbol{\Sigma}$. The second ratio, which accounts for the last $m - d$ dimensions, has the form

$$p(\boldsymbol{x}_{i,d:}|\{\boldsymbol{x}_{j,d:}\}_{j\neq i:z_j=k}, d) = \prod_{j=d+1}^{m} t_{\lambda_{n_k^{-i}}}\left(x_{i,j}\middle|0, s_j^{(k),-i}\right). \tag{12}$$

4.2 Split or merge two communities

To vary the number of communities, move proposals inspired by Sequentially-Allocated Merge-Split sampling (Dahl 2003; Jain and Neal 2004) are used here. Two indices i and j are sampled at random from the n nodes, and without loss of generality assume $z_i \leq z_j$. If $z_i = z_j$, then the single cluster is split, whereas if $z_i > z_j$ the two clusters are merged. In both move types, node i will remain in the same cluster, denoted $z_i^\star = z_i$. In the merge move, all elements of cluster z_j are reassigned to cluster z_i (with any higher indexed clusters subsequently decremented). For the split move, node j is first reassigned to cluster $K^\star = K + 1$ with new allocation $z_j^\star = K^\star$; then, in random order the remaining nodes currently allocated to cluster z_i are randomly reassigned to clusters z_i or $K^\star$ with probability proportional to their predictive distribution from the generative model (10). Denoting the resulting product of renormalised predictive densities from these reallocations by $q(K^\star, \boldsymbol{z}^\star|K, z)$, the acceptance probability for a split move, for example, is

$$\alpha(K^\star, z^\star | K, z) = \min\left\{1, \frac{p(\mathbf{X}|d, K^\star, z^\star)p(d|z^\star, K^\star)p(z^\star, K^\star)}{p(\mathbf{X}|d, K, z)p(d|z, K)p(z, K)q(K^\star, z^\star|K, z)}\right\}. \quad (13)$$

The ratio of densities for $\mathbf{X}$ in (13) will depend only upon the rows of the matrix corresponding to the cluster being split (or similarly, merged), and these expressions will decompose as a products of terms for the first d and remaning $m - d$ components [*cf.* (11), (12)].

4.3 Create or remove an empty community

Adding or removing empty communities whilst fixing z corresponds to proposing $K^\star = K + 1$ or $K^\star = K - 1$ respectively, although the latter proposal is not possible if $K = K_\varnothing$, meaning there are currently no empty communities. The acceptance probability is simply

$$\alpha(K^\star|K) = \min\left\{1, \frac{p(z|K^\star)p(K^\star)q_\varnothing}{p(z|K)p(K)}\right\},$$

where the proposal ratio $q_\varnothing = 2$ if $K^\star = K_\varnothing$, $q_\varnothing = 0.5$ if $K = K_\varnothing$ and $q_\varnothing = 1$ otherwise.

4.4 Change in the latent dimension

This move is only required when the latent dimension is not marginalised out. Given a current value d, a new value $d^\star$ is proposed from a density $q(d^\star|d) \propto \xi^{|d^\star - d|}\mathbb{1}_{\mathcal{D}}(d^\star)$ on a neighbourhood $\mathcal{D} = \{\max\{1, d - l\}, \ldots, d - 1, d + 1, \ldots, \min\{d + l, m\}\}$, typically with $l \leq 5$ and $\xi \in (0, 1)$. The acceptance ratio reduces to

$$\alpha(d^\star|d) = \min\left\{1, \frac{p(\mathbf{X}|d^\star, K, z)p(d^\star|z)}{p(\mathbf{X}|d, K, z)p(d|z)}\frac{q(d|d^\star)}{q(d^\star|d)}\right\}.$$

Notably, if $d^\star > d$, the ratio $p(\mathbf{X}|d^\star, K, z)/p(\mathbf{X}|d, K, z)$ only depends on the first $d^\star$ components of the embedding, since the last $m - d^\star$ components remain independent by (1).

4.5 Inferring communities

Markov Chain Monte Carlo samplers for mixture models with varying number of clusters are well known to be affected by *label switching* (Jasra et al. 2005), since the likelihood is invariant to permutations of the cluster labels. However, the estimated posterior similarity between nodes i and j, $\hat{\pi}_{ij} = \hat{\mathbb{P}}(z_i = z_j|\mathbf{X}) = \sum_{s=1}^{M} \mathbb{1}_{z_i^{(s)}}\{z_j^{(s)}\}/M$ is invariant to label switching. Communities can be estimated from the MCMC chains using the posterior similarity matrix $\{\hat{\pi}_{ij}\}$ and the PEAR method (maximisation of the posterior expected adjusted Rand index, Fritsch and Ickstadt 2009). Alternatively, if a configuration with a fixed number of clusters K is required, the clusters can been estimated using hierarchical clustering with average linkage, using $1 - \hat{\pi}_{ij}$ as distance measure (Medvedovic et al. 2004).

5 Second-level clustering of community variances

Empirical analyses of simulations from the stochastic blockmodel show that identifying and clearly separating the K clusters in $\mathbf{X}_{d:}$ is particularly difficult for the sampler in settings when $d \ll m$. The problem is particularly evident when $m = n$ and d is small. In this case, it has been assessed empirically that the within-cluster variance of the true communities in simulated datasets seems to converge to similar values, such that $\sigma^2_{k,j} \approx \sigma^2_{\ell,j}$ for $j \gg d$ and $k \neq \ell$. Therefore, when m is large enough, the selected model tends to be under-specified: the correct dimension d is identified, but the true number of communities K is underestimated. This is also one of the main reasons why it is not advisable to directly fit a Gaussian mixture model on $\mathbf{X} \in \mathbb{R}^{n \times m}$ and allow K to be random, ignoring the role of d.

The problem is illustrated in Fig. 1, which shows the within-cluster variance for each dimension, obtained from performing ASE for a simulation of $n = 500$ nodes from a stochastic block model containing five communities with well-separated mean locations. In Fig. 1, the clustering structure z is assumed to be known. More details about simulations of SBMs are given in Sect. 7.1. In the plot, the within-cluster variance of three of the five communities of the simulated graph fluctuate around the same values on each dimension larger than d. For a dimension larger than approximately 150, four of the five clusters have approximately the same variance on the subsequent dimensions. Therefore, when $m = n$, the MCMC sampler selects the MAP estimate $\hat{K} = 2$ for parsimony, and increases the variance of the Gaussian distributions on the first two dimensions, on which the clusters are well separated.

The solution proposed here is to assume shared variance parameters between some of the clusters for dimensions larger than d. Specifically, each community $k \in \{1, \ldots, K\}$ is assigned a second-level cluster allocation $v_k \in \{1, \ldots, H\}$, with $H \leq K$. If $v_k = v_\ell$, then for $j > d$, $\sigma^2_{k,j} = \sigma^2_{\ell,j}$. Formally,

$$\boldsymbol{x}_i|d, K, z_i, v_{z_i} \sim \mathbb{N}_m\left(\begin{bmatrix}\boldsymbol{\mu}_{z_i} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{z_i} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{v_{z_i}}\mathbf{I}_{m-d}\end{bmatrix}\right),$$
$$v_k|K, H \sim \text{Multinoulli}(\boldsymbol{\phi}),\ k = 1, \ldots, K,$$
$$\boldsymbol{\phi}|H \sim \text{Dirichlet}(\beta/H, \ldots, \beta/H),$$

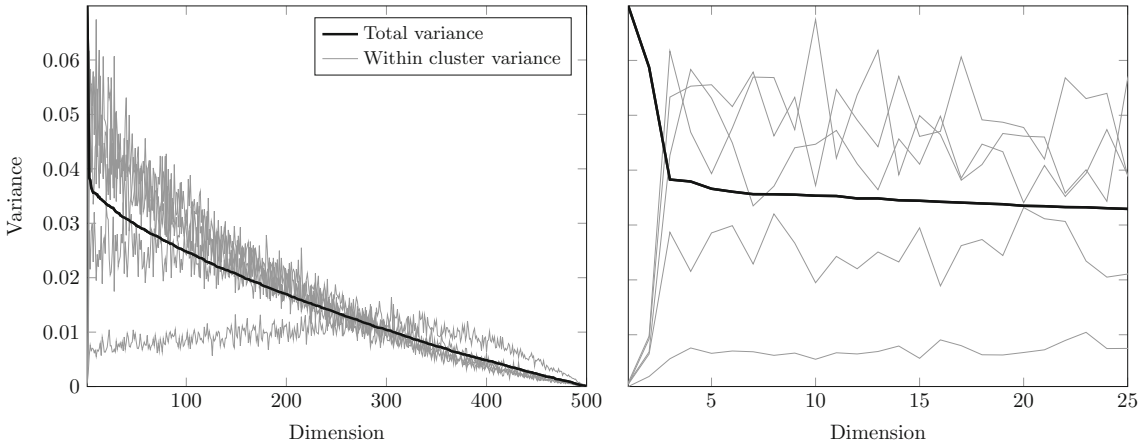

Fig. 1 Empirical within-block variance and total variance for the adjacency embedding obtained from a simulated 5-block SBM with $d = 2$, $n = 500$, d means $\boldsymbol{\mu}_1 = [0.7, 0.4]$, $\boldsymbol{\mu}_2 = [0.1, 0.1]$, $\boldsymbol{\mu}_3 = [0.4, 0.8]$, $\boldsymbol{\mu}_4 = [-0.1, 0.5]$ and $\boldsymbol{\mu}_5 = [0.3, 0.5]$, and $n_k = 100$ for $k = 1, \ldots, K$. The right panel is the left panel plot zoomed in to the first 25 dimensions

$$H|K \sim \text{Uniform}\{1, \ldots, K\}.$$

Essentially, the vector $\boldsymbol{v} = (v_1, \ldots, v_K)$ defines a *clustering of communities*. Note that if $H = 1$, all the communities are assigned to the same second-level cluster, and the problem of selecting d essentially reduces to an *ordinal* version of the feature selection problem in clustering (Raftery and Dean 2006). Also, if $H = 1$, then there is no information about the clusters in the last $m - d$ components of the embedding.

Under this extended model, the posterior distribution for $\sigma^2_{j,k}$ changes due to the aggregation of communities in the second level. Under the Inv-$\chi^2(\lambda_0, \sigma_0^2)$ prior, the posterior is Inv-$\chi^2(\lambda_{n_{\bullet k}}, s_j^{(\bullet k)})$, where $n_{\bullet k} = \sum_{\ell: v_\ell = k} n_\ell$ and

$$s_j^{(\bullet k)} = \left\{ \lambda_0 \sigma_0^2 + \sum\nolimits_{i: z_i = k} x_{ij}^2 \right\} \Big/ \lambda_{n_{\bullet k}}.$$

Calculations similar to (6) give the correct form of the marginal likelihood for the right hand side of the matrix, restricted to a given value of v_k. Clearly, $\boldsymbol{\phi}$ can be again marginalised out, yielding the marginal likelihood

$$p(\boldsymbol{v}|H) = \frac{\Gamma(\beta) \prod_{h=1}^{H} \Gamma(\sum_{k=1}^{K} \mathbb{1}_h\{v_k\} + \beta/H)}{\Gamma(\beta/H)^H \Gamma(K + \beta)}.$$

The MCMC sampler described in Sect. 4 must be slightly adapted. For the Gibbs sampling move in Sect. 4.1, the product of univariate Student's t densities in (12) is modified using the appropriate $(\lambda_{n_{\bullet k}}, s_j^{(\bullet k)})$ pair. For the change in dimension, $p(\mathbf{X}|d, K, \boldsymbol{z}, \boldsymbol{v})$ should be computed using the shared variances and the allocations $\boldsymbol{v}$. When an empty community is proposed, as in Sect. 4.3, the ratio $p(\boldsymbol{v}^\star|K)/p(\boldsymbol{v}|K)$ must be added, limited to the second level allocation of the additional community. The value v_k for the proposed empty cluster can be simply chosen at random from $\{1, \ldots, H\}$. Finally, for the split-merge move in Sect. 4.2, if $z_i = z_j$ for the two selected nodes, then $v_{z_i} = v_{z_j}$ after the split move. Alternatively, if $z_i \neq z_j$, then the new value of v_k is sampled at random from v_{z_i} and v_{z_j}.

Finally, three additional moves are required: resampling the second-level cluster allocations $\boldsymbol{v}$ using a Gibbs sampling step; proposing a second-level split-merge move; and adding or removing an empty second-level cluster. When $\boldsymbol{\phi}$ and the parameters of the Gaussian distributions are marginalised out, the second-level allocations are resampled according to the following equation:

$$p(v_k = h|\boldsymbol{v}_{-k}, \mathbf{X}, \boldsymbol{z}, d, K) \propto p(v_k = h|\boldsymbol{v}_{-k}, K) \times p\left(\mathbf{X}_{d:}^{(k)} \middle| \left\{\mathbf{X}_{d:}^{(\ell)}\right\}_{\ell \neq k: v_\ell = h}, v_k = h, d, K\right), \tag{14}$$

where the independence assumption between $\mathbf{X}_{:d}^{(k)}$ and $\mathbf{X}_{d:}^{(k)}$ is used. Similarly to (9):

$$p(v_k = h|\boldsymbol{v}_{-k}, K) = \frac{\sum_{\ell \neq k} \mathbb{1}_h\{v_\ell\} + \beta/H}{K - 1 + \beta}.$$

The calculations for the second term in (14) are similar to (10):

$$p\left(\mathbf{X}_{d:}^{(k)} \middle| \left\{\mathbf{X}_{d:}^{(\ell)}\right\}_{\ell \neq k: v_\ell = h}, v_k = h, d, K\right) = \frac{p\left(\mathbf{X}_{d:}^{(k)}, \left\{\mathbf{X}_{d:}^{(\ell)}\right\}_{\ell \neq k: v_\ell = h} \middle| v_k = h, d, K\right)}{p\left(\left\{\mathbf{X}_{d:}^{(\ell)}\right\}_{\ell \neq k: v_\ell = h} \middle| d, K\right)},$$

which can be computed using (6). The second-level split-merge move and the proposal of an empty cluster follows the same guidelines in Sects. 4.2 and 4.3.

Potentially, the model could be extended further using the same reasoning: from the plot in Fig. 1, it is clear that the different clusters begin to share the same variance at different points in the plot. Empirically, all the variances approximately converge to the same values at large dimensions, and it is therefore possible to identify a $(K-1)$-vector of discrete points in $\{d, d+1, \ldots, m\}$ at which different community variances coalesce. For the plot in Fig. 1, such vector could be $(d, d, d, 150, n)$, with $d = 2$ and $n = m = 500$.

6 Extension to directed and bipartite graphs

A directed graph $\mathbb{G} = (V, E)$ has the property that $(i, j) \in E \not\Longrightarrow (j, i) \in E$, meaning the corresponding adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$ is not, in general, symmetric. Directed graphs are useful for representing directed interaction networks, such as email traffic patterns; knowing that individual i broadcasts emails to individual j does not immediately imply that j also issues communications to i. In a random dot product graph context, it can be assumed that each node has two underlying latent positions $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$, characterising, respectively, its behaviour as a source and as a destination of the connection. Therefore, $\mathbb{P}(A_{ij} = 1) = \boldsymbol{x}_i^\top \boldsymbol{x}'_j$. For a stochastic blockmodel $\mathbb{P}(A_{ij} = 1) = B_{z_i z_j} = \boldsymbol{\mu}_{z_i}^\top \boldsymbol{\mu}'_{z_j}$ for latent positions $\boldsymbol{\mu}_{z_i}, \boldsymbol{\mu}'_{z_j} \in \mathbb{R}^d$, where the matrix $\mathbf{B} \in [0, 1]^{K \times K}$ is in this case asymmetric.

Definition 3 (Adjacency embedding of the directed graph) Given a directed graph with adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, and an integer $d \in \{1, \ldots, n\}$, consider the singular value decomposition

$$\mathbf{A} = \begin{bmatrix} \hat{\mathbf{U}} & \hat{\mathbf{U}}_\perp \end{bmatrix} \begin{bmatrix} \hat{\mathbf{D}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{D}}_\perp \end{bmatrix} \begin{bmatrix} \hat{\mathbf{V}}^\top \\ \hat{\mathbf{V}}_\perp^\top \end{bmatrix} = \hat{\mathbf{U}} \hat{\mathbf{D}} \hat{\mathbf{V}}^\top + \hat{\mathbf{U}}_\perp \hat{\mathbf{D}}_\perp \hat{\mathbf{V}}_\perp^\top,$$

where $\hat{\mathbf{D}} \in \mathbb{R}_+^{d \times d}$ is diagonal matrix containing the top d singular values in decreasing order, $\hat{\mathbf{U}} \in \mathbb{R}^{n \times d}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$ contain the corresponding left and right singular vectors, and the matrices $\hat{\mathbf{D}}_\perp$, $\hat{\mathbf{U}}_\perp$, and $\hat{\mathbf{V}}_\perp$ contain the remaining $n - d$ singular values and vectors. The d-dimensional directed adjacency embedding of $\mathbf{A}$ in $\mathbb{R}^d$, is defined as the pair $\hat{\mathbf{X}} = \hat{\mathbf{U}} \hat{\mathbf{D}}^{1/2}$, $\hat{\mathbf{X}}' = \hat{\mathbf{V}} \hat{\mathbf{D}}^{1/2}$.

Writing $\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}$ and $\mathbf{X}' = \mathbf{V}\mathbf{D}^{1/2}$, the rows of $\mathbf{X}$ characterise the activities of each node as a source, and the rows of $\mathbf{X}'$ characterise the same nodes as destinations.

The model in (1) can be easily adapted to directed graphs. Treating the embeddings $\mathbf{X}$ and $\mathbf{X}'$ as independent, each is modelled separately using the same Gaussian structure and prior distributions (1), except for three parameters which are initially assumed common to both embeddings: the latent dimension d, the number of communities K and the vector of node assignments to those communities, $\boldsymbol{z}$.

In some contexts it will be more relevant to consider different community membership structures for the same set of nodes when considering them as source nodes or destination nodes. In this case, let K denote the number of source communities and K' denote the number of destination communities; similarly let $\boldsymbol{z}$ denote the assignments of nodes to source communities, and $\boldsymbol{z}'$ the allocations to destination communities. The problem of jointly learning $\boldsymbol{z}$ and $\boldsymbol{z}'$ (as well as d) is commonly known as *co-clustering*, and the corresponding network model is known as the stochastic co-blockmodel (ScBM) (Rohe et al. 2016), or Latent Block Model (LBM) (Govaert and Nadif 2010). Given an asymmetric matrix $\mathbf{B} \in [0, 1]^{K \times K'}$, then $\mathbb{P}(A_{ij} = 1) = B_{z_i z'_j}$. From a random dot product graph perspective, it is assumed that $B_{z_i z'_j} = \boldsymbol{\mu}_{z_i}^\top \boldsymbol{\mu}'_{z'_i}$, for some latent positions $\boldsymbol{\mu}_{z_i}, \boldsymbol{\mu}'_{z'_i} \in \mathbb{R}^d$ and $d = \operatorname{rank}(\mathbf{B}) \leq \min(K, K')$.

The Bayesian model for ScBMs can be easily represented as a separate model for $\mathbf{X}$ and $\mathbf{X}'$, of the form given in (1), with the latent dimension of the embedding d now the only common parameter. Inference via MCMC can be performed in an equivalent way to the method described in Sect. 4; the only difference is in the expression of the acceptance ratio for a change in the shared latent dimension d, but the procedure can exploit the results obtained in Sect. 4.4, using the fact that

$$p(\mathbf{X}, \mathbf{X}' | d, K, K', \boldsymbol{z}, \boldsymbol{z}') = \prod_{k=1}^{K} p\left(\mathbf{X}_{:d}^{(k)} \middle| d\right) p\left(\mathbf{X}_{d:}^{(k)} \middle| d\right) \times \prod_{k'=1}^{K'} p\left(\mathbf{X}'^{(k')}_{:d} \middle| d\right) p\left(\mathbf{X}'^{(k')}_{d:} \middle| d\right),$$

where all the marginal likelihoods can be equivalently obtained from (5). Furthermore, the model can be appropriately modified when $d \ll m$ to include the second-level cluster allocations proposed in Sect. 5.

Finally, in bipartite graphs, the observed nodes can be partitioned into two sets V and V', with $V \cap V' = \varnothing$ and $E \cap (V \times V) \cup (V' \times V') = \varnothing$. Assume that V plays the role of the set of source nodes and V' of the set of destination nodes. Bipartite graphs are usually represented by a rectangular bi-adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n'}$, with $n' = |V'|$. In this case, it is still possible to apply the methods described in this section to the SVD embedding obtained from the rectangular matrix $\mathbf{A}$. Note that the ScBM extends trivially to the bipartite graph case, which is essentially a special case of a directed graph, with the cluster configurations

for source and destination nodes now inescapably unrelated and each node possessing only one latent representation in $\mathbb{R}^d$.

7 Applications and results

The Bayesian latent feature network models described in this article have been applied to both simulated and real world network data from undirected and directed graphs. The real network data analysed are from an undirected network obtained from the Santander bikes hires in London, and the Enron Email Dataset ; details are given in the corresponding sections.

The model and MCMC sampler have been tested using different combinations of the hyperparameters, showing robustness to the prior choice. In absence of strong prior information about the community structure, it is advisable to use the usual *uniformative* values $\kappa_0 = \nu_0 = \lambda_0 = \alpha = \beta = 1$, and $\omega = \delta = 0.1$. For the proposal of change in dimension (*cf.* Sect. 4.4), $\xi = 0.8$. Those values have been used as default settings for the MCMC sampler in the next sections. Inferential performance is sensitive to extreme values of the variance parameters, relative to the prior mean, but otherwise robust. So in practice, the expectation of the prior for the variance parameters should be chosen to be on the same scale as the observed data. The cluster configuration could be suitably initialised using K-means for some pre-specified K, usually chosen according to the scree-plot criterion. The second-level clusters have been initialised setting $H = K$. In order to set the prior covariances to a realistic value, the correlations in the $\mathbf{\Delta}_d$ matrices could be set to zero, and the elements on the diagonal of $\mathbf{\Delta}_d$ to the average within-cluster variance based on the K-means cluster configuration. Similarly, the prior σ^2_{0j} values could be set to the total variance on the corresponding column of the embedding.

In Sects. 7.2 and 7.3, the algorithms were initialised using the above guidelines, and run for a total of $M = 500\,000$ samples with burn-in 25 000, for a number of different chains to check for convergence.

7.1 Synthetic data and model validation

In order to validate the model assumptions in (1), stochastic blockmodels have been simulated and the fit of the proposed model has been evaluated on the estimated latent positions. A stochastic blockmodel can be simulated starting from a matrix $\mathbf{B} \in [0, 1]^{K \times K'}$ containing the probabilities of connection between communities, and a vector $\boldsymbol{\theta}$ of community allocation probabilities. For an undirected graph, $K = K'$ and the constraint $B_{k\ell} = B_{\ell k}$ is imposed; similarly, for directed graphs with a shared cluster configuration (*cf.* Sect. 6), $K = K'$. Each element $B_{k\ell}$ of $\mathbf{B}$ could be generated, for example, from independent beta draws.

Algorithm 1: Simulation of an undirected stochastic blockmodel.

1 for $i = 1, \ldots, n$, sample $z_i \sim \text{Multinoulli}(\boldsymbol{\theta})$,
2 simulate $\mathbf{B} = \{B_{k\ell}\} \in [0, 1]^{K \times K}$, $B_{k\ell} = B_{\ell k}$, where, for $k \leq \ell$, $B_{k\ell} \sim \text{Beta}(a, b)$ for $a, b > 0$,
3 obtain a rank-d truncation of $\mathbf{B}$ using $\tilde{\mathbf{B}} = \mathbf{\Gamma}_d \mathbf{\Lambda}_d \mathbf{\Gamma}_d^\top$, ensuring that $\tilde{B}_{k\ell} \in [0, 1]\ \forall k, \ell$,
4 obtain the adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, $A_{ij} = A_{ji}$, where $A_{ij} \sim \text{Bernoulli}(\tilde{B}_{z_i z_j})$.

The latent dimension d corresponds to the rank of the matrix $\mathbf{B}$, and a random matrix $\mathbf{B}$ generated from independent beta draws has full rank with probability 1. Therefore, to simulate $d < K$ a low-rank approximation of $\mathbf{B}$ must be used to generate the embedding. For undirected graphs, a truncated spectral decomposition can be used: $\tilde{\mathbf{B}} = \mathbf{\Gamma}_d \mathbf{\Lambda}_d \mathbf{\Gamma}_d^\top$ (recall Definition 1). Similarly, for the directed and bipartite graphs, the truncated SVD is an appropriate approximation: $\tilde{\mathbf{B}} = \mathbf{U}_d \mathbf{D}_d \mathbf{V}_d^\top$ (see Definition 3). Note that under this low-rank approximation, it must be checked that each element $\tilde{B}_{k\ell} \in [0, 1]$. The procedure is summarised in Algorithm 1.

The algorithm can be also extended to directed and bipartite graphs. In this section, each element $B_{k\ell}$ of $\mathbf{B}$ was generated from a Beta(1.2, 1.2) distribution, which produces communities with a moderate level of separation. For the given choice of K, the community allocations probabilities in the simulations were chosen to be $\boldsymbol{\theta} = (1/K, \ldots, 1/K)$, providing balanced clusters.

If a stochastic blockmodel is simulated using Algorithm 1, all the parameters are known, and the fit of model (1) can be evaluated using the true underlying cluster allocations $\boldsymbol{z}$.

7.1.1 Emprical analysis of spectral embeddings

Figure 2 illustrates results for a synthetic undirected stochastic blockmodel with $d = 2$ and $K = 5$. Figure 2a shows the scatterplot of the first two columns of the adjacency embedding $\mathbf{X}$, coloured using the true underlying communities. The plot shows well-separated clusters, which can be suitably modelled using a Gaussian mixture. Figure 2b shows the scatterplot of the next two dimensions. Clearly, the community mean locations are significantly different from zero in just the first two dimensions. This is further illustrated in Fig. 2c, which plots the empirical within-cluster means for each dimension, obtained using the known clustering $\boldsymbol{z}$. Fig. 2c gives empirical evidence that the form $[\boldsymbol{\mu}_{z_i}^\top, \mathbf{0}^\top]^\top$ for the mean of $\boldsymbol{x}_i$ in (1) suitably describes what is observed in SBMs. Similarly, Fig. 2d shows the empirical within-cluster variances for each dimension, again obtained using

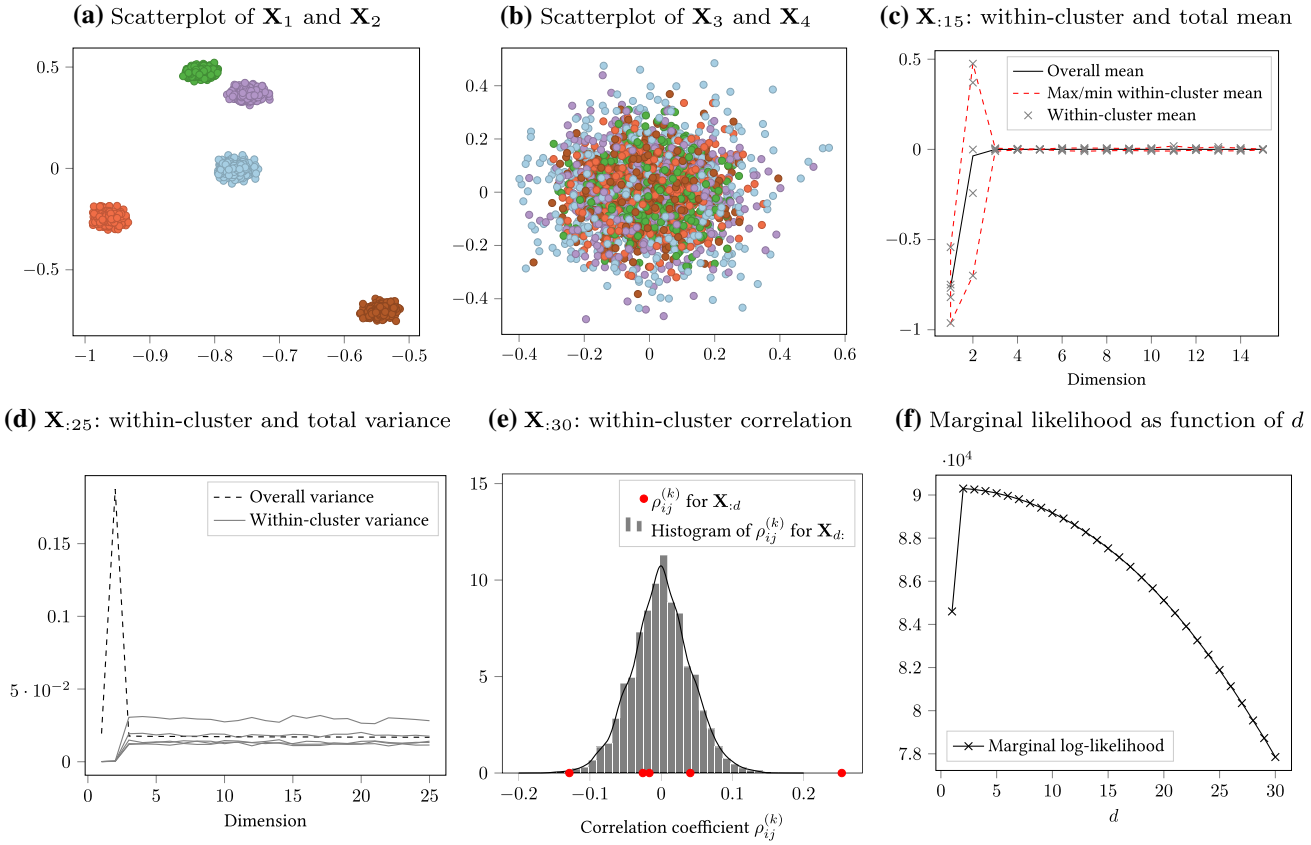

Fig. 2 Adjacency embedding for an undirected graph with $n = 2\,500$ nodes, $K = 5$, obtained from a symmetric $\mathbf{B} \in [0, 1]^{K \times K}$ with $B_{k\ell} \sim \text{Beta}(1.2, 1.2)$, $\boldsymbol{\theta} = (1/K, \ldots, 1/K)$ and $d = 2$. For (e) and (f), $m = 50$ is used

the known values of $\boldsymbol{z}$ from the simulation. In Fig. 2d, the difference between the within-cluster and overall variance is evident only in the first two dimensions, after which the quantities are of the same order of magnitude. The plot also shows that the within-cluster variances differ across communities, suggesting that it is appropriate to have cluster-specific values of $\sigma^2_{j,k}$ for $j > d$; this phenomenon can also be witnessed in Fig. 2b, and Fig. 1 in Sect. 5. Nevertheless, if the MCMC sampler were run on the simulated data in Fig. 2, it could also be appropriate to use a second-level clustering with $H = 3$, since the variances of three of the five communities are approximately the same for dimensions larger than $d = 2$. Furthermore, for small m and fairly large n, it seems from Fig. 2d that the vector $\boldsymbol{\sigma}^2_k$ could be approximated by a constant σ^2_k as in Yang et al. (2019). However, for small values of n and large m, as in Fig. 1, parameter vectors $\boldsymbol{\sigma}^2_k$ are clearly required. If a constant σ^2_k is used, the inferential procedure is essentially identical, but (6) should be modified accordingly.

Figure 2e shows the histogram of the empirical correlation coefficients $\rho_{ij}^{(k)}$, $i, j = 1, \ldots, m$, $i < j$, for each community, obtained from the known $\boldsymbol{z}$. The cluster-specific correlation coefficients $\rho_{12}^{(k)}$ between $\mathbf{X}_1$ and $\mathbf{X}_2$ are represented by bullet points in the plot, suggesting dependence for at least one of the clusters, confirming the result of Rubin-Delanchy et al. (2017), and providing further evidence that cluster-specific full covariance matrices $\boldsymbol{\Sigma}_k$ should be used for the covariance of $\boldsymbol{x}_i$ in (1). On the other hand, the empirical within-cluster correlations for $\mathbf{X}_{d:}$ tend to be small and centred around 0, suggesting that the assumption of independence is appropriate in that part of the model in (1). Finally, Fig. 2f plots the marginal log-likelihood as a function of d, using the known cluster configuration $\boldsymbol{z}$. The marginal likelihood strongly favours the true value $d = 2$, resulting in a posterior distribution essentially consisting of a point mass at the true value.

Figure 3 shows similar results for a simulated bipartite graph with separate community structures for nodes as sources and destinations, with $d = 2$, $K = 5$ and $K' = 3$. Again, the scatterplot for $\mathbf{X}'_1$ and $\mathbf{X}'_2$ in Fig. 3a are well-separated and can be easily estimated using the Gaussian mixture model. Figure 3b, c show the empirical within-cluster means for each dimension, obtained using $\boldsymbol{z}$ and $\boldsymbol{z}'$. The zero-mean assumption for the columns with index larger than d seems to hold even for a relatively small number of nodes per community. Figure 3d, e show the empirical within-

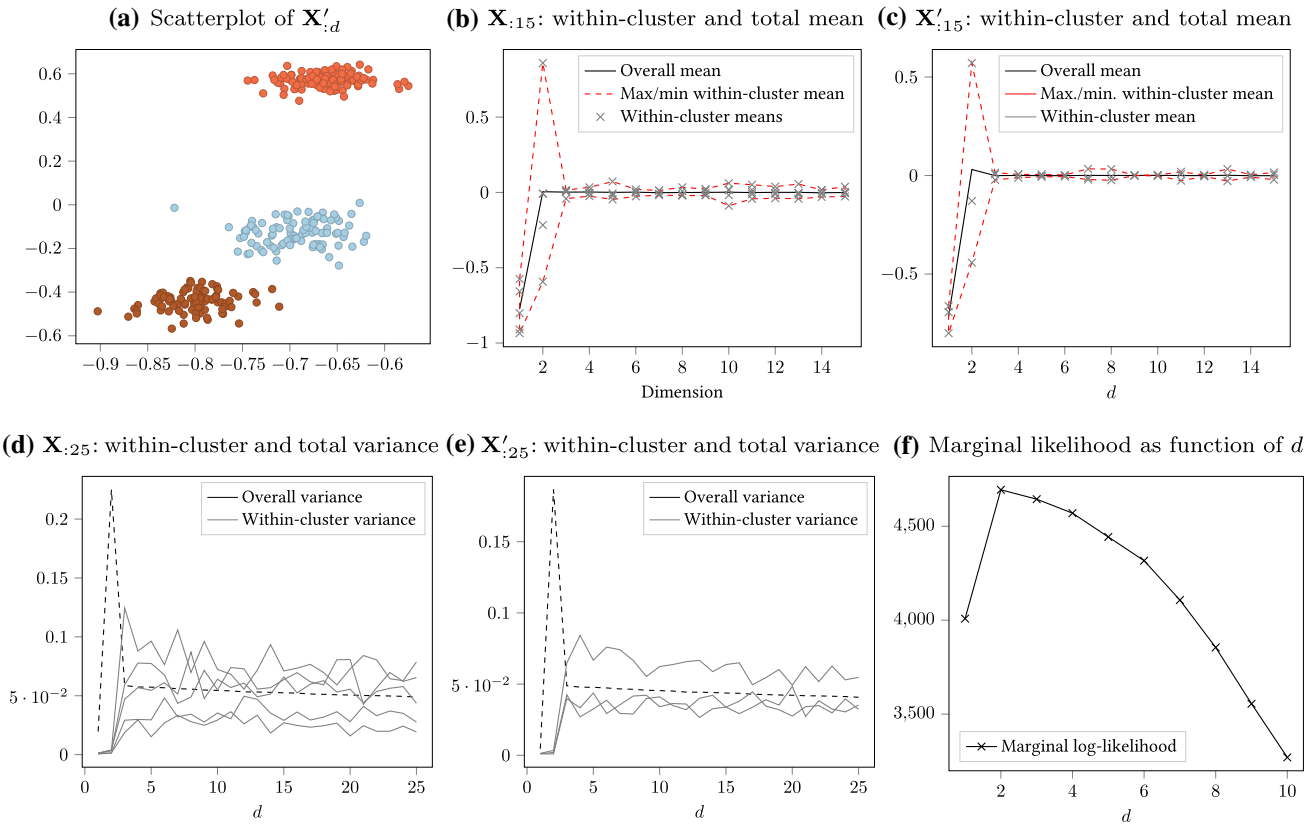

Fig. 3 Simulated adjacency embedding for a bipartite 250 × 300 graph with $K = 5$ and $K' = 3$, obtained from $\mathbf{B} \in [0, 1]^{K \times K'}$ with $B_{k\ell} \sim$ Beta(1.2, 1.2), $\boldsymbol{\theta} = (1/K, \ldots, 1/K)$, $\boldsymbol{\theta}' = (1/K', \ldots, 1/K')$, and $d = 2$. For (f), $m = 50$

cluster variances for each dimension. Again, it is confirmed that the variances are different for each community, even on the last $m - d$ components. The within-cluster variance on the last $m - d$ also seem to be decreasing, showing that for small graphs the parameter vector $\boldsymbol{\sigma}_k^2$ could be preferable to a constant σ_k^2 as in Yang et al. (2019). In Fig. 3f, the marginal likelihood strongly favours the true value $d = 2$, which again results in a point mass posterior centred at the true value. Similar considerations hold for the correlations, with results which are similar to the plots in Fig. 2e for the undirected graph.

7.1.2 Model parameter estimation

Table 1 shows the results from inference for model (1) when applied to SBMs simulated with different values of d and K, for all the possible combinations of the models considered in this article. For each (d, K) pair, an undirected SBM with $n = 1\,000$ nodes is generated, and the MCMC sampler is run three times, each with $M = 10\,000$ samples after a burn-in of 2 500. The samplers are initialised using the guidelines discussed in the introduction to Sect. 7. The table reports the averaged values of d, $K_\varnothing$ and $H_\varnothing$ obtained from the MCMC chains. The posterior mean values of d and K which are obtained are extremely close to the true values, implying that the results support the proposed model.

From Table 1, the performance of adjacency and Laplacian spectral embedding for recovering d, K and $\boldsymbol{z}$ seems to be equivalent. Similar results for $\mathbf{A}$ and $\mathbf{L}$ are also obtained for the application on real data in Sect. 7.2. Note that, for fixed values of K, Priebe et al. (2019) demonstrate that, in practical applications, LSE tends to better capture the affine structure in stochastic blockmodels, whereas ASE better identifies the core-periphery structure. Also, Table 1 shows that the constrained and unconstrained models seem to give an equivalent performance. The difference between the constrained model (2) and unconstrained model (3) will be more evident in Sects. 7.2 and 7.3, where the data might deviate from the stochastic blockmodel assumption. Overall, for synthetic data, the model seems robust and able to detect the correct d and K in a variety of different settings.

7.1.3 Second-level clustering

It is also possible to evaluate the effect of the second-order clustering for estimation of the community structure $\boldsymbol{z}$, and the parameters d and K. For the same simulated graph, the MCMC sampler has been run with and without the second

Table 1 Results of the inferential procedure for undirected SBMs simulated using different (d, K) pairs

(d, K)	Model	$m = 25$		
		$\bar{d}$	$\bar{K}_\varnothing$	$\bar{H}_\varnothing$
(2, 2)	constrained, ASE	2.00	2.00	1.99
	unconstrained, ASE	2.00	2.00	1.99
	constrained, LSE	2.01	2.03	1.99
	unconstrained, LSE	2.02	2.02	1.99
(2, 5)	constrained, ASE	2.00	5.05	1.77
	unconstrained, ASE	2.00	5.07	1.80
	constrained, LSE	2.05	5.10	3.11
	unconstrained, LSE	2.07	5.11	3.10
(6, 7)	constrained, ASE	6.00	7.04	2.10
	unconstrained, ASE	6.00	7.05	2.20
	constrained, LSE	6.00	7.10	2.47
	unconstrained, LSE	6.00	7.07	2.39
(9, 9)	constrained, ASE	8.97	9.01	2.08
	unconstrained, ASE	9.00	9.01	1.98
	constrained, LSE	9.00	9.02	2.12
	unconstrained, LSE	9.00	9.04	2.11
(9, 12)	constrained, ASE	9.00	12.02	1.96
	unconstrained, ASE	9.00	12.01	1.90
	constrained, LSE	9.00	12.03	2.60
	unconstrained, LSE	9.00	12.02	2.53
(10, 15)	constrained, ASE	10.00	14.78	1.25
	unconstrained, ASE	10.00	14.11	1.27
	constrained, LSE	10.00	14.81	1.81
	unconstrained, LSE	10.00	15.01	1.87

Table 2 Results for the MCMC sampler on simulated undirected SBMs for different values of m, with and without second order clustering

(d, K)	m	H random				$H = K$		
		$\hat{d}$	$\hat{K}_\varnothing$	$\bar{H}_\varnothing$	ARI	$\hat{d}$	$\hat{K}_\varnothing$	ARI
(3, 5)	15	3	5	1.669	1.000	3	5	1.000
	50	3	5	1.577	1.000	3	4	0.768
	150	3	5	1.467	1.000	3	4	0.768
	500	3	5	1.006	1.000	3	4	0.768
(9, 12)	15	9	12	1.979	1.000	9	12	1.000
	50	9	12	1.912	1.000	9	12	1.000
	150	9	12	1.875	1.000	9	11	0.942
	500	9	12	1.388	1.000	9	5	0.517

order clustering, using the same settings as the simulation in Table 1. Note that the absence of second order clustering corresponds to the case $H = K$. The procedure is repeated for different values of m, for n fixed, to study the effect of second-order clustering when m is increased. The results of the simulations are summarised in Table 2 for two different simulated graphs. The table reports the *maximum a posteriori* (MAP) values of d and $K_\varnothing$, and the adjusted Rand index (ARI) (Hubert and Arabie 1985) for the estimated clustering of the nodes, obtained setting $K_\varnothing$ to its MAP estimate. For simplicity, only the results for the unconstrained model using ASE are reported. For the case when H is unknown, the table reports the posterior mean of $H_\varnothing$.

In all the simulation settings, the model was able to recover the correct values of d, but when m is large, only the second-order clustering model allows K to be estimated correctly. When the second-order clustering is not used and m is very large relative to d and K, the MAP estimate $\hat{K}_\varnothing$ tends to be underestimated, and the inferred clustering structure is therefore also negatively affected. Also, the table shows that as $m \to n$, the estimated value of $H_\varnothing$ tends to decrease towards 1. Hence, to reduce the computational burden for large m, it would be possible to set $H = 1$, corresponding to the scenario $\boldsymbol{\sigma}_k^2 = \boldsymbol{\sigma}^2\ \forall k$, studied in Raftery and Dean (2006) in the context of variable selection in GMMs. Table 2 also suggests that the model with second-order clustering is robust to the choice of m. This is one of the main advantages of the proposed model: the correct d can be recovered for any choice of m, provided $d \leq m$. Choosing an upper bound m is easier than choosing the correct d, especially because of the robustness of the model. On the other hand, choosing large values of m makes the MCMC sampler computationally more expensive. A suitable choice would be to set m based on common criteria to obtain d (for example, the profile likelihood criterion of Zhu and Ghodsi 2006). For a given value $d^\star$ obtained using such criterion, it would be appropriate to choose $m > d^\star$, and then correct the initial estimate of d using the proposed model.

7.2 Undirected graphs: Santander bikes

The *Santander Cycle hire scheme* is a bike sharing system implemented in central London. Transport for London periodically releases data on the bike hires[1]. Considering this as a network, the nodes correspond to bike sharing stations, and an undirected edge between stations i and j is drawn if at least one ride between the two stations is completed within the time period considered. In this example, one week of data were considered, from 5 September until 11 September, 2018. The total number of stations used, $n = 783$; the total number of undirected edges $|E| = 96\,060$, implying the adjacency matrix is fairly dense. Note that it is possible to collect and return the bike to the same docking station. Those edges, amounting to less than 1% of $|E|$, have been removed, since our modelling framework is specifically developed for hollow binary matrices.

[1] The data are available at the following URL: https://cycling.data.tfl.gov.uk/, powered by TfL Open Data.

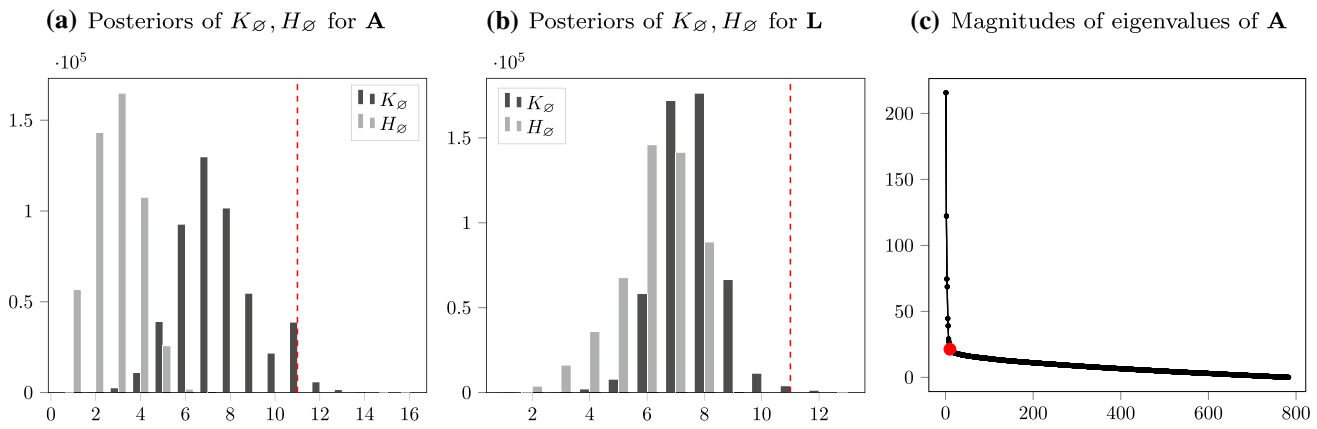

Fig. 4 Posterior distributions and scree-plots for the Santander bike network data using adjacency and Laplacian embeddings, for the unconstrained model (3). The dashed line in **a** and **b**, and the large bullet point in **c** represent MAP estimates of d

The results of the Bayesian inferential procedure, using the unconstrained prior (3) for d, applied to the adjacency and Laplacian embeddings for the Santander bike network are presented in Fig. 4. The initial value of K was set to 10, with $m = 25$, but similar estimates were obtained using different starting points for K and different values of m. It is interesting to note the different shapes of the posterior barplots of $K_\varnothing$ and $H_\varnothing$, Figs. 4a, b showing that the second-level clustering is crucial to obtain a more accurate model fit when the adjacency embedding is used. On the other hand, for the Laplacian embedding, the posteriors for $K_\varnothing$ and $H_\varnothing$ are extremely similar, suggesting that the second-level clustering is not required for $m = 25$. The MAP values $d = 11$ (adjacency) and $d = 12$ (Laplacian) correspond to the elbow in the scree-plots (see Fig. 4c for $\mathbf{A}$).

Note that, especially in the case of the adjacency embedding, d and K have similar values, showing that the two graphs might be well described by a stochastic blockmodel. Similarly, the constrained model with $d \sim \text{Uniform}\{1, \ldots, K_\varnothing\}$ (1) returns the same MAP estimates for d, but the constraint $d \leq K_\varnothing$ results in a larger number of small clusters; the posterior of $K_\varnothing$ essentially reduces to the rescaled probability mass function obtained from the unrestricted model, constrained such that $d \leq K_\varnothing$, since the posterior for d is approximately a point mass.

The resulting estimated clustering for the unconstrained model (3) based on the adjacency embedding and $K = 11$ (the MAP for d), plotted in Fig. 5, shows a clear structure: the largest communities have approximately the same extension, with a few exceptions. This is expected, since the bikes are free for the first 30 minutes and have limited speed, and are therefore used for small distance journeys. Two clusters are significantly smaller than the others, and correspond to touristic areas around Westminster, Covent Garden and Buckingham Palace. On the other hand, two clusters have a large geographical extension, and cover the East and West London areas. For the adjacency embedding, the MAP clustering obtained from the restricted model is almost identical. The PEAR method (Fritsch and Ickstadt 2009) suggests $K = 7$ communities instead. Similarly, if the Laplacian embedding is used, the MAP clustering structure suggested by PEAR has $K = 7$ communities for the unconstrained model (3) and $K = 12$ for the constrained model (1).

7.3 Directed graphs: Enron email dataset

Next, the algorithm is applied to a directed network: the Enron Email Dataset[2]. The Enron database is a corpus of emails sent by the employees of the Enron corporation. The version of the Enron data which has been analysed here is described in Priebe et al. (2005), and consists of $n = 184$ nodes and 3 010 directed edges. A directed edge $i \to j$ is drawn if the employee i sent an email to the employee j.

The results of analysing this network are presented in Fig. 6. The initial value of K was set to 10, with $m = 25$, but again similar results were obtained using different starting points for K and different values of m. The plots in Figs. 6b, c report the estimated posterior distributions of $K_\varnothing$ and $H_\varnothing$ for the constrained (1) and unconstrained (3) models. Interestingly, the MAP estimate for d coincides with the MAP estimate for K in the unconstrained model, which is promising. For the constrained model, the MAP for K exceeds the MAP for d by 1, allowing for $\text{rank}(\mathbf{B}) < K$. Overall, d and K have similar values, showing that the graph might be well described by a directed stochastic blockmodel.

[2] The entire version of the data is available at the following URL: https://www.cs.cmu.edu/~./enron/.

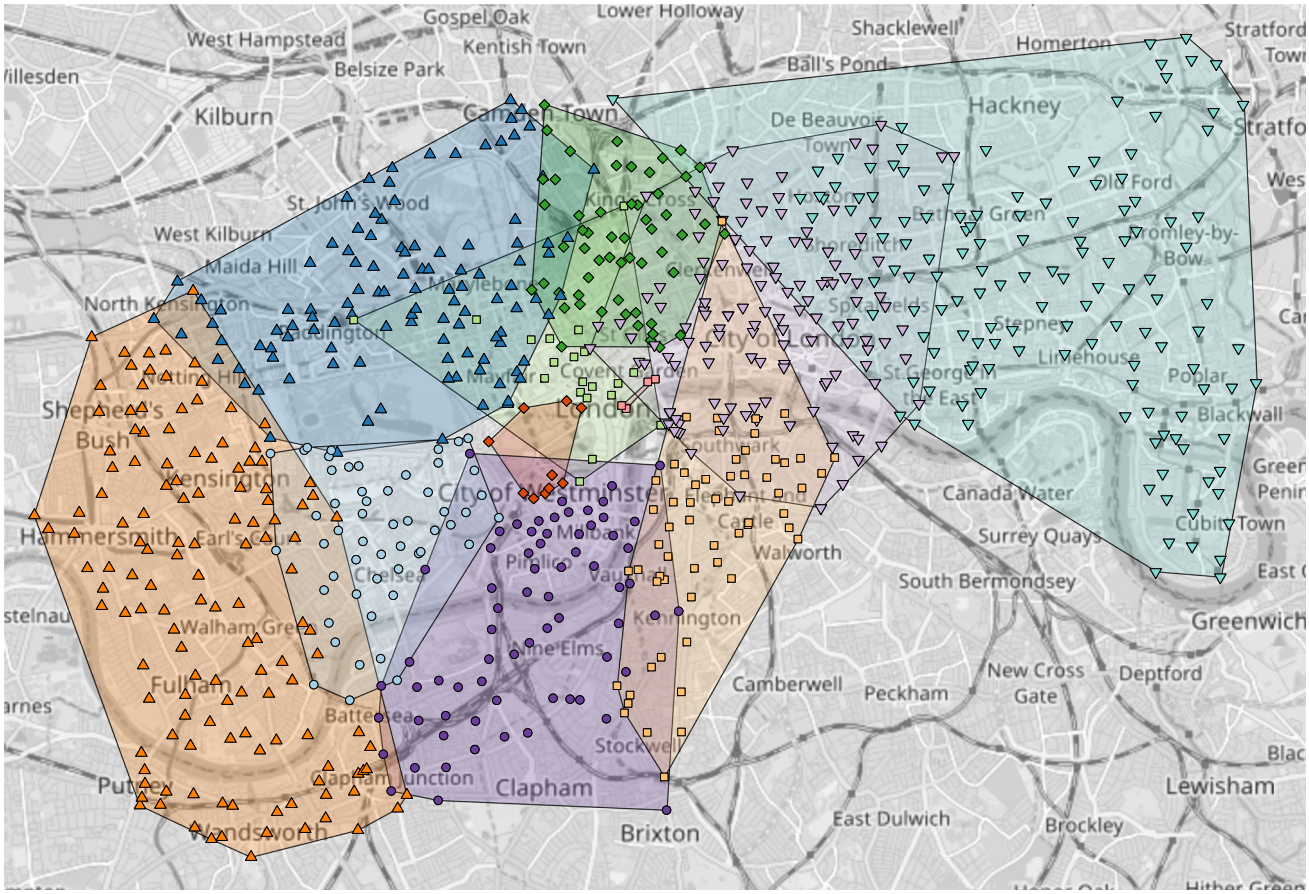

Fig. 5 Santander bike sharing stations in London and maximum a posteriori estimates of the cluster allocations of the stations, obtained using hierarchical clustering with distance $1 - \hat{\pi}_{ij}$, $K = 11$. Stations in the same convex hull share the same cluster

The posterior distributions for $K_\varnothing$ and $H_\varnothing$ in Fig. 6b, c are fairly different, showing that the second-level clustering might be relevant for this model. Inference on the model without second-level clustering confirms this impression: the posteriors for $K_\varnothing$, presented in Fig. 6d, e have a more symmetric shape, and the MAP latent dimension is $d = 6$. As before, the MAP for K is $d+1 = 7$, providing some evidence for the possibility $\text{rank}(\mathbf{B}) < K$.

From Fig. 6a, the selected MAP values $d = 6$ and $d = 9$ for the models with and without second-level clustering seem to be a tradeoff between the two most popular criteria for selection of the appropriate latent dimension: the *eigengap* heuristic suggests $d = 5$ if the second largest difference is considered, and the elbow in the scree-plot is approximately located around $d \approx 15$.

8 Conclusion

In this article, a novel Bayesian model has been proposed for automatic and simultaneous estimation of the number of communities and latent dimension of stochastic block-models, interpreted as special cases of generalised random dot product graphs. The Bayesian framework allows the number of communities K and latent dimension d to be treated as random variables, with associated posterior distributions. The postulated model is based on asymptotic results in the theory of network embeddings and random dot product graphs, and has been validated on synthetic datasets, showing good performance at recovering the latent parameters and communities. The model has been extended to directed and bipartite graphs, using SVD embeddings and allowing for co-clustering.

Overall, the main advantage of the proposed methodology is to allow for an arbitrarily large value of m, the number of columns (dimension) of the embedding at the first stage of the analysis, and then to treat d and K separately, allowing for the case $d = \text{rank}(\mathbf{B}) < K$, which is often overlooked in the literature. Problems arising from overspercification of m are tackled using a second-level clustering procedure. Also, the model provides an automated procedure and criterion to select the dimension of the embedding and an appropriate number of communities. If d is not constrained to be less than or equal to K, the model also provides empirical evi-

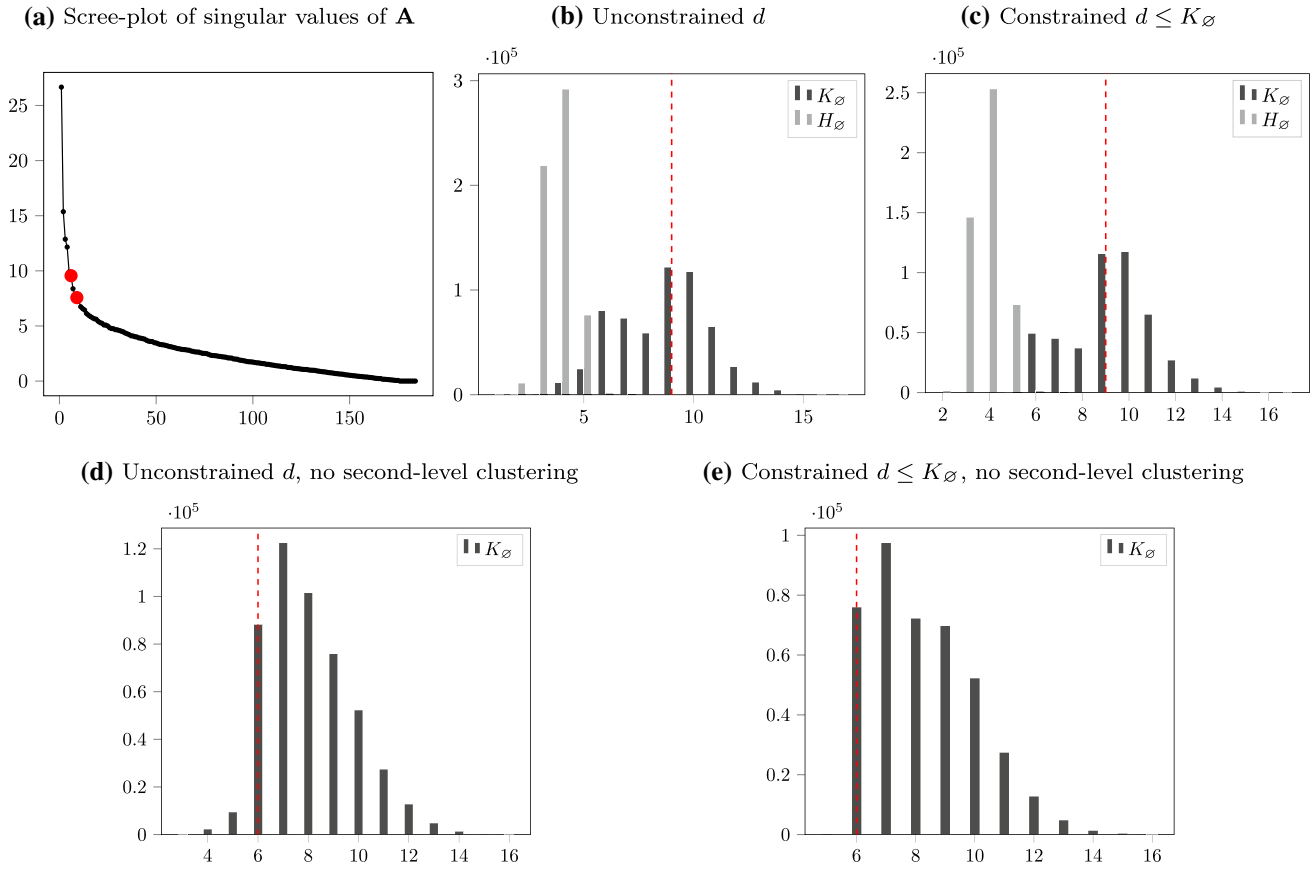

Fig. 6 Scree-plot of singular values of **A** and posterior distributions of $K_\varnothing$, $H_\varnothing$ for the Enron data. The large bullet points in **a**, and the dashed lines in **b**–**e** represent MAP estimates of d

dence of the goodness-of-fit of a stochastic blockmodel for the observed data. Results on real world network data sets show encouraging results in recovering the correct d, when compared to commonly used heuristic methods, and the community structure.

Supplementary material

The *python* code and datasets are available at https://www.github.com/fraspass/sbm.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Amini, A.A., Chen, A., Bickel, P.J., Levina, E.: Pseudo-likelihood methods for community detection in large sparse networks. Ann. Stat. **41**(4), 2097–2122 (2013)

Athreya, A., Fishkind, D.E., Tang, M., Priebe, C.E., Park, Y., Vogelstein, J.T., Levin, K., Lyzinski, V., Qin, Y.: Statistical inference on random dot product graphs: a survey. J. Mach. Learn. Res. **18**(1), 8393–8484 (2017)

Athreya, A., Priebe, C.E., Tang, M., Lyzinski, V., Marchette, D.J., Sussman, D.L.: A limit theorem for scaled eigenvectors of random dot product graphs. Sankhya A **78**(1), 1–18 (2016)

Bickel, P.J., Chen, A.: A nonparametric view of network models and Newman–Girvan and other modularities. Proc. Natl. Acad. Sci. **106**(50), 21068–21073 (2009)

Bickel, P.J., Sarkar, P.: Hypothesis testing for automated community detection in networks. J. R. Stat. Soc. Ser. B **78**(1), 253–273 (2016)

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**(7), 719–725 (2000)

Cai, H., Zheng, V.W., Chang, K.C.: A comprehensive survey of graph embedding: problems, techniques, and applications. IEEE Trans. Knowl. Data Eng. **30**, 1616–1637 (2018)

Celisse, A., Daudin, J., Pierre, L.: Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electron. J. Stat. **6**, 1847–1899 (2012)

Chatterjee, S.: Matrix estimation by universal singular value thresholding. Ann. Stat. **43**(1), 177–214 (2015)

Chen, K., Lei, J.: Network cross-validation for determining the number of communities in network data. J. Am. Stat. Assoc. **113**(521), 241–251 (2018)

Côme, E., Latouche, P.: Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. Stat. Model. **15**(6), 564–589 (2015)

Dahl, D.B.: An improved merge-split sampler for conjugate Dirichlet process mixture models. Tech. Rep. 1086, Department of Statistics, University of Wisconsin, Madison (2003)

Dellaportas, P., Papageorgiou, I.: Multivariate mixtures of normals with unknown number of components. Stat. Comput. **16**(1), 57–68 (2006)

Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. pp. 269–274. KDD '01, ACM, New York, NY, USA (2001)

Fowlkes, E.B., Gnanadesikan, R., Kettenring, J.R.: Variable selection in clustering. J. Classif. **5**(2), 205–228 (1988)

Franco Saldaña, D., Yu, Y., Feng, Y.: How many communities are there? J. Comput. Gr. Stat. **26**(1), 171–181 (2017)

Fritsch, A., Ickstadt, K.: Improved criteria for clustering based on the posterior similarity matrix. Bayesian Anal. **4**(2), 367–391 (2009)

Govaert, G., Nadif, M.: Latent block model for contingency table. Commun. Stat. Theo. Methods **39**(3), 416–425 (2010)

Green, P.J.: Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika **82**(4), 711–732 (1995)

Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. J. Am. Stat. Assoc. **97**(460), 1090–1098 (2002)

Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)

Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)

Jain, S., Neal, R.M.: A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model. J. Comput. Gr. Stat. **13**(1), 158–182 (2004)

Jasra, A., Holmes, C.C., Stephens, D.A.: Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Stat. Sci. **20**(1), 50–67 (2005)

Jolliffe, I.T.: Springer series in statistics. Principal component analysis. Springer, Berlin (2002)

Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1154–1166 (2004)

Lei, J.: A goodness-of-fit test for stochastic block models. Ann. Stat. **44**(1), 401–424 (2016)

Lei, J., Rinaldo, A.: Consistency of spectral clustering in stochastic block models. Ann. Stat. **43**(1), 215–237 (2015)

Liu, J.S.: The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. **89**(427), 958–966 (1994)

Malliaros, F.D., Vazirgiannis, M.: Clustering and community detection in directed networks: a survey. Phys. Rep. **533**(4), 95–142 (2013)

Maugis, C., Celeux, G., Martin-Magniette, M.L.: Variable selection for clustering with Gaussian mixture models. Biometrics **65**(3), 701–709 (2009)

Medvedovic, M., Yeung, K.Y., Bumgarner, R.E.: Bayesian mixture model based clustering of replicated microarray data. Bioinformatics **20**(8), 1222–1232 (2004)

Mengersen, K., Robert, C.: Testing for mixtures: a Bayesian entropic approach (with discussion). In: Berger, J., Bernardo, J., Dawid, A., Lindley, D., Smith, A. (eds.) Bayesian Statistics. Oxford University Press, Oxford (1996)

Miller, J.W., Harrison, M.T.: Mixture models with a prior on the number of components. J. Am. Stat. Assoc. **113**(521), 340–356 (2018)

Murphy, K.P.: Conjugate Bayesian analysis of the Gaussian distribution. Tech. rep. (2007)

Newman, M.E.J., Reinert, G.: Estimating the number of communities in a network. Phys. Rev. Lett. **117**(7), 078301 (2016)

Nobile, A.: On the posterior distribution of the number of components in a finite mixture. Ann. Stat. **32**(5), 2044–2073 (2004)

Nowicki, K., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. J. Am. Stat. Assoc. **96**(455), 1077–1087 (2001)

Priebe, C.E., Conroy, J.M., Marchette, D.J., Park, Y.: Scan statistics on Enron graphs. Comput. Math. Organ. Theo. **11**(3), 229–247 (2005)

Priebe, C.E., Park, Y., Vogelstein, J.T., Conroy, J.M., Lyzinski, V., Tang, M., Athreya, A., Cape, J., Bridgeford, E.: On a two-truths phenomenon in spectral graph clustering. Proc. Nat. Acad. Sci. **116**(13), 5995–6000 (2019)

Raftery, A.E., Dean, N.: Variable selection for model-based clustering. J. Am. Stat. Assoc. **101**(473), 168–178 (2006)

Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. Roy. Stat. Soc. B **59**(4), 731–792 (1997)

Rohe, K., Chatterjee, S., Yu, B.: Spectral clustering and the high-dimensional stochastic blockmodel. Ann. Stat. **39**(4), 1878–1915 (2011)

Rohe, K., Qin, T., Yu, B.: Co-clustering directed graphs to discover asymmetries and directional communities. In: Proceedings of the National Academy of Sciences (2016)

Rubin-Delanchy, P., Adams, N.M., Heard, N.A.: Disassortativity of computer networks. In: 2016 IEEE conference on intelligence and security informatics (ISI). pp. 243–247, (2016)

Rubin-Delanchy, P., Priebe, C.E., Tang, M., Cape, J.: A statistical interpretation of spectral embedding: the generalised random dot product graph. ArXiv e-prints (2017)

Snijders, T.A.B., Nowicki, K.: Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classif. **14**(1), 75–100 (1997)

Stephens, M.: Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. Ann. Stat. **28**(1), 40–74 (2000)

Sussman, D.L., Tang, M., Priebe, C.E.: Consistent latent position estimation and vertex classification for random dot product graphs. IEEE Trans. Pattern Anal. Mach. Intell. **36**(1), 48–57 (2014)

Tadesse, M.G., Sha, N., Vannucci, M.: Bayesian variable selection in clustering high-dimensional data. J. Am. Stat. Assoc. **100**(470), 602–617 (2005)

Tang, M., Priebe, C.E.: Limit theorems for eigenvectors of the normalized Laplacian for random graphs. Ann. Stat. **46**(5), 2360–2415 (2018)

Tang, M., Sussman, D.L., Priebe, C.E.: Universally consistent vertex classification for latent positions graphs. Ann. Stat. **41**(3), 1406–1430 (2013)

von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **1**(4), 395–416 (2007)

Wang, W.J., Wong, G.Y.: Stochastic blockmodels for directed graphs. J. Am. Stat. Assoc. **82**(397), 8–19 (1987)

Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. J. Am. Stat. Assoc. **105**(490), 713–726 (2010)

Wolfe, P.J., Olhede, S.C.: Nonparametric graphon estimation. arXiv e-prints (2013)

Wyse, J., Friel, N.: Block clustering with collapsed latent block models. Stat. Comput. **22**(2), 415–428 (2012)

Wyse, J., Friel, N., Latouche, P.: Inferring structure in bipartite networks using the latent blockmodel and exact ICL. Net. Sci. **5**(1), 45–69 (2017)

Yang, C., Priebe, C.E., Park, Y., Marchette, D.J.: Simultaneous dimensionality and complexity model selection for spectral graph clustering. arXiv e-prints arXiv:1904.02926 (2019)

Young, S.J., Scheinerman, E.R.: Random dot product graph models for social networks. In: Bonato, A., Chung, F.R.K. (eds.) Algorithms Models Web Graph, pp. 138–149. Springer, Berlin (2007)

Zhao, Y., Levina, E., Zhu, J.: Community extraction for social networks. Proc. Nat. Acad. Sci. **108**(18), 7321–7326 (2011)

Zheng, Q., Skillicorn, D.B.: Spectral embedding of directed networks. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) pp. 432–439 (2015)

Zhu, M., Ghodsi, A.: Automatic dimensionality selection from the scree plot via the use of profile likelihood. Comput. Stat. Data Anal. **51**(2), 918–930 (2006)

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