Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities

Adrien Todeschini, Xenia Miscouridou and François Caron

INRIA & Institut de Mathématiques de Bordeaux, France
e-mail: adrien.todeschini@inria.fr

Department of Statistics, University of Oxford, UK
e-mail: xenia.miscouridou@spc.ox.ac.uk; caron@stats.ox.ac.uk

Abstract: We propose a novel statistical model for sparse networks with overlapping community structure. The model is based on representing the graph as an exchangeable point process and naturally generalizes existing probabilistic models with overlapping block-structure to the sparse regime. Our construction builds on vectors of completely random measures, and has interpretable parameters, each node being assigned a vector representing its levels of affiliation to some latent communities. We develop methods for efficient simulation of this class of random graphs and for scalable posterior inference. We show that the proposed approach can recover interpretable structure of real-world networks and can handle graphs with thousands of nodes and tens of thousands of edges.

Keywords and phrases: Networks, Random Graphs, Multiview Networks, Multigraphs, Completely Random Measures, Lévy measure, Multivariate Subordinator, Sparsity, Non-Negative Factorization, Exchangeability, Point Processes.

1. Introduction

There has been a growing interest in the analysis, understanding and modeling of network data over the recent years. A network is composed of a set of nodes, or vertices, with connections between them. Network data arise in a wide range of fields, and include social networks, collaboration networks, communication networks, biological networks, food webs and are a useful way of representing interactions between sets of objects. Of particular importance is the elaboration of random graph models, which can capture the salient properties of real-world graphs. Following the seminal work of Erdös and Rényi (1959), various network models have been proposed; see the overviews of Newman (2009), Kolaczyk (2009) or Goldenberg et al. (2010). In particular, a large body of the literature has concentrated on models that can capture some modular or community structure within the network. The first statistical network model in this line of research is the popular stochastic block-model (Holland et al., 1983; Nowicki and Snijders, 2001). The stochastic block-model assumes that each node belongs to one of p latent communities, and the probability of connection between two nodes is given by a $p \times p$ connectivity matrix. This model has been extended in various directions, by introducing degree-correction parameters (Karrer and Newman, 2011), by allowing the number of communities to be unknown or grow with the size of the network (Kemp et al., 2006; Newman and Reinert, 2016), or by considering overlapping communities (Airoldi et al., 2008; Miller et al., 2009; Latouche et al., 2011; Palla et al., 2012; Yang and Leskovec, 2013). Stochastic block-models and their extensions have shown to offer a very flexible modeling framework, with interpretable parameters, and have been successfully used for the analysis of numerous real-world networks. However, as outlined by Orbanz and Roy (2015), when one makes the usual assumption that the ordering of the nodes is irrelevant in the definition of the statistical network model, the Bayesian probabilistic versions of those models lead to dense networks[1]: that means that the number of edges grows quadratically with the number of nodes. This property is rather undesirable, as many real-world networks are found to be sparse.

Recently, Caron and Fox (2017) proposed an alternative framework for statistical network modeling. The framework is based on representing the graph as an exchangeable random measure on the plane. More precisely, each node is embedded at some location $\theta_i \in \mathbb{R}_+$ and, for simple graphs, a connection exists

[1]We refer to graphs whose number of edges scales quadratically with the number of nodes as dense, and sparse if it scales sub-quadratically.

between two nodes i and j if there exists a point at locations (θ_i, θ_j) and (θ_j, θ_i). An undirected simple graph is therefore represented by a symmetric point process Z on the plane

$$Z = \sum_{i,j} z_{ij} \delta_{(\theta_i, \theta_j)}, \tag{1}$$

where $z_{ij} = z_{ji} = 1$ if i and j are connected, 0 otherwise; see Figure 1 for an illustration. Caron and Fox (2017) noted that jointly exchangeable random measures, a notion to be defined in Eq. (21), admit a representation theorem by Kallenberg (1990), providing a general construction for exchangeable random measures hence for the random graphs represented by such objects. This connection is further explored by Veitch and Roy (2015) and Borgs et al. (2016), who provide a detailed description and extensive theoretical analysis of the associated class of random graphs, which they name *Kallenberg exchangeable graphs* or *graphon processes*. Within this class of models, Caron and Fox (2017) consider the following simple generative model, where two nodes $i \neq j$ connect with probability

$$\Pr(z_{ij} = 1 | (w_\ell)_{\ell=1,2,\ldots}) = 1 - e^{-2 w_i w_j} \tag{2}$$

where the $(w_i, \theta_i)_{i=1,2,\ldots}$ are the points of a Poisson point process on $\mathbb{R}_+^2$. The parameters $w_i > 0$ are interpreted as sociability parameters as explained in Caron and Fox (2017). Depending on the properties of the mean measure of the Poisson process, the authors show that it is possible to generate both dense and sparse graphs, with potentially heavy-tail degree distributions, within this framework. The construction (2) is however rather limited in terms of capturing structure in the network. Herlau et al. (2015) proposed an extension of (2), which can accommodate a community structure. More precisely, introducing latent community membership variables $c_i \in \{1, \ldots, p\}$, two nodes $i \neq j$ connect with probability

$$\Pr(z_{ij} = 1 | (w_\ell, c_\ell)_{\ell=1,2,\ldots}, (\eta_{k\ell})_{1 \leq k, \ell \leq p}) = 1 - e^{-2 \eta_{c_i c_j} w_i w_j} \tag{3}$$

where the $(w_i, c_i, \theta_i)_{i=1,2,\ldots}$ are the points of a (marked) Poisson point process on $\mathbb{R}_+ \times \{1, \ldots, p\} \times \mathbb{R}_+$ and $\eta_{k\ell}$ are positive random variables parameterizing the strength of interaction between nodes in community k and nodes in community ℓ. The model is similar in spirit to the degree-corrected stochastic block-model (Karrer and Newman, 2011), but within the point process framework (1), and can thus accommodate both sparse and dense networks with community structure. The model of Herlau et al. (2015) however shares the limitations of the (degree-corrected) stochastic block-model, in the sense that it cannot model overlapping community structures, since each node belongs to a single community; see Latouche et al. (2011) and Yang and Leskovec (2013) for more discussion along these lines. Other extensions with block structure or mixed membership block structure are also suggested by Borgs et al. (2016).

In this paper, we propose that each node i is assigned a set of latent non-negative parameters w_{ik}, $k = 1, \ldots, p$, and that the probability that two nodes $i \neq j$ connect is given by

$$\Pr(z_{ij} = 1 | (w_{\ell 1}, \ldots, w_{\ell p})_{\ell=1,2,\ldots}) = 1 - e^{-2 \sum_{k=1}^{p} w_{ik} w_{jk}}. \tag{4}$$

These non-negative weights can be interpreted as measuring the level of affiliation of node i to the latent communities $k = 1, \ldots, p$. For example, in a friendship network, these communities can correspond to colleagues, family, or sport partners, and the weights measure the level of affiliation of an individual to each community. Note that as individuals can have high weights in different communities, the model can capture overlapping communities as opposed to (non-overlapping) clustering structure. The link probability (4) builds on a non-negative factorization; it has been used by other authors for network modeling (Yang and Leskovec, 2013; Zhou, 2015) and is also closely related to the model for multigraphs of Ball et al. (2011). The main contribution of this paper is to use the link probability (4) within the point process framework of Caron and Fox (2017). To this aim, we consider that the node locations and weights $(w_{i1}, \ldots, w_{ip}, \theta_i)_{i=1,2,\ldots}$ are drawn from a Poisson point process on $\mathbb{R}_+^{p+1}$ with a given mean measure ν. The construction of such multivariate point processes relies on vectors of completely random measures (or equivalently multivariate subordinators). In particular, we build on the flexible though tractable construction recently introduced by Griffin and Leisen (2017).

The proposed model generalizes that of Caron and Fox (2017) by allowing the model to capture more structure in the network, while retaining its main features, and is shown to have the following properties:

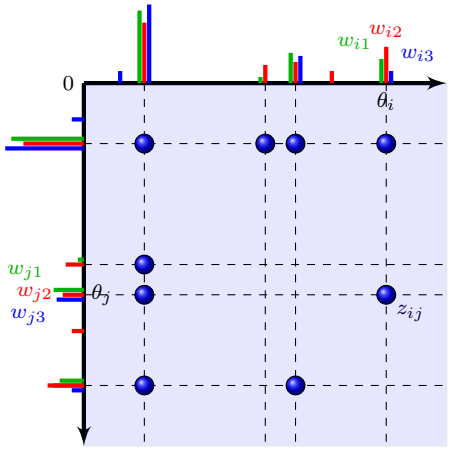


FIG 1. *Representation of a undirected graph via a point process Z. Each node i is embedded in $\mathbb{R}_+$ at some location θ_i and is associated with a set of positive attributes $(w_{i1}, \ldots, w_{ip})$. An edge between nodes θ_i and θ_j is represented by a point at locations (θ_i, θ_j) and (θ_j, θ_i) in $\mathbb{R}^2_+$.*

- Interpretability: each node is assigned a set of positive parameters, which measure the levels of affiliation of each node to each latent community; once those parameters are learned, they can be used to uncover the latent structure in the network.
- Sparsity: we can generate graphs whose number of edges grows subquadratically with the number of nodes, whereas in dense graphs this rate is quadratic.
- Exchangeability: is inherited in the framework of Kallenberg (1990).

Additionally, we develop a Markov chain Monte Carlo (MCMC) algorithm for posterior inference for this model, and perform experiments on two real-world networks with around a thousand of nodes and tens of thousands of edges.

The article is organized as follows. The class of random graph models is introduced in Section 2. Properties of the class of graphs and simulation are described in Section 3. We derive a scalable MCMC algorithm for posterior inference in Section 4. In Section 5 we provide illustrations of the proposed method on simulated data and on two real-world networks: a network of citations between political blogs and a network of connections between US airports. We show that the approach is able to uncover interpretable structure in the data and has competing or better performance when compared to alternative models.

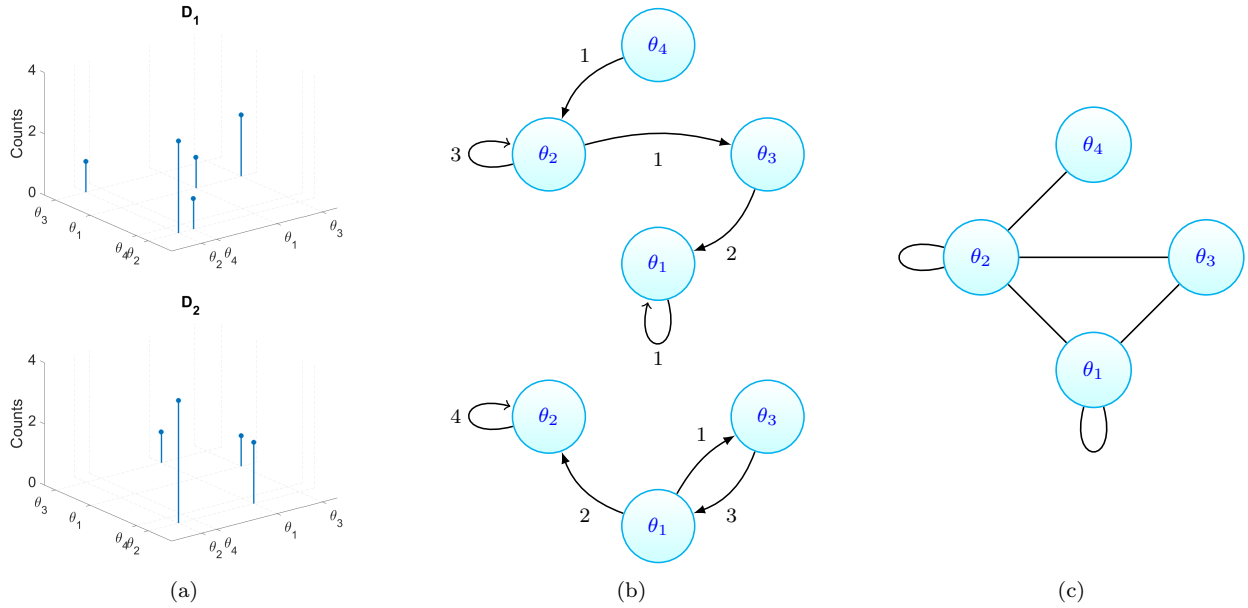
2. Sparse graph models with overlapping communities

In this section, we present the statistical model for simple graphs with self-loops. The construction builds on vectors of completely random measures (CRM, Kingman, 1967). We provide here the necessary material for the definition of the network model; please refer to Section A of the supplementary material for additional background on vectors of CRMs. Additionally, the model described in this section can be extended to bipartite graphs; see Section G of the supplementary material for more details.

2.1. General construction using vectors of CRMs

We consider that each node i is embedded at some location $\theta_i \in \mathbb{R}_+$, and has a set of positive weights $(w_{i1}, \ldots, w_{ip}) \in \mathbb{R}^p_+$. The points $(w_{i1}, \ldots, w_{ip}, \theta_i)_{i=1,\ldots,\infty}$ are assumed to be drawn from a Poisson process with mean measure

$$\nu(dw_1, \ldots, dw_p, d\theta) = \rho(dw_1, \ldots, dw_p)\lambda(d\theta) \tag{5}$$



(a) (b) (c)

FIG 2. *An example of (a) the restriction on* $[0,1]^2$ *of the two atomic measures* D_1 *and* D_2*, (b) the corresponding multiview directed multigraphs (top: view 1; bottom: view 2) and (c) corresponding undirected graph.*

where λ is the Lebesgue measure and ρ is a σ-finite measure on $\mathbb{R}_+^p$, concentrated on $\mathbb{R}_+^p \backslash \{\mathbf{0}\}$, which satisfies

$$\int_{\mathbb{R}_+^p} \min\left(1, \sum_{k=1}^{p} w_k\right) \rho(dw_1, \ldots, dw_p) < \infty. \tag{6}$$

Under this condition (Skorohod, 1991; Barndorff-Nielsen et al., 2001), we can describe the set of weights and locations using a vector of completely random measures $(W_1, \ldots, W_p)$ on $\mathbb{R}_+$:

$$W_k = \sum_{i=1}^{\infty} w_{ik} \delta_{\theta_i}, \text{ for } k = 1, \ldots, p. \tag{7}$$

We simply write

$$(W_1, \ldots, W_p) \sim \text{CRM}(\rho, \lambda). \tag{8}$$

We propose a hierarchical construction, similar to the one of Caron and Fox (2017), and introduce integer-valued random measures D_k on $\mathbb{R}_+^2$, $k = 1, \ldots, p$,

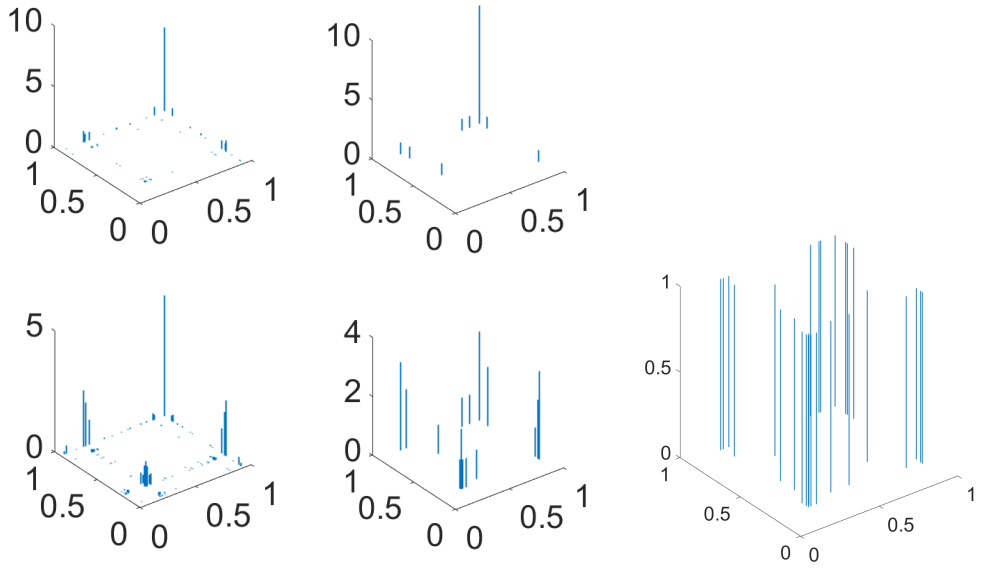
$$D_k = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} n_{ijk} \delta_{(\theta_i, \theta_j)}, \tag{9}$$

where the n_{ijk} are natural integers. The vector of random measures $(D_1, \ldots, D_p)$ can be interpreted as representing a multiview (a.k.a. multiplex or multi-relational) directed multigraph (Verbrugge, 1979; Salter-Townshend and McCormick, 2013), where n_{ijk} represents the number of interactions from node i to node j in the view (community) k; see Figure 2 for an illustration. Conditionally on the vector of CRMs, the measures D_k are independently drawn from a Poisson process[2] with mean measure $W_k \times W_k$

$$D_k | (W_1, \ldots, W_p) \sim \text{Poisson}\left(W_k \times W_k\right) \tag{10}$$

that is, the n_{ijk} are independently Poisson distributed with rate $w_{ik} w_{jk}$.

[2]Note that we consider a generalized definition of a Poisson process, where the mean measure is allowed to have atoms; see e.g. Daley and Vere-Jones (2008, Section 2.4).



(a) $W_k \times W_k$ (b) Integer point processes D_k (c) Point process Z

FIG 3. *An example, for $p = 2$, of (a) the product measures $W_k \times W_k$, (b) a draw of the directed multigraph measures $D_k \mid W_k \sim \text{Poisson}(W_k \times W_k)$ and (c) corresponding undirected measure $Z = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \min(1, \sum_{k=1}^{p} n_{ijk} + n_{jik})\delta_{(\theta_i,\theta_j)}$.*

Finally, the point process Z representing the graph (shown in Eq. (1)) is deterministically obtained from $(D_1, \ldots, D_p)$ by setting $z_{ij} = 1$ if there is at least one directed connection between i and j in any view, and 0 otherwise, therefore $z_{ij} = \min(1, \sum_{k=1}^{p} n_{ijk} + n_{jik})$. Overall, the graph model is described as follows:

$$\begin{array}{ll} W_k = \sum_{i=1}^{\infty} w_{ik}\delta_{\theta_i} & (W_1, \ldots, W_p) \sim \text{CRM}(\rho, \lambda) \\ D_k = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} n_{ijk}\delta_{(\theta_i,\theta_j)} & D_k \mid W_k \sim \text{Poisson}\,(W_k \times W_k) \\ Z = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \min(1, \sum_{k=1}^{p} n_{ijk} + n_{jik})\delta_{(\theta_i,\theta_j)}. & \end{array} \tag{11}$$

The model construction is illustrated in Figure 3. Integrating out the measures D_k, $k = 1, \ldots, p$, the construction can be expressed as, for $i \leq j$

$$z_{ij} | (w_{\ell 1}, \ldots, w_{\ell p})_{\ell=1,2,\ldots} \sim \begin{cases} \text{Ber}(1 - \exp(-2\sum_{k=1}^{p} w_{ik}w_{jk})) & i \neq j \\ \text{Ber}(1 - \exp(-\sum_{k=1}^{p} w_{ik}^2)) & i = j \end{cases} \tag{12}$$

and $z_{ji} = z_{ij}$; see Figure 1.

Graph Restrictions. Except for trivial cases, we have $W_k(\mathbb{R}_+) = \infty$ a.s. and therefore $Z(\mathbb{R}_+^2) = \infty$ a.s., which means that the number of points over the plane is infinite a.s. For $\alpha > 0$, we consider restrictions of the measures W_k, $k = 1, \ldots, p$, to the interval $[0, \alpha]$ and of the measures D_k and Z to the box $[0, \alpha]^2$, and write respectively $W_{k\alpha}$, $D_{k\alpha}$ and Z_α for these restrictions. Note that condition (6) ensures that $W_{k\alpha}([0, \alpha]) < \infty$ a.s. hence $D_{k\alpha}([0, \alpha]^2) < \infty$ and $Z_\alpha([0, \alpha]^2) < \infty$ a.s. As a consequence, for a given $\alpha > 0$, the model yields a finite number of edges a.s., even though there may be an infinite number of points $(w_i, \theta_i) \in \mathbb{R}_+ \times [0, \alpha]$; see Section 3.

Remark 1 *The model defined above can also be used for random multigraphs, where $n_{ij} = \sum_{k=1}^{p} n_{ijk}$ is the number of directed interactions between i and j. Then we have*

$$n_{ij} | (w_{\ell 1}, \ldots, w_{\ell p})_{\ell=1,2,\ldots} \sim \text{Poisson}\left(\sum_{k=1}^{p} w_{ik}w_{jk}\right)$$

which is a Poisson non-negative factorization (Lee, 1999; Cemgil, 2009; Psorakis et al., 2011; Ball et al., 2011; Gopalan et al., 2015).

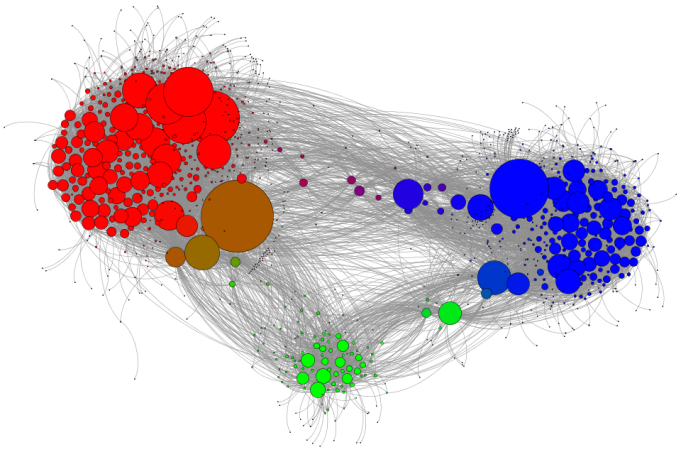


FIG 4. *Graph sampled from the model with three latent communities, identified by colors red, green, blue. For each node, the intensity of each color is proportional to the value of the associated weight in that community. Pure red/green/blue color indicates the node is only strongly affiliated to a single community. A mixture of those colors indicates balanced affiliations to different communities. Graph generated with the software Gephi (Bastian et al., 2009).*

Remark 2 *The model defined by Eq. (12) allows to model networks which exhibit assortativity (Newman, 2003), meaning that two nodes with similar characteristics (here similar set of weights) are more likely to connect than nodes with dissimilar characteristics. The link function can be generalized to (see e.g. Zhou, 2015)*

$$z_{ij} \sim \text{Ber}\left(1 - \exp\left(-\sum_{k=1}^{p}\sum_{\ell=1}^{p} \eta_{k\ell} w_{ik} w_{j\ell}\right)\right)$$

where parameters $\eta_{k\ell} \geq 0$ are used in order to be able to capture both assortative and dissortative mixing in the network. In particular, setting larger values off-diagonal than on the diagonal of the matrix $(\eta_{k\ell})_{1\leq k,\ell\leq p}$ allows to capture dissortativity. The properties and algorithms for simulation and posterior inference can be easily extended to this more general case. For simplicity, we focus here on the simpler link function (12).

2.2. Particular model based on compound CRMs

The key component in our statistical network model is the multivariate Lévy measure ρ in (8). Various approaches have been developed for constructing multivariate Lévy measures (Tankov, 2003; Cont and Tankov, 2003; Kallsen and Tankov, 2006; Barndorff-Nielsen et al., 2001; Skorohod, 1991), or more specifically vectors of completely random measures (Epifani and Lijoi, 2010; Leisen and Lijoi, 2011; Leisen et al., 2013; Griffin et al., 2013; Lijoi et al., 2014). Here we consider the following particular form

$$\rho(dw_1, \ldots, dw_p) = \int_0^\infty w_0^{-p} F\left(\frac{dw_1}{w_0}, \ldots, \frac{dw_p}{w_0}\right) \rho_0(dw_0) \tag{13}$$

where $F(d\beta_1, \ldots d\beta_p)$ is some *score* probability distribution on $\mathbb{R}_+^p$ and ρ_0 is a *base* Lévy measure on $\mathbb{R}_+$. The model defined by (5) and (13) is a special case of the compound completely random measure (CCRM) model proposed by Griffin and Leisen (2017). It admits the following hierarchical construction, which makes interpretability, characterization of the conditionals and analysis of this class of models particularly easy. Let

$$W_0 = \sum_{i=1}^{\infty} w_{i0}\delta_{\theta_i} \sim \text{CRM}(\rho_0, \lambda) \tag{14}$$

and for $k = 1, \ldots, p$ and $i = 1, 2, \ldots$

$$w_{ik} = \beta_{ik} w_{i0}$$

where the scores β_{ik} are independent of (w_{i0}) with the following joint distribution

$$(\beta_{i1}, \ldots, \beta_{ip}) \overset{\text{iid}}{\sim} F. \tag{15}$$

Additionally, the set of points $(w_{i0}, \beta_{i1}, \ldots, \beta_{ip})_{i=1,2,\ldots}$ is a Poisson point process with mean measure

$$F(d\beta_1, \ldots, d\beta_p)\rho_0(dw_0). \tag{16}$$

Dependence between the different CRMs is both tuned by the shared scaling parameter w_{i0} and potential dependency between the scores $(\beta_{i1}, \ldots, \beta_{ip})$. The hierarchical construction has the following interpretation:

- The weight w_{i0} is an individual scaling parameter for node i with distribution tuned by the base Lévy measure ρ_0. It is considered as a degree correction parameter, a concept that is often met in network models (Karrer and Newman, 2011; Zhao et al., 2012; Herlau et al., 2015). As shown in Section 3, ρ_0 tunes the overall sparsity properties of the network.
- The community-related scores β_{ik} tune the level of affiliation of node i to community k; this is controlled by the score distribution F. These parameters tune the overlapping block-structure of the network.
- Finally, the parameter $\alpha > 0$ tunes the size of the network, the number of edges increasing quadratically with α.

Note that the variables z_{ij} are independent of the locations θ_i which are therefore not likelihood-identifiable, and only introduced for the construction of the model. See (Caron and Fox, 2017, Section 3.5) for more discussion and interpretation of these parameters. An example of such a graph with three communities is displayed in Figure 4.

Specific choices for F and ρ_0. We now give specific choices for the score distribution F and base Lévy measure ρ_0, which lead to scalable inference algorithms. As in Griffin and Leisen (2017), we consider that F is a product of independent gamma distributions

$$F(d\beta_1, \ldots, d\beta_p) = \prod_{k=1}^{p} \beta_k^{a_k - 1} e^{-b_k \beta_k} \frac{b_k^{a_k}}{\Gamma(a_k)} d\beta_k \tag{17}$$

where $a_k > 0$, $b_k > 0$, $k = 1, \ldots, p$. ρ_0 is set to be the mean measure of the jump part of a generalized gamma process (GGP) (Hougaard, 1986; Brix, 1999), which has been extensively used in BNP models due to its generality, the interpretability of its parameters and its attractive conjugacy properties (James, 2002; Lijoi et al., 2007; Saeedi and Bouchard-Côté, 2011; Caron, 2012; Caron et al., 2014). The Lévy measure in this case is

$$\rho_0(dw_0) = \frac{1}{\Gamma(1-\sigma)} w_0^{-1-\sigma} \exp(-w_0 \tau) dw_0 \tag{18}$$

where the parameters (σ, τ) satisfy

$$\sigma \in (0,1), \tau \geq 0 \quad \text{or} \quad \sigma \in (-\infty, 0], \tau > 0. \tag{19}$$

The gamma process ($\sigma = 0$), the inverse Gaussian process ($\sigma = \frac{1}{2}$) and the stable process ($\sigma \in (0,1)$, $\tau = 0$) are special cases. Using (17) and (18), we can write the multivariate Lévy measure in the following analytic form

$$\rho(dw_1, \ldots, dw_p) = \frac{2}{\Gamma(1-\sigma)} \left[\prod_{k=1}^{p} \frac{w_k^{a_k - 1} b_k^{a_k}}{\Gamma(a_k)} \right] \left(\frac{\tau}{\sum_{k=1}^{p} b_k w_k} \right)^{\frac{\kappa}{2}} K_\kappa \left(2\sqrt{\tau \sum_k b_k w_k} \right) dw_1 \ldots dw_p \tag{20}$$

where $\kappa = \sigma + \sum_{k=1}^{p} a_k$ and $K_\nu(z)$ is the modified Bessel function of the second kind. Throughout the paper we call the overall model a compound generalized gamma process (CGGP).

3. Properties and Simulation

In this section we provide the theorems for the graph properties and the simulation algorithm for our class of random graph models. The first theorem provides expressions for the expected number of edges in the multigraph and simple graph, and for the expected number of nodes. The proof is given in Section D of the supplementary material.

Theorem 3 *The expected number of edges in the multigraph D^*_α, edges in the undirected graph $N^{(e)}_\alpha$ and observed nodes N_α are given as follows*

$$\begin{aligned}
\mathbb{E}[D^*_\alpha] &= \alpha^2 \mu^T \mu + \alpha tr(\Sigma) \\
\mathbb{E}[N^{(e)}_\alpha] &= \alpha \int_{\mathbb{R}^p_+} \left(1 - e^{-w^T w}\right) \rho(dw_1, \ldots, dw_p) + \frac{\alpha^2}{2} \int_{\mathbb{R}^p_+} \psi(2w_1, \ldots, 2w_p) \rho(dw_1, \ldots, dw_p) \\
\mathbb{E}[N_\alpha] &= \alpha \int_{\mathbb{R}^p_+} \left(1 - e^{-w^T w - \alpha \psi(2w)}\right) \rho(dw_1, \ldots, dw_p)
\end{aligned}$$

where $\mu = \int_{\mathbb{R}^p_+} w \rho(dw_1, \ldots, dw_p)$, $\Sigma = \int_{\mathbb{R}^p_+} w w^T \rho(dw_1, \ldots, dw_p)$ *and* $\psi(t_1, \ldots, t_p) = \int_{\mathbb{R}^p_+} (1 - e^{-\sum_{k=1}^p t_i w_i}) \rho(dw_1, \ldots, dw_p)$ *is the multivariable Laplace exponent for* $(t_1, \ldots, t_p) \in (0, \infty)^p$.

3.1. Exchangeability

The point process Z defined by (11) is jointly exchangeable in the sense of Kallenberg (1990, 2005). For any $h > 0$ and any permutation π of $\mathbb{N}$

$$(Z(A_i \times A_j)) \overset{d}{=} (Z(A_{\pi(i)} \times A_{\pi(j)})) \text{ for } (i, j) \in \mathbb{N}^2 \tag{21}$$

where $A_i = [h(i-1), hi]$. This follows directly from the fact that the vector of CRMs $(W_1, \ldots, W_p)$ has independent and identically distributed increments, hence

$$(W_1(A_i), \ldots, W_p(A_i)) \overset{d}{=} (W_1(A_{\pi(i)}), \ldots, W_p(A_{\pi(i)})). \tag{22}$$

The model thus falls into the general representation theorem for exchangeable point processes (Kallenberg, 1990).

3.2. Sparsity

Here, we derive the sparsity properties of our graph model, first for the general construction given in Section 2.1, then for the specific construction based on compound CRMs of Section 2.2. Following the notation in Caron and Fox (2017), let Z_α be the restriction of Z to the box $[0, \alpha]^2$. Let $(N_\alpha)_{\alpha \geq 0}$ and $(N^{(e)}_\alpha)_{\alpha \geq 0}$ be counting processes corresponding to the number of nodes and edges in Z_α respectively:

$$\begin{aligned}
N_\alpha &= \sum_i \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{(\sum_j z_{ij} \mathbb{1}_{\theta_j \leq \alpha}) \geq 1} \\
N^{(e)}_\alpha &= \sum_{i \leq j} z_{ij} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_j \leq \alpha}.
\end{aligned}$$

Note that in the propositions below, we discard the trivial case $\int_{\mathbb{R}^p_+} \rho(dw_1, \ldots, dw_p) = 0$ which implies $N^{(e)}_\alpha = N_\alpha = 0$ a.s.

General construction. The next proposition characterizes the sparsity properties of the random graph depending on the properties of the Lévy measure ρ. In particular, if

$$\int_{\mathbb{R}_+^p} \rho(dw_1, \ldots, dw_p) = \infty, \tag{23}$$

then for any $\alpha > 0$, there is a.s. an infinite number of $\theta_i \in [0, \alpha]$ for which $\sum_k w_{ik} > 0$ and the vector of CRMs is called infinite-activity. Otherwise, it is called finite-activity. The proof is given in Section B of the supplementary material.

Proposition 4 *Assume that, for any $k = 1, \ldots, p$, the following condition is satisfied*

$$\int_{\mathbb{R}_+^p} w_k \rho(dw_1, \ldots, dw_p) < \infty. \tag{24}$$

Then it holds that,

$$N_\alpha^{(e)} = \begin{cases} \Theta(N_\alpha^2) & \text{if } (W_1, \ldots, W_p) \text{ is finite-activity} \\ o(N_\alpha^2) & \text{otherwise} \end{cases}$$

a.s. as α tends to ∞.

Construction based on CCRMs. For the CCRM Lévy measure (13), the sparsity properties are solely tuned by the base Lévy measure ρ_0. Ignoring trivial degenerate cases for the score distribution F, it is easily shown that the CCRM model defined by (5) and (13) is infinite-activity iff the Lévy measure ρ_0 verifies

$$\int_0^\infty \rho_0(dw) = \infty. \tag{25}$$

In this case all CRMs $W_0, W_1, \ldots, W_p$ are infinite-activity. Otherwise they are all finite-activity and the vector of CRMs is finite-activity. In the particular case of a CCRM with independent gamma distributed scores (17) and generalized gamma process base measure (18), the condition (25) is satisfied whenever $\sigma \geq 0$. The next proposition characterizes the sparsity of the network depending on the properties of the base Lévy measure ρ_0.

Proposition 5 *Assume that*

$$\int_0^\infty w_0 \rho_0(dw_0) < \infty \tag{26}$$

and F is not degenerated at 0. Then

$$N_\alpha^{(e)} = \begin{cases} \Theta(N_\alpha^2) & \text{if } \int_0^\infty \rho_0(dw) < \infty \\ o(N_\alpha^2) & \text{otherwise} \end{cases}$$

a.s. as α tends to ∞. Furthermore, if the tail Lévy intensity $\overline{\rho}_0$ defined by

$$\overline{\rho}_0(x) = \int_x^\infty \rho_0(dw), \tag{27}$$

is a regularly varying function, i.e.

$$\frac{\overline{\rho}_0(x)}{x^{-\sigma}\ell(1/x)} \longrightarrow 1 \text{ as } x \to 0$$

for some $\sigma \in (0, 1)$ where ℓ is a slowly varying function verifying $\lim_{t\to\infty} \ell(at)/\ell(t) = 1$ for any $a > 0$ and $\lim_{t\to\infty} \ell(t) > 0$, then

$$N_\alpha^{(e)} = O(N_\alpha^{2/(1+\sigma)})$$

a.s. as α tends to ∞. In the particular case of a CCRM with independent gamma distributed scores (17) and generalized gamma process base measure (18), condition (26) is equivalent to having $\tau > 0$. In this case, we have

$$N_\alpha^{(e)} = \begin{cases} \Theta(N_\alpha^2) & \text{if } \sigma < 0 \\ o(N_\alpha^2) & \text{if } \sigma \geq 0 \\ O(N_\alpha^{2/(1+\sigma)}) & \text{if } \sigma \in (0, 1). \end{cases}$$

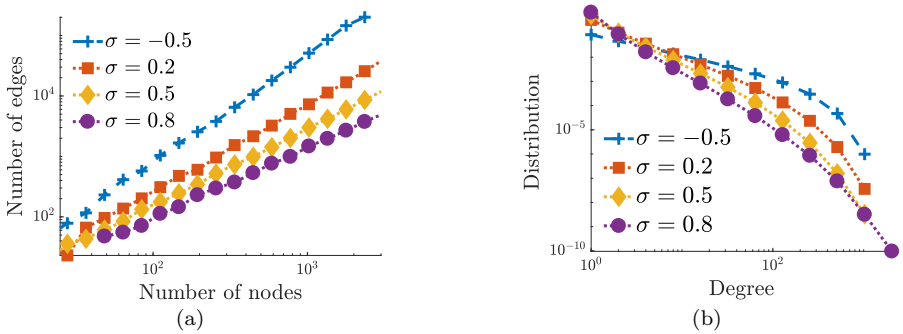


FIG 5. *Empirical analysis of the properties of CCRM based graphs generated with parameters* $p = 2$, $\tau = 1$, $a_k = 0.2$, $b_k = \frac{1}{p}$ *and averaging over various* α. *(a) Number of edges versus the number of nodes and (b) degree distributions on a log-log scale for various* σ: *one finite-activity CCRM* ($\sigma = -0.5$) *and three infinite-activity CCRMs* ($\sigma = 0.2$, $\sigma = 0.5$ *and* $\sigma = 0.8$). *In (a) we note growth at a rate* $\Theta(N_\alpha^2)$ *for* $\sigma = -0.5$ *and* $O(N_\alpha^{2/(1+\sigma)})$ *for* $\sigma \in (0, 1)$.

The proof is given in Section B of the supplementary material. Figure 5(a) provides an empirical illustration of Proposition 5 for a CCRM with independent gamma scores and generalized gamma based Lévy measure. Figure 5(b) shows empirically that the degree distribution exhibits a power-law behaviour when $\sigma \in (0, 1)$.

3.3. Simulation

Having presented the properties of our constructions we now explain how to simulate such graphs. The point process Z is defined on the plane. We describe in this section how to sample a realization Z_α which is a restriction of Z to the box $[0, \alpha]^2$.

General construction. The hierarchical construction given by Eq. (11) suggests a direct way to sample from the model:

1. Sample $(w_{i1}, \ldots, w_{ip}, \theta_i)_{i=1,2,\ldots}$ from a Poisson process with mean measure $\nu(dw_1, \ldots, dw_p, d\theta)\mathbb{1}_{\theta\in[0,\alpha]}$.
2. For each pair of points, sample z_{ij} from (12).

There are two caveats to this strategy. First, for infinite-activity CRMs, the number of points in $\mathbb{R}_+^p \times [0, \alpha]$ is a.s. infinite. Even for finite-activity CRMs, it may be so large that it is not practically feasible. We need therefore to resort to an approximation, by sampling from a Poisson process with a truncated mean measure $\nu^\varepsilon(dw_1, \ldots, dw_p, d\theta)\mathbb{1}_{\theta\in[0,\alpha]} = \rho^\varepsilon(dw_1, \ldots, dw_p)\lambda(d\theta)\mathbb{1}_{\theta\in[0,\alpha]}$ where

$$\int_{\mathbb{R}_+^p} \rho^\varepsilon(dw_1, \ldots, dw_p) < \infty$$

and $\varepsilon > 0$ controls the level of approximation. The approximation is specific to the choice of the mean measure, and we describe such an approximation for CCRMs below. The second caveat is that, in order to apply Eq. (12), we need to consider all pairs $i \leq j$, a task that can be computationally problematic. We can instead, similarly to Caron and Fox (2017), use the hierarchical Poisson construction below:

1. Sample $(w_{i1}, \ldots, w_{ip}, \theta_i)_{i=1,2,\ldots,K}$ from a Poisson process with mean measure $\nu^\varepsilon(dw_1, \ldots, dw_p, d\theta)\mathbb{1}_{\theta\in[0,\alpha]}$. Let $W_{k,\alpha}^\varepsilon = \sum_{i=1}^K w_{ik}\delta_{\theta_i}$ be the associated truncated CRMs and $W_{k,\alpha}^{\varepsilon *} = \sum_{i=1}^K w_{ik}$ their total masses.
2. For $k = 1, \ldots, p$, sample $D_{k,\alpha}^* | W_{k,\alpha}^{\varepsilon *} \sim \text{Poisson}\left(\left(W_{k,\alpha}^{\varepsilon *}\right)^2\right)$.
3. For $k = 1, \ldots, p$, $\ell = 1, \ldots, D_{k,\alpha}^*$, $j = 1, 2$, sample $U_{k\ell j} | W_{k,\alpha}^\varepsilon \overset{\text{ind}}{\sim} \frac{W_{k,\alpha}^\varepsilon}{W_{k,\alpha}^{\varepsilon *}}$.
4. Set $D_{k,\alpha}^\varepsilon = \sum_{\ell=1}^{D_{k,\alpha}^*} \delta_{U_{k\ell 1}, k\ell 2}$.
5. Obtain Z from $(D_1, \ldots, D_p)$ as in (11).

Construction based on CCRMs. The hierarchical construction of compound CRMs suggests an algorithm to simulate a vector of CRMS. We consider the following (truncated) mean measure

$$\rho^{\varepsilon}(dw_1,\ldots,dw_p)=\int_{\varepsilon}^{\infty} w_0^{-p}F\left(\frac{dw_1}{w_0},\ldots,\frac{dw_p}{w_0}\right)\rho_0(dw_0) \tag{28}$$

with $\varepsilon \geq 0$. We can sample from the (truncated) CCRM in step 1 above as follows

1. (a) Sample $(w_{i0},\theta_i)_{i=1,\ldots,K}$ from a Poisson point process with mean measure $\rho_0(dw_0)\lambda(d\theta)\mathbb{1}_{\{w_0>\varepsilon,\theta\in[0,\alpha]\}}$.
   (b) For $i=1,\ldots,K$ and $k=1,\ldots,p$, set $w_{ik}=\beta_{ik}w_{i0}$ where $(\beta_{i1},\ldots,\beta_{ip})$ is drawn from F.

The truncation level ε is set to 0 for finite-activity CCRMs, and $\varepsilon>0$ otherwise. Note that the generalized gamma process can be extended to the case of a tilted generalized gamma process. We explain in Section F of the supplementary material how to perform step 1.(a) in the case of a tilted generalized gamma process.

4. Posterior inference

In this section, we describe a MCMC algorithm for posterior inference of the model parameters and hyperparameters in the statistical network model defined in Section 2. We first describe the data augmentation scheme and characterization of conditionals. We then describe the sampler for a general Lévy measure ρ, and finally derive the sampler for compound CRMs.

4.1. Characterization of conditionals and data augmentation

Assume that we have observed a set of connections $(z_{ij})_{1\leq i,j\leq N_\alpha}$, where N_α is the number of nodes with at least one connection. We aim at inferring the positive parameters $(w_{i1},\ldots,w_{ip})_{i=1,\ldots,N_\alpha}$ associated to the nodes with at least one connection. We also want to estimate the positive parameters associated to the other nodes with no connection. The number of such nodes may be large, and even infinite for infinite-activity CRMs; but under our model, these parameters are only identifiable through their sum, denoted by $(w_{*1},\ldots,w_{*p})$. Note that the node locations θ_i are not likelihood-identifiable, and we will not try to infer them. We assume that there is a set of unknown hyperparameters ϕ of the mean intensity ρ, with prior $p(\phi)$. We assume that the Lévy measure ρ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, and write simply $\rho(dw_1,\ldots,dw_p;\phi)=\rho(w_1,\ldots,w_p;\phi)dw_1\ldots dw_p$. The parameter α is also assumed to be unknown and we place a prior $\alpha\sim\text{Gamma}(a_\alpha,b_\alpha)$ with $a_\alpha>0,b_\alpha>0$. Our objective is to approximate $p\left((w_{1k},\ldots,w_{N_\alpha k},w_{*k})_{k=1,\ldots,p},\phi,\alpha|(z_{ij})_{1\leq i,j\leq N_\alpha}\right)$.

As a first step, we characterize the conditional distribution of the restricted vector of CRMs $(W_{1\alpha},\ldots,W_{p\alpha})$ given the restricted measures $(D_{1\alpha},\ldots,D_{p\alpha})$. Proposition 6 below extends Theorem 12 in Caron and Fox (2017) to the multivariate setting.

Proposition 6 *Let $(\theta_1,\ldots,\theta_{N_\alpha})$, $N_\alpha\geq 0$ be the support points of $(D_{1\alpha},\ldots,D_{p\alpha})$, with*

$$D_{k\alpha}=\sum_{1\leq i,j\leq N_\alpha} n_{ijk}\delta_{(\theta_i,\theta_j)}.$$

The conditional distribution of $(W_{1\alpha},\ldots,W_{p\alpha})$ given $(D_{1\alpha},\ldots,D_{p\alpha})$ is equivalent to the distribution of

$$\left(\widetilde{W}_1+\sum_{i=1}^{N_\alpha}w_{i1}\delta_{\theta_i},\ldots,\widetilde{W}_p+\sum_{i=1}^{N_\alpha}w_{ip}\delta_{\theta_i}\right) \tag{29}$$

where $(\widetilde{W}_1,\ldots,\widetilde{W}_p)$ is a vector of discrete random measures, which depends on $(D_{1\alpha},\ldots,D_{p\alpha})$ only through the total masses $w_{*k}=\widetilde{W}_k([0,\alpha])$.

The set of weights $(w_{ik})_{i=1,\dots,N_\alpha;k=1,\dots,p}$ *and* $(w_{*k})_{k=1,\dots,p}$ *are dependent, with joint conditional distribution*

$$p((w_{1k},\dots,w_{N_\alpha k},w_{*k})_{k=1,\dots,p}|(n_{ijk})_{1\leq i,j\leq N_\alpha;k=1,\dots,p},\phi,\alpha)$$
$$\propto \left[\prod_{i=1}^{N_\alpha}\prod_{k=1}^{p} w_{ik}^{m_{ik}}\right] e^{-\sum_{k=1}^{p}(w_{*k}+\sum_{i=1}^{N_\alpha} w_{ik})^2}\left[\prod_{i=1}^{N_\alpha}\rho(w_{i1},\dots,w_{ip};\phi)\right]\alpha^{N_\alpha} g_{*\alpha}(w_{*1},\dots,w_{*p};\phi) \tag{30}$$

where $m_{ik}=\sum_{j=1}^{N_\alpha} n_{ijk}+n_{jik}$ *and* $g_{*\alpha}(w_{*1},\dots,w_{*p};\phi)$ *is the probability density function of the random vector* $(W_1([0,\alpha]),\dots,W_p([0,\alpha]))$.

The proof can be straightforwardly adapted from that of Caron and Fox (2017), or from Proposition 5.2 of James (2014) and is omitted here. It builds on other posterior characterizations of Bayesian nonparametric models (Prünster, 2002; James, 2002, 2005; James et al., 2009).

Data augmentation. Similarly to Caron and Fox (2017), we introduce latent count variables $\widetilde{n}_{ijk}=n_{ijk}+n_{jik}$

$$(\widetilde{n}_{ij1},\dots,\widetilde{n}_{ijp})|w,z\sim\begin{cases}\delta_{(0,\dots,0)} & \text{if } z_{ij}=0\\ \text{tPoisson}(2w_{i1}w_{j1},\dots,2w_{ip}w_{jp}) & \text{if } z_{ij}=1,\ i\neq j\end{cases}$$
$$\left(\frac{\widetilde{n}_{ij1}}{2},\dots,\frac{\widetilde{n}_{ijp}}{2}\right)|w,z\sim \text{tPoisson}(w_{i1}^2,\dots,w_{ip}^2) \quad \text{if } z_{ij}=1,\ i=j \tag{31}$$

where $\text{tPoisson}(\lambda_1,\dots,\lambda_p)$ is the multivariate Poisson distribution truncated at zero, whose probability mass function (pmf) is

$$\text{tPoisson}(x_1,\dots x_p;\lambda_1,\dots,\lambda_p)=\frac{\prod_{k=1}^{p}\text{Poisson}(x_k;\lambda_k)}{1-\exp(-\sum_{k=1}^{p}x_k\lambda_k)}\mathbb{1}_{\{\sum_{k=1}^{p}x_k>0\}}.$$

One can sample from this distribution by first sampling $x=\sum_{k=1}^{p}x_k$ from a zero-truncated Poisson distribution with rate $\sum_{k=1}^{p}\lambda_k$, and then $(x_1,\dots,x_p)|(\lambda_1,\dots,\lambda_p),x\sim\text{Multinomial}\left(x,\left(\frac{\lambda_1}{\sum\lambda_k},\dots\frac{\lambda_p}{\sum\lambda_k}\right)\right)$.

4.2. Markov chain Monte Carlo algorithm: General construction

Using the data augmentation scheme together with the posterior characterization (30), we derive the following MCMC sampler, which uses Metropolis-Hastings (MH) and Hamiltonian Monte Carlo (HMC) updates within a Gibbs sampler, and iterates as described in Algorithm 1.

Algorithm 1 Markov chain Monte Carlo sampler for posterior inference.

At each iteration

1. Update $(w_{i1},\dots,w_{ip})$, $i=1,\dots,N_\alpha$ given the rest using MH or HMC.
2. Update hyperparameters (ϕ,α) and total masses $(w_{*1},\dots,w_{*p})$ given the rest using MH.
3. Update the latent variables given the rest using (31).

In general, if the Lévy intensity ρ can be evaluated pointwise, one can use a MH update for step 1., but it would scale poorly with the number of nodes. Alternatively, if the Lévy intensity ρ is differentiable, one can use a Hamiltonian Monte Carlo update (Duane et al., 1987; Neal, 2011).

The challenging part of Algorithm 1 is Step 2. From Eq. (30) we have the conditional distribution

$$p((w_{*k})_{k=1,\dots,p},\phi,\alpha|\text{rest})\propto p(\phi)p(\alpha)e^{-\sum_{k=1}^{p}(w_{*k}+\sum_{i=1}^{N_\alpha}w_{ik})^2}\left[\prod_{i=1}^{N_\alpha}\rho(w_{i1},\dots,w_{ip};\phi)\right]\alpha^{N_\alpha}g_{*\alpha}(w_{*1},\dots,w_{*p};\phi)$$

which is not of standard form and involves the multivariate pdf $g_{*\alpha}(w_{*1},\dots,w_{*p})$ of the random vector $(W_1([0,\alpha]),\dots,W_p([0,\alpha]))$ for which there is typically no analytic expression. What is available is its Laplace transform,

$$\mathbb{E}\left[e^{-\sum_{k=1}^{p}t_kW_k([0,\alpha])}\right]=e^{-\alpha\psi(t_1,\dots,t_p;\phi)} \tag{32}$$

where

$$\psi(t_1,\dots,t_p;\phi)=\int_{\mathbb{R}_+^p}\left(1-e^{-\sum_{k=1}^p t_k w_k}\right)\rho(dw_1,\dots,dw_p;\phi) \tag{33}$$

is the multivariate Laplace exponent, which involves a p-dimensional integral. We propose to use a Metropolis-Hastings step, with proposal

$$q(\widetilde{w}_{*1:p},\widetilde{\phi},\widetilde{\alpha}|w_{*1:p},\phi,\alpha)=q(\widetilde{w}_{*1:p}|w_{*1:p},\widetilde{\phi},\widetilde{\alpha})\times q(\widetilde{\phi}|\phi)\times q(\widetilde{\alpha}|\alpha,\widetilde{\phi},w_{*1:p}) \tag{34}$$

where

$$q(\widetilde{\alpha}|\alpha,\widetilde{\phi},w_{*1:p})=\text{Gamma}(\widetilde{\alpha};a_\alpha+N_\alpha,b_\alpha+\psi(\lambda_1,\dots,\lambda_p;\widetilde{\phi})). \tag{35}$$

The proposal for $w_{*1:p}$ is an exponentially tilted version of $g_{*\alpha}$

$$q((\widetilde{w}_{*k})_{k=1,\dots,p}|(w_{*k})_{k=1,\dots,p},\widetilde{\phi})=\frac{e^{-\sum_{k=1}^p\lambda_k\widetilde{w}_{*k}}g_{*\widetilde{\alpha}}(\widetilde{w}_1,\dots,\widetilde{w}_p;\widetilde{\phi})}{e^{-\widetilde{\alpha}\psi(\lambda_1,\dots,\lambda_p;\widetilde{\phi})}} \tag{36}$$

where $\lambda_k=w_{*k}+2\sum_{i=1}^{N_\alpha}w_{ik}$ and $q(\widetilde{\phi}|\phi)$ can be freely specified by the user. This leads to the following acceptance ratio

$$r=\frac{p(\widetilde{\phi})q(\phi|\widetilde{\phi})}{p(\phi)q(\widetilde{\phi}|\phi)}\left[\prod_{i=1}^{N_\alpha}\frac{\rho(w_{i1},\dots,w_{ip};\widetilde{\phi})}{\rho(w_{i1},\dots,w_{ip};\phi)}\right]\left[\frac{b_\alpha+\psi(\widetilde{\lambda}_1,\dots,\widetilde{\lambda}_p;\phi)}{b_\alpha+\psi(\lambda_1,\dots,\lambda_p;\widetilde{\phi})}\right]^{a_\alpha+N_\alpha}e^{\sum_{k=1}^p[w_{*k}^2-\widetilde{w}_{*k}^2]}$$

where $\widetilde{\lambda}_k=\widetilde{w}_{*k}+2\sum_{i=1}^{N_\alpha}w_{ik}$. This acceptance ratio involves evaluating the multivariate exponent (33).

In the general case, the MCMC algorithm 1 requires

(a) pointwise evaluation of the Lévy intensity ρ, and potentially differentiating it,
(b) pointwise evaluation of the Laplace exponent (33) and
(c) sampling from the exponentially tilted distribution (36).

Regarding requirement (c), the random variable with pdf (36) has the same distribution as the random vector $\left(W_1'([0,\alpha]),\dots,W_p'([0,\alpha])\right)$ where $(W_1',\dots,W_p')\sim\text{CRM}(\rho',\lambda)$ and ρ' is an exponentially tilted version of ρ

$$\rho'(w_1,\dots,w_p)=e^{-\sum_k\lambda_k w_k}\rho(w_1,\dots,w_p). \tag{37}$$

By considering an approximate tilted intensity $\rho^{\varepsilon\,\prime}(w_1,\dots,w_p)$, one can approximately sample from (36) by simulating points from a Poisson process with mean measure $\alpha\rho^{\varepsilon\,\prime}(w_1,\dots,w_p)$ and summing them up.

4.3. Markov chain Monte Carlo algorithm: Compound CRMs

The hierarchical construction of CCRMs enables to derive a certain number of simplifications in the algorithm described in the previous section. Using the construction $w_{ik}=\beta_{ik}w_{i0}$ where the points $(w_{i0},\beta_{i1},\dots,\beta_{ip})_{i=1,2,\dots}$ have Lévy measure (16), we aim at approximating the posterior

$$p\left((w_{10},\dots,w_{N_\alpha 0}),(\beta_{1k},\dots,\beta_{N_\alpha k},w_{*k})_{k=1,\dots,p},\phi,\alpha|(z_{ij})_{1\leq i,j\leq N_\alpha}\right). \tag{38}$$

Conditionally on the latent count variables defined in (31), we have the following conditional characterization, similar to (30)

$$\begin{aligned}p((w_{10},\dots,w_{N_\alpha 0}),(\beta_{1k},\dots,\beta_{N_\alpha k},w_{*k})_{k=1,\dots,p}|(n_{ijk})_{1\leq i,j\leq N_\alpha;k=1,\dots,p},\phi,\alpha)\\ \propto\left[\prod_{i=1}^{N_\alpha}w_{i0}^{m_i}\prod_{k=1}^{p}\beta_{ik}^{m_{ik}}\right]e^{-\sum_{k=1}^p(w_{*k}+\sum_{i=1}^{N_\alpha}w_{ik})^2}\\ \times\left[\prod_{i=1}^{N_\alpha}f(\beta_{i1},\dots,\beta_{ip};\phi)\rho_0(w_{i0};\phi)\right]\alpha^{N_\alpha}g_{*\alpha}(w_{*1},\dots,w_{*p};\phi)\end{aligned} \tag{39}$$

where $m_i = \sum_{k=1}^p m_{ik}$ and f and ρ_0 are the densities of F and ρ_0 with respect to the Lebesgue measure.

If f and ρ_0 are differentiable, one can use a HMC update for Step 1 of Algorithm 1. In particular, when they take the form (17) and (18), we obtain the following simple expressions for the gradient:

$$\frac{\partial U(q)}{d(\log w_{i0})} = m_i - \sigma - w_{i0}\left[\tau + 2\sum_{k=1}^p \beta_{ik}\left(w_{*k} + \sum_{j=1}^{N_\alpha} w_{j0}\beta_{jk}\right)\right], \quad i = 1, \ldots, N_\alpha,$$

$$\frac{\partial U(q)}{d(\log \beta_{ik})} = m_{ik} + a_k - \beta_{ik}\left[b_k + 2w_{i0}\left(w_{*k} + \sum_{j=1}^{N_\alpha} w_{j0}\beta_{jk}\right)\right], \quad i = 1, \ldots, N_\alpha, \quad k = 1, \ldots, p,$$

where $U(q) = \log p\,(q|\text{rest})$ with $q = (\log w_{i0}, \log \beta_{i1}, \ldots, \log \beta_{ip})_{i=1,\ldots,N_\alpha}$.

Regarding Step 2 of Algorithm 1, the Laplace exponent for CCRM takes the simple form

$$\psi(t_1, \ldots, t_p) = \int_0^\infty \left[1 - M(-w_0 t_1, \ldots, -w_0 t_p)\right] \rho_0(dw_0) \tag{40}$$

where M is the (multivariate) Laplace transform of the score distribution F. This only requires evaluating a one-dimensional integral, whatever the number p of communities is, and this can be done numerically. For the specific model defined by (17) and (18), we obtain

$$\psi(t_1, \ldots, t_p) = \frac{1}{\Gamma(1-\sigma)} \int_0^\infty \left[1 - \prod_{k=1}^p \left(1 + \frac{w_0 t_k}{b_k}\right)^{-a_k}\right] w_0^{-1-\sigma} e^{-w_0 \tau} dw_0. \tag{41}$$

Finally, we need to sample total masses $(w_{*1}, \ldots, w_{*p})$ from (36), and this can be done by simulating points $(w_{i0}, \beta_{i1}, \ldots, \beta_{ip})_{i=1,2,\ldots}$ from a Poisson process with exponentially tilted Lévy intensity

$$\alpha e^{-w_0 \sum_{k=1}^p \lambda_k \beta_k} f(\beta_1, \ldots, \beta_p) \rho_0(w_0) \tag{42}$$

and summing up the weights $w_{*k} = \sum_{i=1,2,\ldots} w_{i0}\beta_{ik}$ for $k = 1, \ldots, p$. For infinite-activity CRMs, this is not feasible, and we suggest to resort to the approximation of Cohen and Rosinski (2007). More precisely, we write

$$(w_{*1}, \ldots, w_{*p}) = X_\varepsilon + X^\varepsilon,$$

where the random vectors $X_\varepsilon \in \mathbb{R}_+^p$ and $X^\varepsilon \in \mathbb{R}_+^p$ are defined as $X_\varepsilon = \sum_{i|w_{i0}<\varepsilon} w_{i0}(\beta_{i1}, \ldots, \beta_{ip})$ and $X^\varepsilon = \sum_{i|w_{i0}>\varepsilon} w_{i0}(\beta_{i1}, \ldots, \beta_{ip})$. We can sample a realization of the random vector X^ε exactly by simulating the points of a Poisson process with mean intensity

$$\alpha e^{-w_0 \sum_{k=1}^p \lambda_k \beta_k} f(\beta_1, \ldots, \beta_p) \rho_0(w_0) \mathbb{1}_{w_0 > \varepsilon}. \tag{43}$$

See Section F of the supplementary material for details. The positive random vector X_ε is approximated by a truncated Gaussian random vector with mean μ_ε and variance Σ_ε such that

$$\mu_\varepsilon = \alpha \int_{\mathbb{R}_+^p} w_{1:p} \rho_\varepsilon(dw_1, \ldots, dw_p)$$

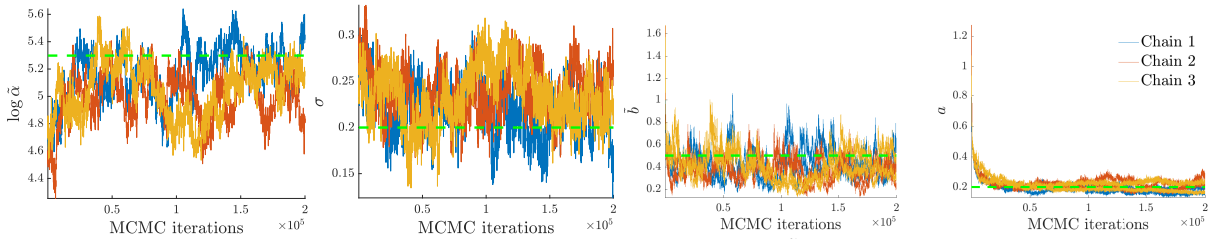
$$\Sigma_\varepsilon = \alpha \int_{\mathbb{R}_+^p} w_{1:p} w_{1:p}^T \rho_\varepsilon(dw_1, \ldots, dw_p)$$

where

$$\rho_\varepsilon(dw_1, \ldots, dw_p) = e^{-\sum_{k=1}^p \lambda_k w_k} \int_0^\varepsilon w_0^{-p} F\left(\frac{dw_1}{w_0}, \ldots, \frac{dw_p}{w_0}\right) \rho_0(dw_0).$$

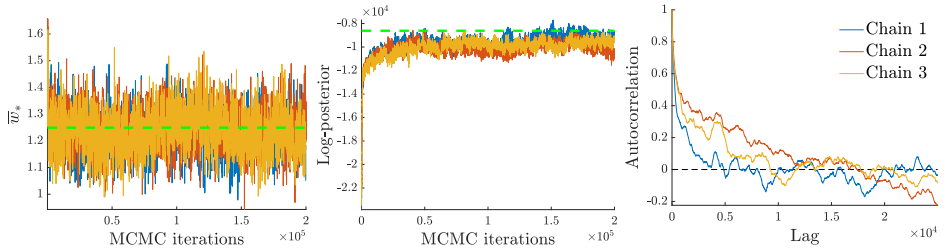
Note that μ_ε and Σ_ε can both be expressed as one-dimensional integrals using the gradient and Hessian of the moment generating function M of F. Theorem S1 in Section H of the supplementary material, which is an adaptation of the results of Cohen and Rosinski (2007) to CCRMs, gives the conditions on the parameters of CCRM under which

$$\Sigma_\varepsilon^{-1/2}(X_\varepsilon - \mu_\varepsilon) \xrightarrow{d} \mathcal{N}(0, I_p) \text{ as } \varepsilon \to 0$$



(a) $\log \tilde{\alpha}$ (b) σ (c) $\tilde{b}$ (d) a

FIG 6. *MCMC trace plots of parameters (a) $\log \tilde{\alpha}$, (b) σ, (c) $\tilde{b}$ and (d) a for a graph generated with parameters $p = 2$, $\alpha = 200$, $\sigma = 0.2$, $\tau = 1$, $b = \frac{1}{p}$, $a = 0.2$. The dotted green line corresponds to the value of the parameters used to generate the graph.*



(a) $\overline{w}_*$ (b) Log-posterior (c) Autocorrelation

FIG 7. *MCMC trace plots of parameters (a) $\overline{w}_*$, (b) trace plot of the log-posterior, (c) autocorrelation of the log-posterior for a graph generated with parameters $p = 2$, $\alpha = 200$, $\sigma = 0.2$, $\tau = 1$, $b = \frac{1}{p}$, $a = 0.2$. The dotted green line in (a) corresponds to the value of $\overline{w}_*$ of the sampled graph, in (b) to the value of the approximate log-posterior (up to a constant) under the model parameters used to generate the graph.*

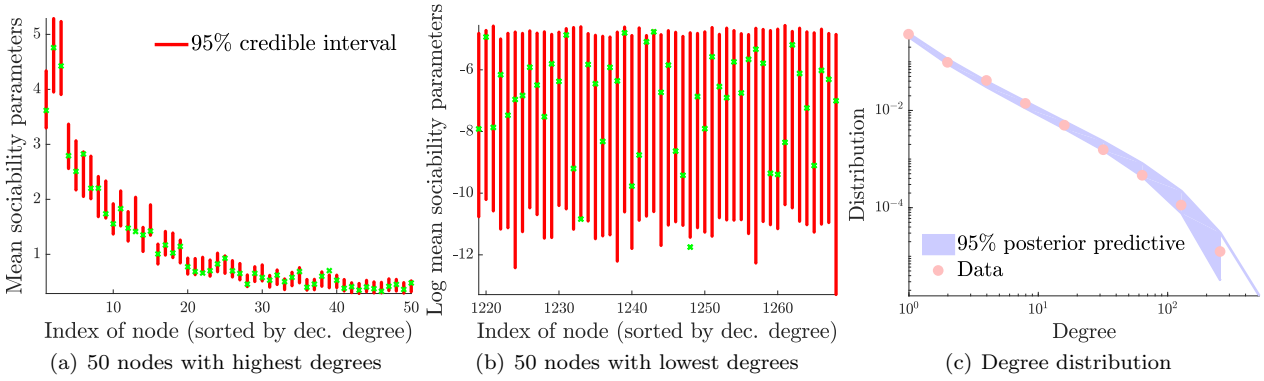
and thus the approximation is asymptotically valid. The Gaussian approximation is asymptotically valid for the CCRM defined by (17) and (18) when $\sigma \in (0, 1)$, hence is valid for all infinite-activity cases except $\sigma = 0$.

Note that due to the Gaussian approximation in the proposal distribution for $(w_{*\alpha})$, Algorithm 1 does not actually admit the posterior distribution (38) as invariant distribution, and is an approximation of an exact MCMC algorithm targeting this distribution. We observe in the next section that this approximation provides very reasonable results for the experiments considered. See also Sections J and K.1 in the supplementary material for additional empirical analysis of this approximation.

5. Experiments

5.1. Simulated data

We first study the convergence of the MCMC algorithm on synthetic data simulated from the CCRM based graph model described in Section 2 where F and ρ_0 take the form (17) and (18). We generate an undirected graph with $p = 2$ communities and parameters $\alpha = 200$, $\sigma = 0.2$, $\tau = 1$, $b_k = b = \frac{1}{p}$, $a_k = a = 0.2$. The sampled graph has 1268 nodes and 7228 edges. For the inference, we consider that b is known and we assume a vague prior Gamma$(0.01, 0.01)$ on the unknown parameters α and $\phi = (1-\sigma, \tau, a)$. We run 3 parallel MCMC chains with different initial values. Each chain starts with 10,000 iterations using our model with only one community where the scores β are fixed to 1, which is equivalent to the model of Caron and Fox (2017). We then run 200,000 iterations using our model with p communities. We use $\varepsilon = 10^{-3}$ as a truncation level for simulating $w_{*1:p}$ and $\mathcal{L} = 10$ leapfrog steps for the HMC. The stepsizes of both the HMC and the random walk MH on $(\log(1-\sigma), \log\tau, \log a)$ are adapted during the first 50,000 iterations so that to target acceptance ratios of 0.65 and 0.23 respectively. We give the trace plots of the identifiable parameters which are $\log\tilde{\alpha}$, σ, a, $\tilde{b}$ and $\overline{w}_*$, where $\tilde{\alpha} = \alpha\tau^\sigma$ and $\tilde{b} = b\tau$. The trace plots for the MCMC output are given in Figure 6 and Figure 7(a). The dotted green line in these, corresponds to the value of the model parameters used for generating the graph. Figure 7(b) and (c) show the trace and autocorrelation plots of the log-posterior



(a) 50 nodes with highest degrees (b) 50 nodes with lowest degrees (c) Degree distribution

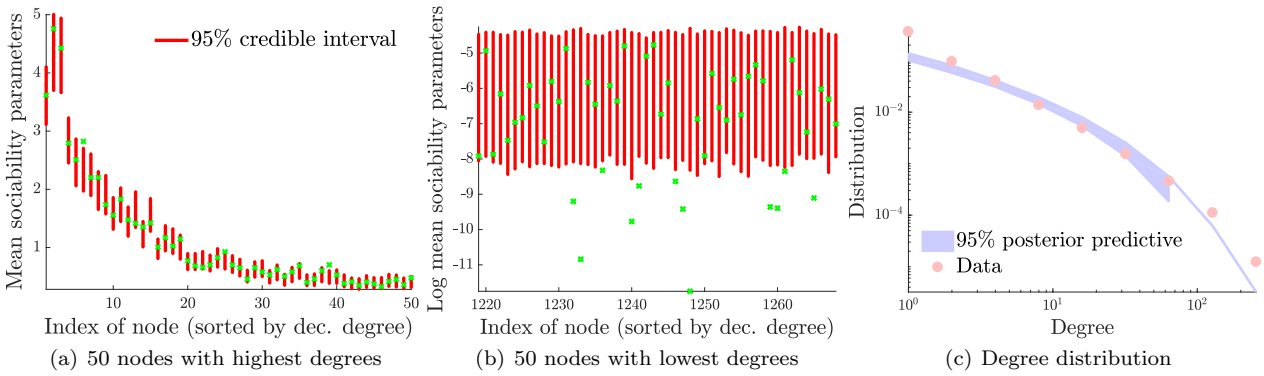
FIG 8. *95% posterior credible intervals (red) and true values (green) of (a) the mean parameters* $\overline{w}_i = \frac{1}{p}\sum_{k=1}^{p} w_{ik}$ *of the 50 nodes with highest degrees and (b) the log mean parameters* $\log \overline{w}_i$ *of the 50 nodes with lowest degrees. (c) Empirical degree distribution (red) and 95% posterior predictive credible interval (blue). Results obtained for a graph generated with parameters* $p = 2$, $\alpha = 200$, $\sigma = 0.2$, $\tau = 1$, $b = \frac{1}{p}$, $a = 0.2$.

probability density function (up to a normalizing constant). The dotted green line corresponds to the value of the approximate log-posterior computed using the model parameters. For details on the calculation of the approximate log-posterior, see Eq. (S14) in Section E in the supplementary material. The log-posterior trace suggests convergence of the chains and the autocorrelation plot shows that the correlation of samples decreases quickly with increasing lag.

Our model is able to accurately recover the mean parameters of both low and high degree nodes and to provide reasonable credible intervals, as shown in Figure 8(a)-(b). By generating 500 graphs from the posterior predictive we assess the performance of our model regarding degree distribution fit. As shown in Figure 8(c), the model's posterior degree distribution fits the empirical power-law degree distribution of the sparse generated graph. We demonstrate the interest of our nonparametric approach by comparing these results to the ones obtained with the parametric version of our model. To achieve this, we fix $w_{*k} = 0$ and force the model to lie in the finite-activity domain by assuming $\sigma \in (-\infty, 0)$ and using the prior distribution $-\sigma \sim \text{Gamma}(0.01, 0.01)$. Note that in this case, the model is equivalent to that of Zhou (2015). As shown in Figure 9(a)-(b), the parametric model is able to recover the mean parameters of nodes with high degrees, and credible intervals are similar to those obtained with the full model; however, it fails to provide reasonable credible intervals for nodes with low degree. In addition, as shown in Figure 9(c), the posterior predictive degree distribution does not fit the data, illustrating the lack of ability of this parametric model to capture power-law behaviour.

Aiming to study further empirically the accuracy of our approximate MCMC algorithm, we explore the impact of a lower threshold value ε. This is described in Section K.1 of the supplementary material. We perform inference using the same generated graph and MCMC settings and we vary ε to the values $100, 1, 10^{-4}$ and 10^{-7}. Note that this sensitivity analysis test is relevant only to the positive values of σ since for $\sigma < 0$ we set $\varepsilon = 0$. We show the results for the values of $\sigma = 0.2$ and $\sigma = 0.5$ in Section K.1 of the supplementary material. We provide the trace plots of the parameters $\log \tilde{\alpha}$, σ, a, $\tilde{b}$ and $\overline{w}_*$ as well as the trace and autocorrelation function of the approximate log-posterior density. The plots suggest the samplers converge for all the values of the threshold ε, with no noticeable differences in the mixing time nor limit distributions. Similarly we perform a sensitivity analysis on the number of leapfrog steps $\mathcal{L}$. In particular, we use the values of $\mathcal{L} = 1, 2, 10$ and 20 and perform inference on the same generated graph. For details and the plots see Section K.2 of the supplementary. The plots of the log-posterior trace indicate that the chains with less leapfrog steps ($\mathcal{L} = 1$ and $\mathcal{L} = 2$) need more time to converge, whereas the log-posterior for $\mathcal{L} = 10$ and $\mathcal{L} = 20$ are very close to each other, suggesting that $\mathcal{L} = 10$ offers a good trade-off.

Finally, we also study the convergence of our methods on denser simulated graphs and in particular for negative values of σ. We provide the results in Section L of the supplementary material.



(a) 50 nodes with highest degrees (b) 50 nodes with lowest degrees (c) Degree distribution

FIG 9. *See Figure 8. Results obtained on the same generated graph by inferring a finite-activity model with* $w_{*k} = 0$ *and* $\sigma \leq 0$.

5.2. *Real-world graphs*

We now apply our methods to learn the latent communities of two real-world undirected simple graphs. The first network to be considered, the **polblogs** network (Adamic and Glance, 2005), is the network of the American political blogosphere in February 2005[3]. Two blogs are considered as connected if there is at least one hyperlink from one blog to the other. Additional information on the political leaning of each blog (left/right) is also available. The second network, named **USairport**, is the network of airports with at least one connection to a US airport in 2010[4]. Two airports are considered as connected if there is at least one flight from one airport to the other. Information on the sizes of the two networks are given in Table 1.

TABLE 1
Size of the networks and number of communities.

Name	Nb nodes	Nb edges	Nb communities p
`polblogs`	1224	16,715	2
`USairport`	1574	17,215	4

We assume a vague prior Gamma$(0.01, 0.01)$ on the unknown parameters α, $1-\sigma$, τ,a_k and b_k. In all the experiments we initialize the MCMC posterior from the estimates obtained under the GGP model (corresponding to $p = 1$). We then estimate all parameters under the CGGP model apart from the parameter τ, which we do not estimate to avoid overparameterization (we fix its value to the estimate obtained from the initialization). For the main MCMC inference, we take $p = 2$ communities for `polblogs` and $p = 4$ communities for `USairport`. For `USairport` we run 3 MCMC chains in parallel, each with $10{,}000 + 10{,}000{,}000$ iterations, using the same procedure as used for the simulated data; see Section 5.1. For `polblogs` we run 3 MCMC chains in parallel, each with $10{,}000 + 2{,}000{,}000$ iterations, using the same procedure as used for the simulated data; see Section 5.1.

We interpret the communities based on the minimum Bayes risk point estimate where the cost function is a permutation-invariant absolute loss on the weights $w = (w_{ik})_{i=1,\ldots,N_\alpha;k=1,\ldots,p}$. Let $\mathcal{S}_p$ be the set of permutations of $\{1,\ldots,p\}$ and consider the cost function

$$C(w, w^\star) = \min_{\pi \in \mathcal{S}_p} \left[\sum_{k=1}^{p} \sum_{i=1}^{N_\alpha} \left| w_{i\pi(k)} - w^\star_{ik} \right| + \sum_{k=1}^{p} \left| w_{*\pi(k)} - w^\star_{*k} \right| \right].$$

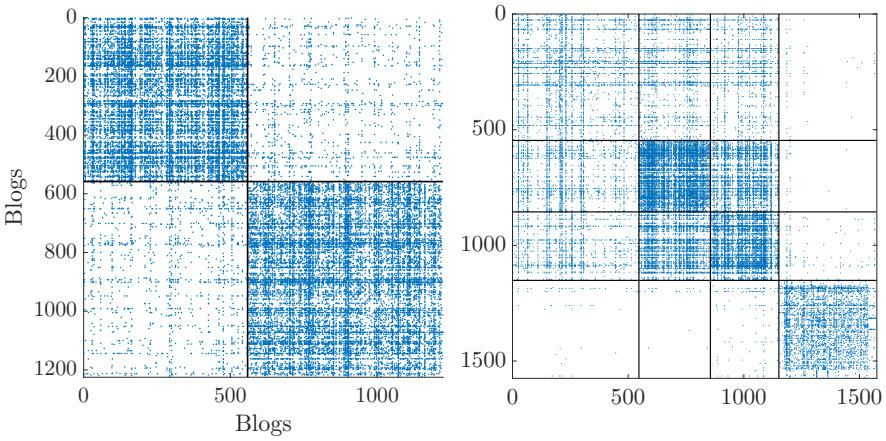
Its evaluation requires solving a combinatorial optimization problem in $O\left(p^3\right)$ using the Hungarian method. We therefore want to solve

$$\widehat{w} = \arg\min_{w^\star} \mathbb{E}\left[C(w, w^\star) \,|Z \right]$$

[3] http://www.cise.ufl.edu/research/sparse/matrices/Newman/polblogs
[4] http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=292

where $\mathbb{E}\left[C\left(w, w^{\star}\right)|Z\right] \simeq \frac{1}{N}\sum_{t=1}^{N} C\left(w^{(t)}, w^{\star}\right)$ and $\left(w^{(t)}\right)_{t=1,\ldots,N}$ are from the MCMC output. For simplicity, we limit the search of $\widehat{w}$ to the set of MCMC samples giving

$$\widehat{w} = \underset{w^{\star}\in\left\{w^{(1)},\ldots,w^{(N)}\right\}}{\arg\min} \frac{1}{N}\sum_{t=1}^{N} C\left(w^{(t)}, w^{\star}\right).$$



(a) `polblogs` (b) `USairport`

FIG 10. *Adjacency matrices of the (a)* `polblogs` *and (b)* `USairport` *networks, reordered by associating each node to the community where it has the highest weight.*

TABLE 2

Nodes with highest weight in each community for the `polblogs` *network. Blog URLs are followed by known political leaning: (L) for left-wing and (R) for right-wing.*

Community 1: "Liberal"	**Community 2:** "Conservative"
dailykos.com (L)	instapundit.com (R)
atrios.blogspot.com (L)	blogsforbush.com (R)
talkingpointsmemo.com (L)	littlegreenfootballs.com/weblog (R)
liberaloasis.com (L)	powerlineblog.com (R)
washingtonmonthly.com (L)	drudgereport.com (R)
politicalstrategy.org (L)	lashawnbarber.com (R)
digbysblog.blogspot.com (L)	hughhewitt.com (R)
bodyandsoul.typepad.com (L)	michellemalkin.com (R)
newleftblogs.blogspot.com(L)	wizbangblog.com (R)
juancole.com (L)	truthlaidbear.com (R)

Table 2 reports the nodes with highest weights in each community for the `polblogs` network. Figure 12 also shows the weight associated to each of the two communities alongside the true left/right class for each blog. The two learned communities, which can be interpreted as "Liberal" and "Conservative", clearly recover the political leaning of the blogs. Figure 10 shows the adjacency matrices obtained by reordering the nodes by community membership, where each node is assigned to the community in which it has the largest weight, showing the block-structure of this network. Figure 11(a)-(b) show the trace plot of the log-posterior for the two datasets. Additional trace plots are reported in Section M of the supplementary material. Note that while the approximate log-posterior density of the different chains is of similar order for the `USairport` network (Figure 11(b)), the posterior output exhibits multiple modes, as seen from Figure S15 in the supplementary material. Figure 11(c)-(d) shows the autocorrelation of the log-posterior which decreases with increasing lag. The obtained clustering yields a 93.95% accuracy when compared to the ground truth classification. Figure 13(a) shows the relative community proportions for a subset of the blogs. Clearly, `dailykos.com` and `washingtonmonthly.com` are described as liberal while `blogsforbush.com`, `instapundit.com` and `drudgereport.com` as conservative. Other more moderate blogs such as`andrewsullivan.com` have more balanced values in both communities. Figure 15(a) shows that the

TABLE 3
Nodes with highest weights in each community for the `USairport` *network.*

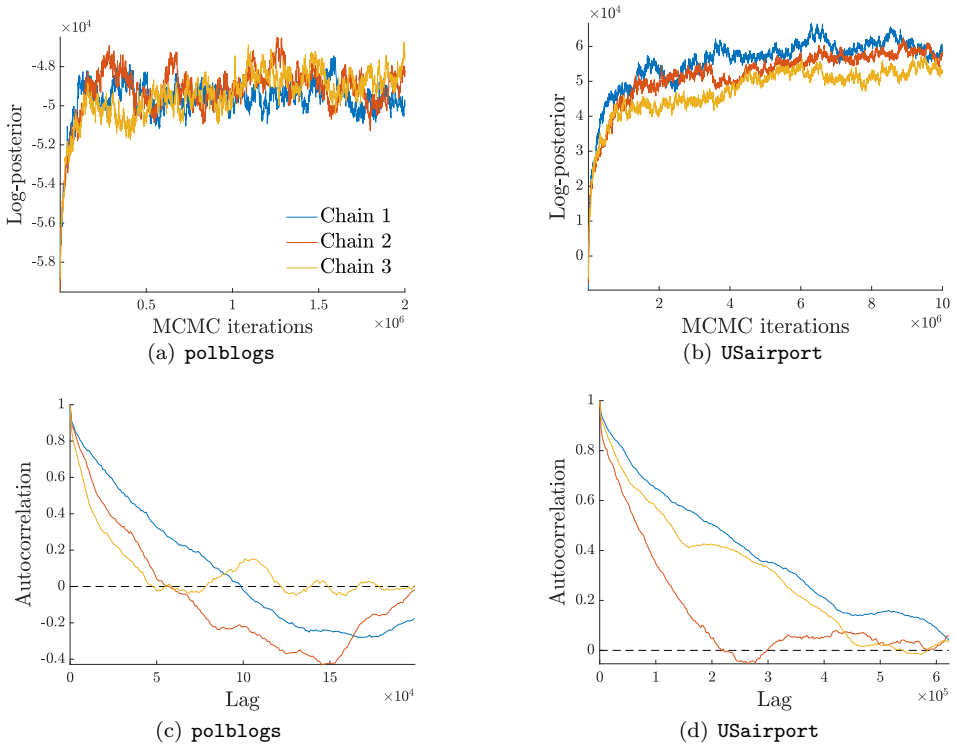
Community 1: "Hub"	**Community 2:** "East"	**Community 3:** "West"	**Community 4:** "Alaska"
New York, NY	Detroit, MI	Denver, CO	Anchorage, AK
Miami, FL	Chicago, IL	Las Vegas, NV	Fairbanks, AK
Boston, MA	Atlanta, GA	Burbank, CA	Bethel, AK
Washington, DC	Cleveland, OH	Salt Lake City, UT	Unalakleet, AK
Newark, NJ	Washington, DC	Phoenix, AZ	St. Mary's, AK
Los Angeles, CA	Philadelphia, PA	Los Angeles, CA	King Salmon, AK
Atlanta, GA	Nashville, TN	Dallas/Fort Worth, TX	Galena, AK
Fort Lauderdale, FL	Pittsburgh, PA	Seattle, WA	Dillingham, AK
Chicago, IL	Columbus, OH	San Francisco, CA	Kotzebue, AK
Minneapolis, MN	Indianapolis, IN	Sacramento, CA	Nome, AK

TABLE 4
Kolmogorov Smirnov statistics for assessing the degree distribution for the `USairport` *network.*

Model	KS reweighted	KS heavy-tail	KS standard
`CCRM`	0.279	1.962	0.0541
`MLFM`	0.466	6.964	0.192
`MMSB`	7.147	26.783	0.675

TABLE 5
Kolmogorov Smirnov statistics for assessing the degree distribution for the `polblogs` *network.*

Model	KS reweighted	KS heavy-tail	KS standard
`CCRM`	0.151	1.581	0.045
`MLFM`	2.455	10.546	0.340
`MMSB`	3.622	21.197	0.606



(a) `polblogs` (b) `USairport`

(c) `polblogs` (d) `USairport`

FIG 11. *Trace plot (left) and autocorrelation plot (right) of the approximated log-posterior probability density (up to a constant) of the (left)* `polblogs` *and (right)* `USairport`.

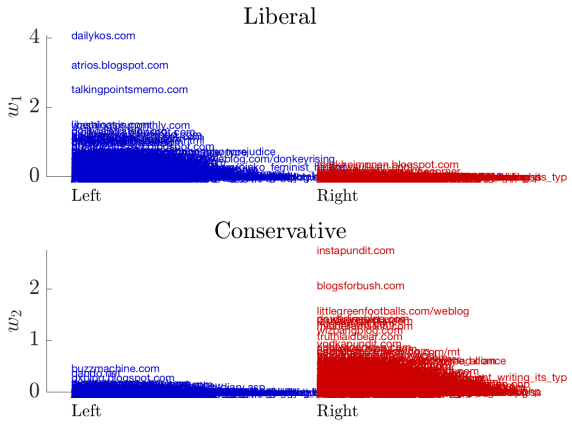


FIG 12. *Level of affiliation of each blog of the* `polblogs` *network to the communities identified as "Liberal" (top) and "Conservative" (bottom). The names of the blogs are grouped according to the left-right wing ground truth. Left-wing blogs are represented in blue on the left, right-wing blogs in red on the right.*

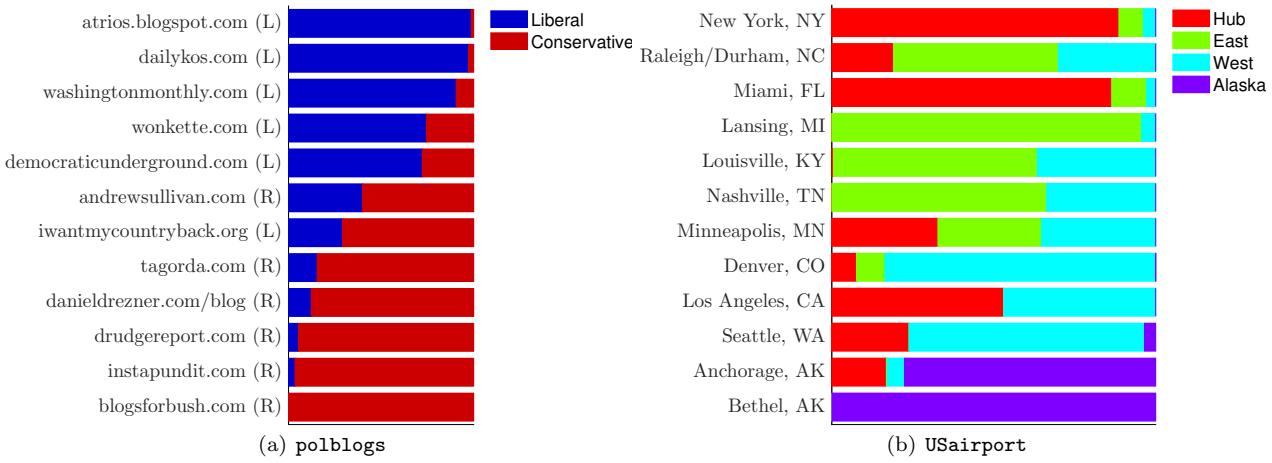


FIG 13. *Relative values of the weights in each community for a subset of the nodes of the (a)* `polblogs` *and (b)* `USairport` *networks.*

posterior predictive degree distribution under our model provides a good fit to the empirical degree of the network dataset.

For `USairport`, the four learned communities can also be easily interpreted, as shown in Table 3. The first community, labeled "Hub", represents highly connected airports with no preferred location, while the three others, labeled "East", "West" and "Alaska", are communities based on the location of the airport. Figure 14 illustrates the affiliations of the airports and shows how the model detects the underlying network communities. In Figure 14 we place the airports on the map and for each airport we normalize the weights to show them in a pie chart form. The size of each pie chart is proportional to the degree of that node. In Figure 13(b), we can see that some airports have a strong level of affiliation in a single community: New York and Miami for "Hub", Lansing for "East", Seattle for "West" and Bethel and Anchorage for "Alaska". Other airports have significant weights in different communities: Raleigh/Durham and Los Angeles are hubs with strong regional connections, Nashville and Minneapolis share a significant number of connections with both East and West of the USA. Anchorage has a significant "Hub" weight, while most airports in Alaska are disconnected from the rest of the world as can be seen in Figure 10(b). "Alaska" appears as a separate block while substantial overlaps are observed between the "Hub", "East" and "West" communities. Under our model the posterior predictive degree distribution provides a good fit to the data, as illustrated in

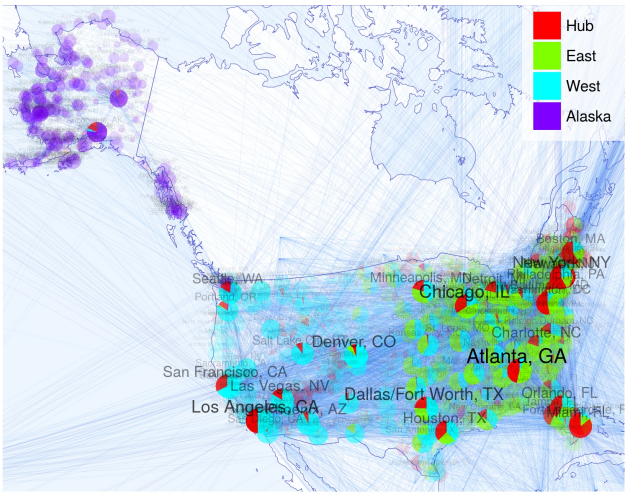


FIG 14. *Usairport network. Each pie chart represents an airport (node) and the size of the pie chart is proportional to the number of connections of that airport. The estimated weights of each airport are normalized so that to sum to 1 and their colors correspond to each uncovered community respectively.*

Figure 15(b).

5.3. Comparisons

We compare the fit and performance of our model to the mixed membership stochastic blockmodel (MMSB) of Airoldi et al. (2008) and the multiplicative latent factor model (MLFM) of Hoff (2009). The two models are briefly explained below.

Multiplicative latent factor model. Let $X \in \mathbb{R}^{N\times N}$ denote a (symmetric) random matrix of effects for a set of N nodes. Under this model, X is explained as the sum of systematic patterns and random noise. Y denotes the adjacency matrix. For $1 \leq i < j \leq N$

$$X_{ij} = \xi_0 + \zeta_i + \zeta_j + M_{ij} + E_{ij}$$
$$Y_{ij} = \mathbb{1}_{X_{ij}>0}$$

where $\xi_0 \in \mathbb{R}$ is the intercept, $\zeta_i \in \mathbb{R}$ is the additive node effect, $M \in \mathbb{R}^{N\times N}$ is the matrix of multiplicative effects. The square symmetric matrix M has a latent decomposition of the form $M = U\Lambda U^T$, where $U \in \mathbb{R}^{N\times p}$ and $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix. $E \in \mathbb{R}^{N\times N}$ is the matrix of standard normal noise; $E_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Denoting by u_i the i^{th} row of U we have $M_{ij} = u_i^T \Lambda u_j$. We use the priors implemented in the amen R package (Hoff et al., 2017). For the additive effect ζ_i it is assumed that $\zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\zeta^2)$, with $1/\sigma_\zeta^2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$. For the multiplicative effects it is assumed that $u_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_j^2)$ with $1/\sigma_j^2 \sim \text{Gamma}(2,1)$. For ξ_0 we use an improper prior $p(\xi_0) \propto 1/\xi_0$.

Mixed membership stochastic blockmodel. Let N be the number of nodes in the network and Y the adjacency matrix. For each node $i = 1, \ldots, N$, let

$$\pi_i \sim \text{Dirichlet}(\varsigma, \ldots, \varsigma)$$

be a p dimensional mixed membership probability vector for node i, with $\varsigma > 0$. For each pair of nodes $i < j$, let

$$c_{ij}|\pi_i \sim \pi_i$$
$$c_{ji}|\pi_j \sim \pi_j$$
$$Y_{ij}|c_{ij}, c_{ji}, B \sim \text{Bernoulli}\left\{(1-\rho)\, B_{c_{ij}c_{ji}}\right\}$$

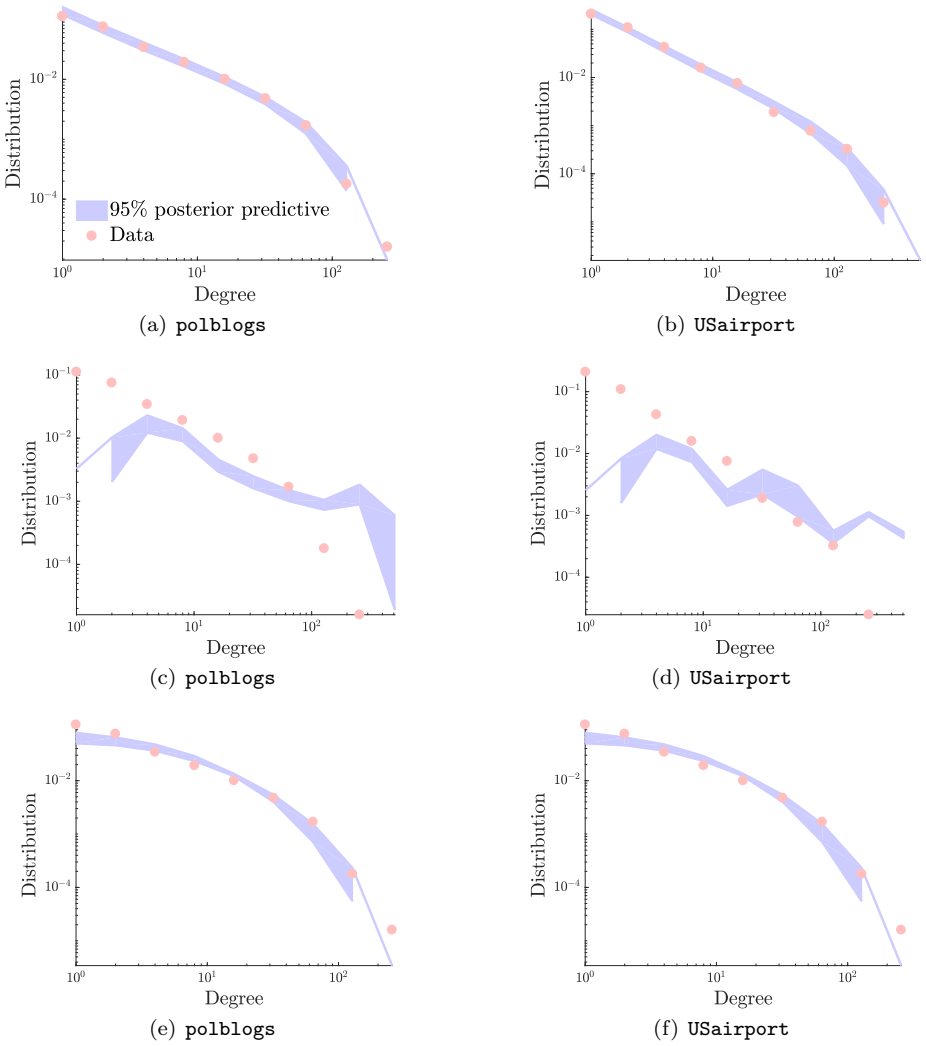


FIG 15. *Empirical degree distribution (red) and posterior predictive (blue) of the (left)* `polblogs` *and (right)* `USairport` *networks under our (top row) CCRM model, (middle row) MMSB and the (bottom row) MLFM.*

where $c_{ij} \in \{1, \dots, p\}$ is the emission indicator variable, ρ is a parameter that controls the proportion of zeros that should not be explained by the blockmodel and B is the $p \times p$ matrix that contains the Bernoulli rates of the link probabilities between different communities, i.e. $B_{k,l}$ is the probability of a connection between a member of group k and one of group l. We assume $\varsigma \sim \text{Gamma}(1,1)$, $\rho \sim \text{Beta}(1/2, 1/2)$ and $B_{k,\ell} \sim \text{Beta}(1,1)$.

For each model, we run three MCMC chains for posterior inference, using the `amen` package (Hoff et al., 2017) for MLFM. Under MMSB, we use 200,000 MCMC iterations, of which 100,000 were discarded as burn in. Under MLFM, which required more iterations to converge, we ran $1M$ iterations, of which 500,000 were discarded as burn-in. In both cases we thinned the output to obtain 500 samples approximately distributed from the posterior distribution.

Posterior predictive checks. In order to evaluate the goodness of fit of each model, we look at some statistics of a replicated network for a **new** set of nodes, sampled under the posterior predictive. Let Y^* be

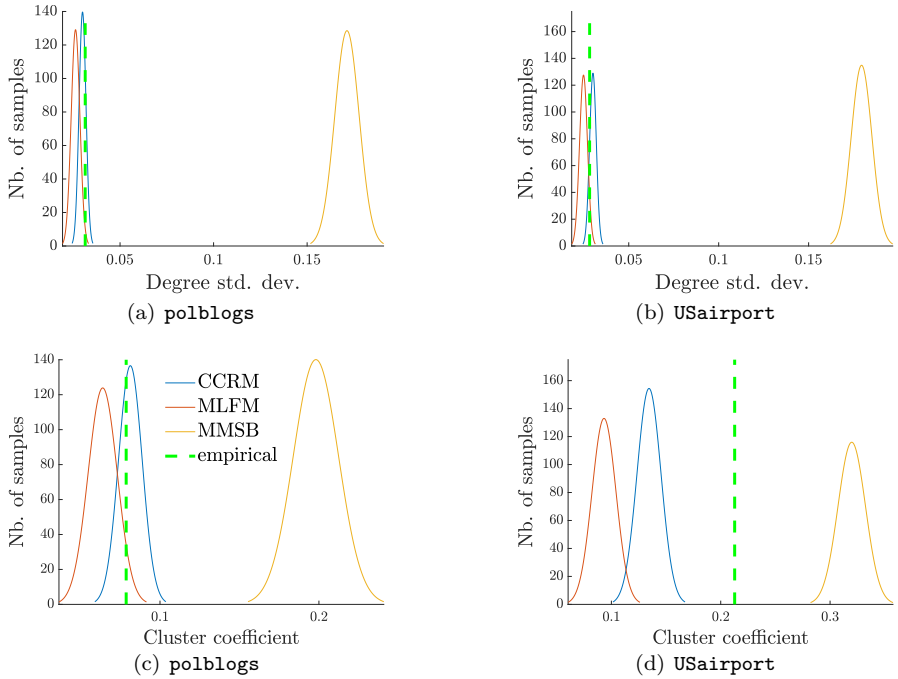


FIG 16. *Posterior predictive plots for the* `polblogs` *and* `USairport` *networks under the CCRM, MLFM and MMSB models. Top row: standard deviation of degree; bottom row: cluster coefficient (or triadic dependence).*

a N by N adjacency matrix corresponding to the edges of a set of N new nodes, sampled from

$$\Pr(Y^*|Y) = \begin{cases} \int \Pr(Y^*|\xi_0, \Lambda) \Pr(d\xi_0, d\Lambda|Y) & \text{for MLFM} \\ \int \Pr(Y^*|\varsigma, \rho, B) \Pr(d\varsigma, d\rho, dB|Y) & \text{for MMSB.} \end{cases}$$

Note that the posterior predictive setting is different from that of Hoff (2009), which replicates a network for the same set of nodes and therefore conditions on the parameters u_i, ζ_i. We generate 500 samples from the posterior predictive under each model. We are interested in some standard summary statistics: the degree distribution, standard deviation of the degrees and cluster coefficient. Posterior degree distributions under each model for `polblogs` and `USairport` are presented in Figure 15 and posterior predictive distributions of the standard deviation of the degree and clustering coefficient in Figure 16. In addition, we compare the models by providing a quantitative assessment for the degree distribution using modified versions of the Kolmogorov Smirnov (KS) test statistics. The standard KS test is the most common test for non-normal data, however when assessing differences between distributions, this test may not be sensitive enough at the extreme limits. Therefore, we use modified KS statistics; namely the reweighted KS statistic as proposed in Clauset et al. (2009) and a heavy-tail reweighted KS proposed in Mason and Schuenemeyer (1983). Let $S(x)$ denote the empirical CDF of the data for the observations and $P(x)$ the CDF under the model of interest. The three statistics are defined in Eq. (44).

$$D^* = \left\{ \begin{array}{l|l} D = \sup_x |S(x) - P(x)| & \text{for KS} \\ D_1 = \sup_x \frac{|S(x)-P(x)|}{\sqrt{P(x)(1-P(x))}} & \text{for reweighted KS} \\ D_2 = \max\{rL, D, rU\} & \text{for heavy-tail KS} \end{array} \right. \tag{44}$$

where $L = \sup_x \frac{P(x)}{S(x)}, U = \sup_x \frac{1-P(x)}{1-S(x)}$ and r is a nonnegative scalar weight. In these tests the null hypothesis is that samples from the empirical and the posterior predictive degree distribution come from the same distribution. The choice of the value of r is set so that the test assigns equal probability of a false rejection (at some level alpha) due to deviations that occur in the lower, middle and upper sample quantiles. A table for the appropriate values for r is provided in Mason and Schuenemeyer (1983). For our comparisons we choose a value of r that corresponds to a significance level alpha of 0.05.

The two modifications of the KS statistic as well as the standard KS statistic, are all shown in Tables 4 and 5, for the `USairport` and `polblogs` respectively. In all three tests the model CCRM has a better fit of the degree distribution compared to the other two models.

Overall, the MMSB did not perform well on the two datasets considered. While being a flexible model, that has been applied successfully to a wide range of real-world networks, the MMSB does not explicitly capture degree heterogeneity; the latent communities recovered (not shown) do not correspond to those recovered by CCRM or MLFM: it tends to cluster nodes according to their degree, explaining the shape of the posterior degree predictive in Figure 15(c)-(d) . Such limitations, due to the lack of degree correction, have also been acknowledged by previous authors (Karrer and Newman, 2011; Gopalan et al., 2013). Note that degree-corrected versions of the MMSB have been proposed by Gopalan et al. (2013), similar to the general class of models discussed in Hoff (2009).

MLFM on the other hand incorporates degree heterogeneity and thus gives a better fit on the degree predictive distribution. It also gives similar results than CCRM and recovers similar latent communities. It underestimates the proportion of nodes with degree one as shown in Figure 15(e)-(f). As a general comment, the CCRM model gives a better fit to the degree distribution overall. In Figure 16, we study further the degree distribution fit, plotting the goodness of fit statistics used in the analyses of Hoff (2009). The first one is the standard deviation of the posterior predictive degree, shown in the first row. For this, the MMSB performs poorly, whereas CCRM and MLFM have similar results, marginally close to the empirical value. The bottom row reports the cluster coefficient, also known as triadic dependence. This measures how much the nodes in a graph tend to cluster together. Although our model's interest is on the overlapping blocks and not in a disjoint clustering, we still report this statistic. In this case, CCRM gives a better fit than the other two in both datasets. In particular, for the dense `polblogs` dataset CCRM is centered on the empirical value with the MLFM being quite close as well, whereas the MMSB is far from the empirical value. For the sparse `USairport` dataset, although none of the models is centered on the empirical value, the CCRM model is the one most close to it whereas the other two are further away on opposite directions.

Supplementary material

The Matlab package SNetOC (Sparse Networks with Overlapping Communities) contains the code for performing simulation and posterior inference as well as the datasets used in this article. It can be downloaded from the url `https://github.com/0xCSML-BayesNP/SNetOC`.

Acknowledgments

The authors thank George Deligiannidis for pointing out the article of Asmussen and Rosiński (2001) and Peter Hoff for his useful comments on the paper. FC acknowledges the support of the European Commission under the Marie Curie Intra-European Fellowship Programme. Part of this work has been supported by the BNPSI ANR project no ANR-13-BS-03-0006-01, the EPSRC grant no EP/M508111/1. and the "Rich, Structured and Efficient Learning of Big Bayesian Models" project no 617071. XM acknowledges the support the A. G. Leventis Foundation.

References

L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

E. M Airoldi, D. Blei, S. E Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

S. Asmussen and J. Rosiński. Approximations of small jumps of Lévy processes with a view towards simulation. *Journal of Applied Probability*, pages 482–493, 2001.

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

O. E. Barndorff-Nielsen, J. Pedersen, and K.-I. Sato. Multivariate subordination, self-decomposability and stability. *Advances in Applied Probability*, 33:160–187, 2001.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *ArXiv preprint arXiv:1601.07134*, 2016.

A. Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4): 929–953, 1999.

F. Caron. Bayesian nonparametric models for bipartite graphs. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2051–2059. Curran Associates, Inc., 2012.

F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 2017.

F. Caron, Y.W. Teh, and T.B. Murphy. Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2):1145–1181, 2014.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

S. Cohen and J. Rosinski. Gaussian approximation of multivariate Lévy processes with applications to simulation of tempered stable processes. *Bernoulli*, 13(1):195–210, 2007.

R. Cont and P. Tankov. *Financial modelling with jump processes*, volume 2. CRC press, 2003.

D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*. Springer Verlag, 2nd edition edition, 2008.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2): 216–222, 1987.

I. Epifani and A. Lijoi. Nonparametric priors for vectors of survival functions. *Statistica Sinica*, pages 1455–1484, 2010.

P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

A. Goldenberg, A.X. Zheng, S.E. Fienberg, and E.M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical Poisson factorization. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

P. K. Gopalan, C. Wang, and D. Blei. Modeling overlapping communities with node popularities. In *Advances in neural information processing systems*, pages 2850–2858, 2013.

J. E. Griffin and F. Leisen. Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):525–545, 2017.

J. E. Griffin, M. Kolossiatis, and M. F. J. Steel. Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B*, 75(3):499–529, 2013.

T. Herlau, M. N. Schmidt, and M. Mørup. Completely random measures for modelling block-structured sparse networks. Technical report, arXiv preprint arXiv:1507.02925, 2015.

P. Hoff, B. Fosdick, A. Volfovsky, and Y. He. *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*, 2017. URL https://CRAN.R-project.org/package=amen. R package version 1.3.

P. D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):261–272, 2009.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2): 109–137, 1983.

P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.

L. James. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics*, pages 1771–1799, 2005.

L. James. Poisson latent feature calculus for generalized Indian buffet processes. Technical report, arXiv:1411.2936, 2014.

L. F. James. Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*, 2002.

L.F. James, A. Lijoi, and I. Prünster. Posterior analysis for normalized random measures with independent

increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.
O. Kallenberg. Exchangeable random measures in the plane. *Journal of Theoretical Probability*, 3(1):81–136, 1990.
O. Kallenberg. *Probabilistic symmetries and invariance principles*. Springer, 2005.
J. Kallsen and P. Tankov. Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis*, 97(7):1551–1572, 2006.
B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.
C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 21, page 381, 2006.
J.F.C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.
E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.
P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, pages 309–336, 2011.
H. S. Lee, D. D.and Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755):788–791, 1999.
F. Leisen and A. Lijoi. Vectors of two-parameter Poisson–Dirichlet processes. *Journal of Multivariate Analysis*, 102(3):482–495, 2011.
F. Leisen, A. Lijoi, and D. Spanó. A vector of Dirichlet processes. *Electronic Journal of Statistics*, 7:62–90, 2013.
A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.
A. Lijoi, B. Nipoti, and I. Prünster. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291, 2014.
D. M. Mason and J. H. Schuenemeyer. A modified kolmogorov-smirnov test sensitive to tail alternatives. 1983.
K. Miller, T. Griffiths, and M. Jordan. Nonparametric latent feature models for link prediction. In *NIPS*, 2009.
R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, volume 2. Chapman & Hall / CRC Press, 2011.
M. Newman. *Networks: an introduction*. OUP Oxford, 2009.
M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.
M. E. J. Newman and G. Reinert. Estimating the number of communities in a network. *Physical review letters*, 117(7):078301, 2016.
K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.
P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intelligence (PAMI)*, 37(2):437–461, 2015.
K. Palla, D. A. Knowles, and Z. Ghahramani. An infinite latent attribute model for network data. In *ICML*, 2012.
I. Prünster. *Random probability measures derived from increasing additive processes and their application to Bayesian statistics*. PhD thesis, University of Pavia, 2002.
I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping community detection using Bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114, 2011.
A. Saeedi and A. Bouchard-Côté. Priors over recurrent continuous time processes. In *Advances in Neural Information Processing Systems*, pages 2052–2060, 2011.
M. Salter-Townshend and T. H. McCormick. Latent space models for multiview network data. Technical report, Technical Report, 2013.
A. V. Skorohod. *Random processes with independent increments*, volume 47. Springer, 1991.
P. Tankov. Dependence structure of spectrally positive multidimensional Lévy processes. *Unpublished manuscript*, 2003.
V. Veitch and D. M. Roy. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.
L. M. Verbrugge. Multiplexity in adult friendships. *Social Forces*, 57(4):1286–1309, 1979.

J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM, 2013.

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

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Artificial Intelligence and Statistics (AISTATS2015), JMLR W&CP*, volume 38, 2015.