Submitted to the Annals of Statistics

NON-EXCHANGEABLE RANDOM PARTITION MODELS FOR MICROCLUSTERING

By Giuseppe Di Benedetto[*], François Caron[*] and Yee Whye Teh[*,†]

University of Oxford[*] and Google Deepmind[†]

Many popular random partition models, such as the Chinese restaurant process and its two-parameter extension, fall in the class of exchangeable random partitions, and have found wide applicability in various fields. While the exchangeability assumption is sensible in many cases, it implies that the size of the clusters necessarily grows linearly with the sample size, and such feature may be undesirable for some applications. We present here a flexible class of non-exchangeable random partition models which are able to generate partitions whose cluster sizes grow sublinearly with the sample size, and where the growth rate is controlled by one parameter. Along with this result, we provide the asymptotic behaviour of the number of clusters of a given size, and show that the model can exhibit a power-law behaviour, controlled by another parameter. The construction is based on completely random measures and a Poisson embedding of the random partition, and inference is performed using a Sequential Monte Carlo algorithm. Experiments on real datasets emphasise the usefulness of the approach compared to a two-parameter Chinese restaurant process.

1. Introduction. Random partitions arise in a wide range of different applications such as Bayesian model-based clustering [43, 56], population genetics [40], ecology [47] or network modelling [7]. A partition of a set $[n] = \{1, \dots, n\}$ is a set of disjoint non-empty subsets $A_{n,j} \subseteq [n]$, $j = 1, \dots, K_n$ with $\cup_j A_{n,j} = [n]$ where $K_n \leq n$ is the number of clusters and $A_{n,j}$ denotes the set of integers in cluster j. A random partition Π_n of $[n]$ is a random variable taking values in the finite set of partitions of $[n]$. A random partition of $\mathbb{N}$ is a sequence $\Pi = (\Pi_n)_{n\geq 1}$ of random partitions of $[n]$, defined on a common probability space, that satisfy the Kolmogorov consistency condition: for every $1 \leq m < n$, Π_n restricted to $[m]$ is Π_m [39, 1, 61, 59]. For many applications, it is important to characterise the properties of the random partition model as the number of items n grows. Of particular importance are the asymptotic behaviour of (i) the number of clusters, (ii) the proportion of clusters of a given size, and (iii) the cluster sizes.

In some contexts a natural and useful assumption is the *exchangeability* of the random partition: Π is said to be exchangeable if for every $n \geq 1$

the distribution of Π_n is invariant to the group of permutations of $[n]$. Arguably the best known exchangeable random partition model is the Chinese Restaurant Process (CRP) [1]. This model has a single parameter, a very simple generative process and well established asymptotic properties; the number of clusters K_n grows logarithmically with n [42], while the proportion of clusters of any given size goes to zero. Such behaviour is not appropriate for some applications such as natural language processing or image segmentation [67, 66], where these proportions typically exhibit a power-law behaviour. This asymptotic property can be achieved by considering the two-parameter CRP [62], another exchangeable random partition model which generalizes the one-parameter CRP. Beyond these two popular models, the class of exchangeable random partitions offers a rich, flexible and tractable framework, including models based on normalised random measures [64, 33, 48, 35], Poisson-Kingman processes [60] or Gibbs-type priors [31, 25, 20, 3].

Although exchangeability is a sensible assumption in many applications, it has strong implications regarding the growth rate of the cluster's sizes: Kingman's representation theorem indeed implies that the size of each cluster grows linearly with the sample size n. As recently noted by Miller et al. [55] and Betancourt et al. [68], this assumption may be unrealistic for some applications, such as entity resolution, which require the construction of random partition models where the cluster sizes grow sublinearly with the sample size; Miller et al. [55] call it the *microclustering* property.

The objective of this article is to present a general class of models for non-exchangeable random partitions of $\mathbb{N}$ which retains the wide range of asymptotic properties of exchangeable partition models, while capturing the microclustering property. The model allows:

- Flexibility in the asymptotic growth rates of (i) the number of cluster, and (ii) the proportion of clusters of a given size, tuned by interpretable parameters; in particular, it is possible to obtain the same growth rates as with the two-parameter CRP, including the power-law regime.
- Flexibility in the asymptotic sublinear growth rates of the cluster sizes, tuned by interpretable parameters.

The paper is organised as follows. In Section 2 we provide background on completely random measures (CRM), exchangeable random partitions, and give a derivation of the partition associated to a normalised completely random measure via a Poissonisation technique. In Section 3 we present our novel class of non-exchangeable random partition models that builds on the same Poissonisation idea. Section 4 develops the properties of this

class of models and posterior inference. In Section 5, we describe how our model can also be used to build sparse random multigraph models with an asymptotic power-law degree distribution and sublinear degree growth. Section 6 discusses related approaches in the literature. Section 7 provides comparisons between the proposed non-exchangeable model and the two-parameter CRP on two datasets. Most proofs and some definitions can be found in the Appendix.

2. Background material.

2.1. *Completely random measures.* Completely random measures, introduced by Kingman [38], have found wide applicability as priors over functional spaces in Bayesian nonparametrics [64, 50], due to their flexibility and tractability; the reader can refer to [19, Chapter 10.1] or [50] for an extended coverage. A homogeneous CRM on $\mathbb{R}_+$ without fixed atoms nor deterministic component is almost surely discrete and takes the form

$$W = \sum_{j \geq 1} \omega_j \, \delta_{\vartheta_j}$$

where $\{(\omega_j, \vartheta_j)\}_{j \geq 1}$ are the points of a Poisson process on $(0, \infty) \times \mathbb{R}_+$ with mean measure $\nu(d\omega, d\theta)$. The measure decomposes as $\nu(d\omega, d\theta) = \rho(d\omega)\alpha(d\theta)$ where α is a non-atomic Borel measure on $\mathbb{R}_+$, called the base measure, such that $\alpha(A) < \infty$ for any bounded Borel set A, and ρ is a Lévy measure on $(0, \infty)$. We write $W \sim \mathrm{CRM}(\alpha, \rho)$. We will also assume in the following that the base measure $\alpha(d\theta)$ is absolutely continuous with respect to the Lebesgue measure with density $\tilde{\alpha}$ and

$$\int_{(0,\infty)\times\mathbb{R}_+} \rho(dw)\alpha(d\theta) = \infty. \tag{2.1}$$

Let

$$\psi(t) = \int_0^\infty \left\{1 - e^{-wt}\right\} \rho(dw) \tag{2.2}$$

be the Laplace exponent and define, for any integer $m \geq 1$ and any $u > 0$

$$\kappa(m, u) = \int_0^\infty \omega^m e^{-u\omega} \rho(d\omega).$$

A remarkable example of CRM is the generalised gamma process [8] (GG) with mean measure

$$\nu(d\omega, d\theta) = \frac{1}{\Gamma(1 - \sigma_0)} \omega^{-1-\sigma_0} e^{-\zeta_0 \omega} d\omega \, \alpha(d\theta)$$

with $\sigma_0 \in (0,1)$ and $\zeta_0 \geq 0$ or $\sigma_0 \in (-\infty, 0]$ and $\zeta_0 > 0$. We write $W \sim \mathrm{GG}(\alpha, \sigma_0, \zeta_0)$. The GG has been a popular model in Bayesian nonparametrics due to its flexibility and attractive conjugacy properties [33, 49, 48, 14]. It includes several important models as special cases: the gamma process for $\sigma_0 = 0$, $\zeta_0 > 0$; the inverse Gaussian process for $\sigma_0 = 1/2$, $\zeta_0 > 0$ and the stable process for $\sigma_0 \in (0,1)$ and $\zeta_0 = 0$.

2.2. *Exchangeable random partitions.* For an exchangeable partition $\Pi = (\Pi_n)_{n\geq 1}$ of $\mathbb{N}$ we have, for every $n \geq 1$

$$\mathbb{P}(\Pi_n = \{A_{n,1}, \ldots, A_{n,k_n}\},\, K_n = k_n) = p(|A_{n,1}|, \ldots, |A_{n,k_n}|)$$

where the sets $A_{n,j}$ are considered in order of appearance and p is a symmetric function of its arguments called *exchangeable partition probability function* (EPPF). Therefore, by definition, the ordering in which we observe the data is not taken into account and the only information that affects the distribution of the random partition is the size of the clusters. For an infinite sequence of random variables $(\theta_{(1)}, \theta_{(2)}, \ldots)$ taking values in $\mathbb{R}_+$, let $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$ be the random partition of $\mathbb{N}$ defined by the equivalence relation "i and j are in the same cluster" if and only if $\theta_{(i)} = \theta_{(j)}$ [59]. By Kingman's representation theorem [39], every exchangeable random partition has the same distribution as $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$, where the random variables $\theta_{(1)}, \theta_{(2)}, \ldots$ are conditionally independent and identically distributed from some random probability distribution $\mathbb{P}$.

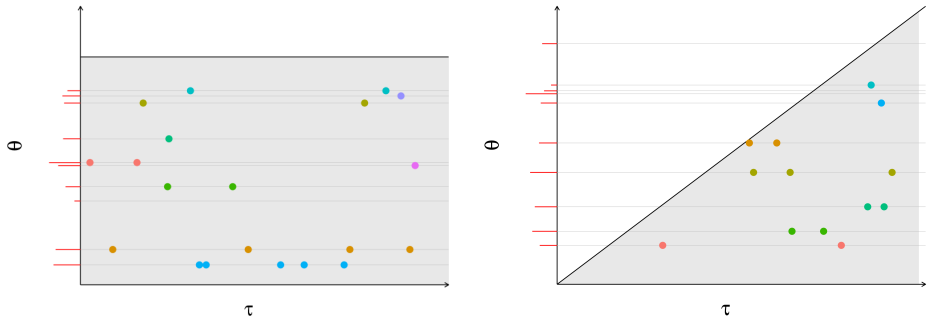
A popular model for this random probability distribution is a normalised completely random measure [64, 34, 35], defined as $P = W/W(\mathbb{R}_+)$, where $W \sim \mathrm{CRM}(\rho, \alpha)$ and $\alpha(\mathbb{R}_+) < \infty$. This condition, together with the condition (2.1), ensures that $0 < W(\mathbb{R}_+) < \infty$ almost surely, and the model is thus properly defined.

2.3. *Continuous-time embedding of exchangeable random partitions via Poissonisation.* Let $(\theta_{(1)}, \theta_{(2)}, \ldots)$ be an infinite sequence of random variables taking values in $\mathbb{R}_+$ and $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$ be the random partition of $\mathbb{N}$ defined by the equivalence relation "i and j are in the same cluster" if and only if $\theta_{(i)} = \theta_{(j)}$. Let $0 < \tau_{(1)} < \tau_{(2)} < \ldots$ be an infinite sequence of arrival times. Define the continuous-time partition-valued process $(\Pi(t))_{t\geq 0}$ as

$$\Pi(t) := \Pi_{N(t)} = \Pi(\theta_{(1)}, \ldots, \theta_{(N(t))}), \quad t \geq 0$$

where $N(t) = \sum_i \mathbb{1}_{\tau_{(i)} \leq t}$ with $\mathbb{1}_{\tau \leq t} = 1$ if $\tau \leq t$ and 0 otherwise. $(\Pi(t))_{t\geq 0}$ defines a continuous-time embedding of the partition Π. Note that we have

$$\Pi_n = \Pi(\tau_{(n)}) \quad n \geq 1.$$



(a) Poisson embedding of an exchangeable random partition

(b) Poisson embedding of the non-exchangeable random partition

FIG 1. *(a) The Chinese restaurant process obtained via a Poisson embedding. Points* (τ_i, θ_i) *are drawn from a Poisson point process on* $\mathbb{R}_+ \times [0,1]$ *with mean measure* $\mu(d\tau, d\theta) = d\tau\, W(d\theta)\mathbb{1}_{\theta\leq 1}$ *where* W *is a gamma random measure with base measure* $\alpha(d\theta) = \alpha_0 d\theta$ *and* $\zeta = 1$*. The red sticks on the* θ*-axis represent the jumps of the gamma random measure* W*. Points on the same horizontal line are in the same cluster. The random partition* $\Pi(\theta_{(1)}, \theta_{(2)}, \ldots)$ *of* $\mathbb{N}$ *induced by the sequence of points* $\theta_{(1)}, \theta_{(2)}, \ldots$ *ordered by their arrival times* $\tau_{(1)} < \tau_{(2)} < \ldots$ *is the CRP. (b) Non-exchangeable random partition model via a Poisson embedding. Points* (τ_i, θ_i) *are drawn from a Poisson point process on* $\mathbb{R}_+ \times \mathbb{R}_+$ *with mean measure* $\mu(d\tau, d\theta) = d\tau\, W(d\theta)\mathbb{1}_{\theta\leq\tau}$ *where* W *is a CRM. Note that the CRM is not normalised. The red sticks on the* θ*-axis represent the jumps of the CRM* W*. Points on the same horizontal line are in the same cluster.*

A remarkable feature of exchangeable random partitions is that they admit a continuous-time embedding via a Poisson process. Poissonisation is a classical technique used in combinatorial problems in order to derive analytical properties of exchangeable partitions and urn schemes [37, 30, 9].

We focus on the important case where the partition is obtained from a normalised CRM [64, 57, 45, 46, 48, 35], which includes as a special case the one-parameter CRP. Let $Q = \{(\tau_i, \theta_i)\}_{i\geq 1}$ be a Poisson (Cox) process on $\mathbb{R}_+ \times \mathbb{R}_+$ with random mean measure

$$\mu(d\tau, d\theta) = d\tau\, W(d\theta)\mathbb{1}_{\theta\leq 1}$$

where $W \sim \mathrm{CRM}(\alpha, \rho)$ with base measure $\alpha(d\theta) = \alpha_0\, d\theta$ with $\alpha_0 > 0$ and Lévy measure ρ satisfying $\int_0^\infty \rho(d\omega) = \infty$. The point process Q has support on $\mathbb{R}_+ \times [0,1]$ and we can write it as follows

$$\begin{aligned} W &\sim \mathrm{CRM}\,(\alpha, \rho) \\ Q \,|\, W &\sim \mathrm{Poisson}(d\tau\, W(d\theta)\mathbb{1}_{\theta\leq 1}) \end{aligned}$$

where Poisson(μ) denotes a Poisson point process with mean μ. This is illustrated in Figure 1(a). Note, importantly, that the CRM is not normalised in the above construction. The distribution of the first n points given W is

$$\mathbb{P}(\theta_{(i)} \in dx_i, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n \mid W) = \left[\prod_{i=1}^{n} W\{dx_i\}\right] e^{-t_n \overline{W}(1)} \mathbb{1}_{t_1<t_2<\dots<t_n} dt_1 \dots dt_n \tag{2.3}$$

where $\overline{W}(t) = \int_0^t W(d\theta) = \sum_{i\geq 1} \omega_i \mathbb{1}_{\vartheta_i \leq t}$. Let us denote by $m_{n,j}$ the number of points in the j-th cluster $A_{n,j}$ after having observed n points, and by $(\theta_j^*)_{j=1,\dots,K_n}$ the unique values in $(\theta_{(1)}, \dots, \theta_{(n)})$, ordered by arrival times. Using the results in [33, Proposition 3.1 page 18], we can obtain the expectation of (2.3) with respect to the CRM W

$$\begin{aligned}\mathbb{P}(\Pi_n = \{A_{n,1}, \dots, A_{n,k_n}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \dots, k_n, K_n = k_n,\\ \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n)\\ = e^{-\alpha_0 \psi(t_n)} \alpha_0^{k_n} \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, t_n)\, dx_j^*\right] \mathbb{1}_{t_1<\dots<t_n}, dt_1 \dots dt_n.\end{aligned}$$

Integrating over the arrival times $\tau_{(i)}$ and the cluster locations θ_j^* gives

$$\begin{aligned}\mathbb{P}(\Pi_n = \{A_{n,1}, \dots, A_{n,k_n}\}, K_n = k_n)\\ = \int_0^\infty e^{-\alpha_0 \psi(u)} \alpha_0^{k_n} \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, u)\right] \frac{u^{n-1}}{\Gamma(n)} du,\end{aligned} \tag{2.4}$$

and one recovers the EPPF of the exchangeable random partition associated to a normalised completely random measure [60, Corollary 6], [35, Proposition 3]. In the gamma process case, $\kappa(m, u) = \Gamma(m)/(1+u)^m$ and $\psi(u) = \log(1+u)$, and the right-hand side of (2.4) reduces to

$$\frac{\alpha_0^{k_n}}{\Gamma(n)} \left[\prod_{j=1}^{k_n} \Gamma(m_{n,j})\right] \int_0^\infty \frac{u^{n-1}}{(1+u)^{n+\alpha_0}} du = \frac{\alpha_0^{k_n} \Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \left[\prod_{j=1}^{k_n} \Gamma(m_{n,j})\right]$$

which is the EPPF of the Chinese Restaurant process.

3. Non-exchangeable random partitions. In this section, we build on the Poissonisation idea in order to derive a class of non-exchangeable random partitions. This class is shown to have the microclustering property

in the next section. The Cox Process $Q = \{(\tau_i, \theta_i)\}_{i\geq 1}$ that defines our non-exchangeable random partition model has the following random mean measure

$$\mu(d\tau, d\theta) = \mathbb{1}_{\theta\leq\tau} W(d\theta)d\tau \tag{3.1}$$

therefore the points will lie under the bisector as shown in Figure 1(b). The overall model is therefore defined as

$$\begin{aligned} W &\sim \mathrm{CRM}\,(\alpha, \rho) \\ Q \,|\, W &\sim \mathrm{Poisson}(\mathbb{1}_{\theta\leq\tau}\, d\tau\, W(d\theta)) \end{aligned}$$

The random partition $\Pi = (\Pi_n)_{n\geq 1}$ of $\mathbb{N}$ is obtained by considering the points $((\tau_{(i)}, \theta_{(i)}))_{i\geq 1}$ of the point process Q ordered by their arrival time, and let $\Pi_n = \Pi(\theta_{(1)}, \ldots, \theta_{(n)})$ be the partition induced by the first n points for any $n \geq 1$. The random partition model is completely specified by the base measure α and the Lévy measure ρ.

The crucial difference with the previous construction is the support of the point process. In the continuous time version of the CRP, every atom of W in $[0, 1]$ was allowed to be chosen at any time, hence the set of potential cluster labels was constant over time. Now, for every fixed $t > 0$, all the clusters whose θ are greater than t cannot be chosen before that time, therefore the set of potential cluster labels increases with t, if for instance the base measure α has unbounded support on $\mathbb{R}_+$. This property intuitively leads to both the non-exchangeability of the random partition induced on $\mathbb{N}$, but also to the microclustering property.

Samples from the process are represented in Figure 2 when $W \sim \mathrm{GG}(\alpha, \sigma, 1)$ with base measure $\alpha(d\theta) = \xi\theta^{\xi-1}d\theta$, for different values of ξ and σ.

PROPOSITION 1. *Let W be a homogeneous* $\mathrm{CRM}(\alpha, \rho)$ *and a point process $Q = \{(\tau_i, \theta_i)\}_{i\geq 1}$ on $\mathbb{R}_+^2$ with mean measure $\mu(d\tau\, d\theta) = \mathbb{1}_{\theta\leq\tau} d\tau\, W(d\theta)$. Let $((\tau_{(i)}, \theta_{(i)}))_{i\geq 1}$ be the sequence of points ordered in time, that is such that $\tau_{(1)} < \tau_{(2)} < \ldots$. For any $n \geq 1$, let Π_n be the random partition of $\{1, \ldots, n\}$ defined by the equivalence relation "i and j are in the same cluster" if and*

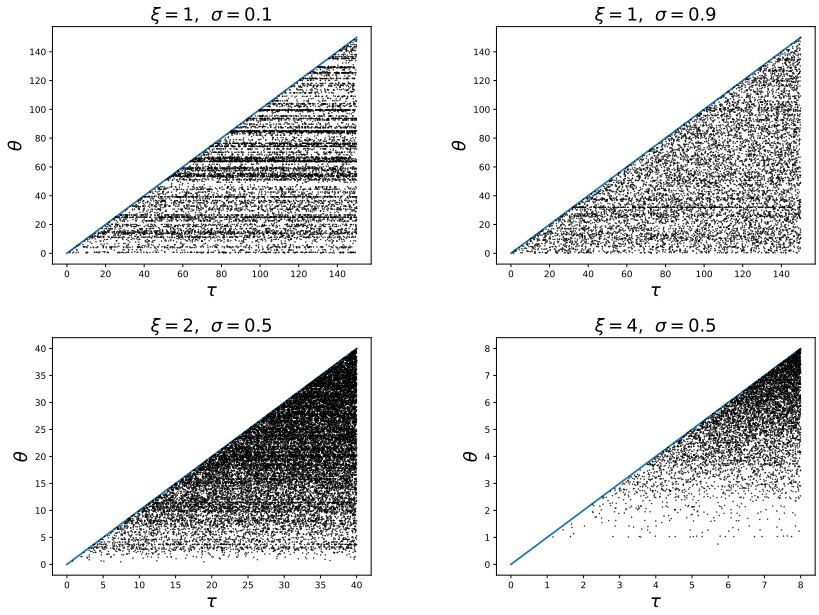


FIG 2. *Samples from the Cox Process Q where $W \sim \mathrm{GG}(\alpha, \sigma, 1)$ with base measure $\alpha(d\theta) = \xi\, \theta^{\xi-1}\, d\theta$, for different values of σ and ξ.*

only if $\theta_{(i)} = \theta_{(j)}$. For any $n \geq 1$, we have

$$\begin{aligned} &\mathbb{P}(\Pi_n = \{A_{n,1}, \ldots, A_{n,k_n}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \ldots, k_n, K_n = k_n, \\ &\quad \tau_{(i)} \in dt_i \text{ for } i = 1, \ldots, n) \\ &\qquad = \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, t_n - x_j^*)\alpha(dx_j^*)\right] \\ &\qquad \times e^{-\int_0^{t_n} \psi(t_n-\theta)\alpha(d\theta)} \left[\prod_{i=1}^{n} \mathbb{1}_{x_{c_i}^* \leq t_i}\right] \mathbb{1}_{t_1<t_2<\ldots<t_n}\, dt_1 \ldots dt_n \end{aligned} \tag{3.2}$$

where θ_j^*, $j = 1, \ldots, k_n$, are the unique values of $(\theta_{(1)}, \ldots, \theta_{(n)})$, $m_{n,j} = |A_{n,j}|$ their multiplicities and $c_i \in \{1, \ldots, k_n\}$ is such that $i \in A_{n,c_i}$.

PROOF. The derivation is similar to the derivation for the exchangeable case described in the previous section. Given W, the set of points $(\tau_i)_{i\geq 1}$ is an inhomogeneous Poisson point process on $\mathbb{R}_+$ with rate $\overline{W}(t)$, hence

$$\mathbb{P}(\tau_{(i)} \in dt_i \text{ for } i = 1, \ldots, n \mid W) = e^{-\int_0^{t_n} \overline{W}(t)dt} \left[\prod_{i=1}^{n} \overline{W}(t_i)\right] \mathbb{1}_{t_1<\ldots<t_n} dt_1 \ldots dt_n.$$

Given the n time variables, the $\theta_{(i)}$'s are distributed as follows

$$\mathbb{P}(\theta_{(i)} \in dx_i \text{ for } i = 1, \dots, n \mid \tau_{(1:n)}, W) = \prod_{i=1}^{n} \frac{W(dx_i)}{\overline{W}(\tau_{(i)})} \mathbb{1}_{x_i < \tau_{(i)}}.$$

It follows that

$$\begin{aligned} &\mathbb{P}(\tau_{(i)} \in dt_i, \theta_{(i)} \in dx_i \text{ for } i = 1, \dots, n \mid W) \\ &\quad = \left[\prod_{i=1}^{n} W(dx_i) \mathbb{1}_{x_i \leq t_i} \right] e^{-\int_0^{t_n} \overline{W}(t)dt} \mathbb{1}_{t_1 < \dots < t_n} \, dt_1 \dots dt_n \end{aligned} \tag{3.3}$$

where $\int_0^{\tau_{(n)}} \overline{W}(t)dt = \sum_j \omega_j (\tau_{(n)} - \vartheta_j)_+ = W(g_{\tau_{(n)}})$ with $g_t(x) = (t - x)_+ = \max(0, t - x)$. Using [33, Proposition 3.1], we can integrate over W to obtain the final result. □

Integrating Equation (3.2) over the cluster allocations $(\theta_j^*)_{j=1,\dots,k_n}$ and the arrival times $\tau_{(1:n)}$, we would obtain the distribution of the random partition Π_n. To the best of our knowledge, there is however no analytical expression for this distribution. We can nonetheless simulate random partitions by using the cluster allocations and arrival times as latent variables. In particular, for the generalised gamma process, we have the following result.

PROPOSITION 2. *Let $W \sim \mathrm{GG}(\alpha, \sigma_0, \zeta_0)$, and $Q = \{(\tau_i, \theta_i)\}_{i \geq 1}$ be the points of a Cox process with mean measure $\mu(d\tau, d\theta) = \mathbb{1}_{\theta \leq \tau} W(d\theta)d\tau$. Then the predictive distribution of $\tau_{(n)}$ is given by*

$$\begin{aligned} \mathbb{P}(\tau_{(n)} \in dt_n \,|\, (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}) \propto &\left[\prod_{j=1}^{k_{n-1}} \frac{1}{(t_n - \theta_j^* + \zeta_0)^{m_{n-1,j} - \sigma_0}} \right] e^{-\int_0^{t_n} \psi(t_n - \theta)\alpha(d\theta)} \\ &\times \left(\sum_{j=1}^{k_{n-1}} \frac{m_{n-1,j} - \sigma_0}{t_n - \theta_j^* + \zeta_0} + \int_0^{t_n} \frac{\alpha(\theta)}{(t_n - \theta + \zeta_0)^{1-\sigma_0}} \, d\theta \right) \mathbb{1}_{t_n > \tau_{(n-1)}} dt_n \end{aligned}$$

where $\psi(t) = \log(1 + t/\zeta_0)$ for $\sigma_0 = 0$, while $\psi(t) = \left((t + \zeta_0)^{\sigma_0} - \zeta_0^{\sigma_0}\right)/\sigma_0$ for $\sigma_0 \in (0, 1)$. The conditional distribution for $\theta_{(n)}$ is a convex combination of a discrete distribution and a diffuse one,

$$\mathbb{P}(\theta_{(n)} \in dx_n \,|\, (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}, \tau_{(n)}) \propto H_{\tau_{(n)}}(dx_n) + \sum_{i=1}^{k_{n-1}} \frac{m_{n-1,i} - \sigma_0}{\tau_{(n)} - \theta_i^* + \zeta_0} \delta_{\theta_i^*}(dx_n)$$

where H_t is a diffuse distribution defined as

$$H_t(A) = \int_A \frac{\mathbb{1}_{\theta \le t}}{(t-\theta+\zeta_0)^{1-\sigma_0}}\,\alpha(d\theta) \tag{3.4}$$

for every Borel set $A \subset \mathbb{R}_+$.

For example, if W is a gamma process ($\sigma_0 = 0$) and $\alpha(d\theta) = d\theta$, we obtain

$$\begin{aligned}\mathbb{P}(\tau_{(n)} \in dt_n \mid (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}) \\ \propto \left[\prod_{j=1}^{k_{n-1}} \frac{1}{(t_n - \theta_j^* + \zeta_0)^{m_{n-1,j}}}\right] e^{-t_n}\left(1+\frac{t_n}{\xi_0}\right)^{-t_n-\xi_0} \\ \times \left(\sum_{j=1}^{k_{n-1}} \frac{m_{n-1,j}}{t_n-\theta_j^*+\zeta_0} + \log(1+t_n/\zeta_0)\right)\mathbb{1}_{t_n>\tau_{(n-1)}}dt_n,\end{aligned}$$

and

$$\begin{cases} \mathbb{P}(\theta_{(n)} = \theta_j^* \mid (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}, \tau_{(n)}) = C_n \frac{m_{n-1,j}}{\tau_{(n)}-\theta_j^*+\zeta_0} & \text{for } j = 1,\dots,k_{n-1} \\ \mathbb{P}(\theta_{(n)} \text{ is new} \mid (\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n-1}, \tau_{(n)}) = C_n \log(1+\tau_{(n)}/\zeta_0) \end{cases}$$

where C_n is the appropriate normalising constant.

4. Properties and inference.

4.1. *Asymptotic properties.* In this section, denote $X_t \sim Y_t$, $X_t = o(Y_t)$ and $X_t = O(Y_t)$ respectively for $X_t/Y_t \to 1$, $X_t/Y_t \to 0$ and $\limsup_t X_t/Y_t < \infty$. The notation $X_t \asymp Y_t$ means both $X_t = O(Y_t)$ and $Y_t = O(X_t)$ hold. When X_t and/or Y_t are random variables the asymptotic relation is meant to hold almost surely.

The properties we are most interested in are the asymptotic behaviour of the cluster sizes $m_{n,j}$, of the number K_n of clusters and the number $K_{n,r}$ of clusters of size r in the random partition. We show in this section that our non-exchangeable model allows for a sublinear growth of the clusters' sizes while retaining desirable properties for the other quantities. Let us list the assumptions on the CRM W to derive the asymptotic results.

(A1) W has finite first two moments, that is

$$\kappa(1,0) = \int_0^\infty \omega\rho(d\omega) < \infty \text{ and } \kappa(2,0) = \int_0^\infty \omega^2\rho(d\omega) < \infty.$$

(A2) The Lévy tail intensity $\bar{\rho}(x) = \int_x^\infty \rho(d\omega)$ is a *regularly varying* function at 0, that is
$$\overline{\rho}(x) \sim \ell\,(1/x)\, x^{-\sigma}$$
as $x \to 0^+$, where ℓ is a slowly varying function at infinity and $\sigma \in [0,1]$.

(A3) The improper cumulative distribution $\overline{\alpha}(t) = \int_0^t \alpha(dx)$ of the base measure α is a regularly varying function at infinity, that is
$$\overline{\alpha}(t) \sim L(t)\, t^{\xi}$$
as $t \to \infty$, where $\xi > 0$ and L is a slowly varying function. Assume additionally that the base measure α is dominated by the Lebesgue measure, and admits a continuous density denoted $\tilde{\alpha}(\theta)$.

The moment assumption (A1) excludes the stable process that has infinite first moment. (A2) controls, through the parameter σ, the power-law behaviour of the proportion of clusters of a given size, while condition (A3) is used to prove the microclustering property and control the sublinear rate of the clusters' size. Assumptions (A1-A2) are satisfied for the GG with parameters $\sigma_0 \in (-\infty, 1)$ and $\zeta_0 > 0$. In this case, we have $\sigma = \max(\sigma_0, 0)$ and $\ell(t) \propto \log t$ for $\sigma = 0$ and $\ell(t)$ is constant otherwise.

Recall that
$$N(t) = \sum_{i\geq 1} \mathbb{1}_{\tau_i \leq t}$$
denotes the number of points of Q such that $\tau_i \leq t$. For each atom ϑ_j, $j \geq 1$, of the CRM W, let
$$X_j(t) = \sum_{i\geq 1} \mathbb{1}_{\tau_i \leq t} \mathbb{1}_{\theta_i = \vartheta_j}.$$
For $j \geq 1$, let
$$M_j(t) = \sum_{i\geq 1} \mathbb{1}_{\tau_i \leq t} \mathbb{1}_{\theta_i = \theta_j^*}$$
the size of cluster j, ordered by appearance, at time t. Note that $N(\tau_{(n)}) = n$ and $M_j(\tau_{(n)}) = m_{n,j}$.

PROPOSITION 3. *Let $W = \sum_{j\geq 1} \omega_j \delta_{\vartheta_j}$ be a CRM with mean measure $\alpha(d\theta)\rho(d\omega)$ satisfying Assumptions (A1-A3). Let $\{(\tau_i, \theta_i)\}_{i\geq 1}$ be a Poisson point process with mean measure $\mu(d\tau\, d\theta) = \mathbb{1}_{\theta \leq \tau} d\tau\, W(d\theta)$. We have, almost surely as t tends to infinity,*
$$N(t) \sim \frac{\kappa(1,0)}{\xi + 1} t^{\xi+1} L(t)$$

and, for $j \geq 1$

$$X_j(t) \sim W(\{\vartheta_j\})t$$
$$M_j(t) \sim W(\{\theta_j^*\})t.$$

Proposition 3 implies the microclustering property for the random partition Π_n: almost surely, $M_j(t)/N(t) \to 0$ as $t \to \infty$, hence $m_{n,j}/n \to 0$ as $n \to \infty$. In the following theorem, which follows from properties of inverse of regularly varying functions [5, Proposition 1.5.15] or [30, Lemma 22], we obtain exact rates of growth for the cluster sizes.

THEOREM 4 (Microclustering property). *We have*

$$t \sim \left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} L^*_{\xi+1}(N(t))\, N(t)^{1/(\xi+1)} \tag{4.1}$$

almost surely as $t \to \infty$, *where* $L^*_{\xi+1}$ *is a slowly varying function defined in equation* (B.1) *in the Appendix. It follows that the cluster sizes* $m_{n,j} = M_j(\tau_{(n)})$ *verify, for any* $j \geq 1$,

$$m_{n,j} \sim W(\{\theta_j^*\}) \left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} L^*_{\xi+1}(n)\, n^{1/(\xi+1)} \tag{4.2}$$

almost surely as n *tends to infinity. The random weights* $\omega_j^* = W(\{\theta_j^*\})$ *admit the following construction. Denoting by* τ_j^* *the arrival time of the* j*-th cluster, we have that*

$$\tau_j^* = \Phi^{-1}(\gamma_j)$$
$$\mathbb{P}(\theta_j^* \in dx \mid \tau_j^*) = \frac{H_{\tau_j^*}(dx)}{H_{\tau_j^*}(\mathbb{R}_+)}$$
$$\mathbb{P}(\omega_j^* \in dw_j \mid \theta_j^*, \tau_j^*) = \frac{w_j e^{-(\tau_j^* - \theta_j^*)w_j} \rho(dw_j)}{\kappa(1, \tau_j^* - \theta_j^*)}$$

where $H_\tau(d\theta) = \kappa(1, \tau - \theta)\mathbb{1}_{\theta<\tau}\alpha(d\theta)$, $\gamma_j = \sum_{k=1}^j e_k$ *where* $e_1, e_2, \ldots$ *are iid* Exp(1) *and* $\Phi(t) = \int_0^\infty \int_0^\infty \left(1 - e^{-\omega(t-\theta)_+}\right) \rho(d\omega)\alpha(d\theta) = \int_0^\infty \psi(t-\theta)\alpha(d\theta)$. *In the specific case of the GG with parameters* α, σ_0 *and* ζ_0, H_t *takes the form* (3.4) *and*

$$\omega_j^* \,|\, \theta_j^*, \tau_j^* \sim \text{Gamma}(1 - \sigma_0, \tau_j^* - \theta_j^* + \zeta_0)$$

Note that the growth rate of the cluster sizes only depends on the parameters ξ and L of the base measure α, and not on the properties of the Lévy measure ρ. For example, taking $\overline{\alpha}(t) = \gamma t^{\xi}$, with $\xi, \gamma > 0$, we have $L(t) = \gamma$ and $m_{n,j} \asymp n^{1/(\xi+1)}$ and the cluster sizes grow at a rate of n^a where $0 < a < 1$.

We now provide results on the asymptotic rates of the number of clusters and number of clusters of a given size, showing that we can have the same range of behaviour as for exchangeable random partitions. Let

$$K(t) = \sum_{j \geq 1} \mathbb{1}_{X_j(t)>0}$$

be the number of different clusters in $\Pi(t)$ at time t and

$$K_r(t) = \sum_{j \geq 1} \mathbb{1}_{X_j(t)=r}$$

the number of clusters of size r at time t.

PROPOSITION 5. *Let $W = \sum_{j\geq 1} \omega_j \delta_{\vartheta_j}$ be a CRM with mean measure $\alpha(d\theta)\rho(d\omega)$ satisfying Assumptions (A1-A3). Define*

$$\begin{cases} \ell_\sigma(t) = \Gamma(1-\sigma)\ell(t) & \text{if } \sigma \in [0,1) \\ \ell_1(t) = \int_t^\infty y^{-1}\ell(y)dy & \text{if } \sigma = 1. \end{cases}$$

Let $\{(\tau_i, \theta_i)\}_{i\geq 1}$ be a Poisson point process with mean measure $\mu(d\tau\, d\theta) = \mathbb{1}_{\theta\leq\tau} d\tau\, W(d\theta)$. We have, almost surely at t tends to infinity,

$$K(t) \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}.$$

For $r \geq 1$, if $\sigma = 0$ then $K_r(t) = o(K(t))$, if $\sigma \in (0,1)$,

$$K_r(t) \sim \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} K(t)\,.$$

If $\sigma = 1$, $K_1(t) \sim K(t)$ and $K_r(t) = o(K(t))$ for all $r \geq 2$.

By noting that $K_n = K(\tau_{(n)})$ and $K_{n,r} = K_r(\tau_{(n)})$, we can combine the results of Proposition 5 and Equation (4.1) to obtain asymptotic expressions for the number K_n of clusters and the number $K_{n,j}$ of clusters of size j in Π_n.

COROLLARY 6. *We have, almost surely as n tends to infinity,*

$$K_n \sim \widetilde{\ell}(n)\, n^{(\sigma+\xi)/(\xi+1)}$$

where $\widetilde{\ell}$ is a slowly varying function defined in equation (B.2) in the Appendix.

For $r \geq 1$, if $\sigma = 0$ then $K_{n,r} = o(K_n)$; if $\sigma \in (0,1)$,

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}\,. \tag{4.3}$$

This corresponds to a power-law behaviour for the proportion of clusters of size r, as

$$\frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} \asymp \frac{1}{r^{1+\sigma}}$$

for large r. If $\sigma = 1$, $K_{n,1} \sim K_n$ and $K_{n,r} = o(K_n)$ for all $r \geq 2$. In this case, the proportion of clusters of size 1 tends to one almost surely.

EXAMPLE 7. *If $W \sim \mathrm{GG}(\alpha, \sigma, 1)$ with $\sigma \in (0,1)$ and base measure $\alpha(d\theta) = \gamma\xi\theta^{\xi-1}d\theta$ with $\xi, \gamma > 0$ we have $\ell(t) = \frac{1}{\sigma\,\Gamma(1-\sigma)}$ and $L(t) = \gamma$, therefore*

$$K_n \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\sigma\Gamma(\sigma+\xi+1)}(\xi+1)^{\frac{\sigma+\xi}{1+\xi}}\,\gamma^{1-\frac{\sigma+\xi}{1+\xi}}\,n^{\frac{\sigma+\xi}{1+\xi}}$$

and for all $r \geq 1$

$$\frac{K_{n,r}}{K_n} \to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}$$

almost surely as n tends to infinity. This power-law behaviour is illustrated on Figure 3.

It is worth noting that although the asymptotic behaviour of the number of clusters and the number of clusters of a given size depend also on the base measure α, the power-law exponent in the proportion of clusters of a given size is solely tuned by the Lévy measure ρ through the parameter σ. The asymptotic proportion of clusters of a given size, given by Equation (4.3), is indeed the same as that obtained with a two-parameter Chinese restaurant process [61] with strictly positive discounting parameter, with a normalised generalised gamma process [48] with parameter $\sigma_0 > 0$ and more generally with a Poisson-Kingman partitions exhibiting α-diversity [60].

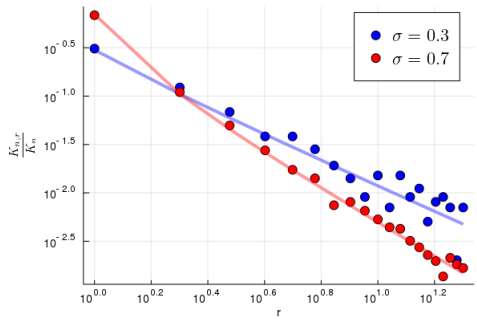


FIG 3. *Log-log plot of the proportions of clusters of given size r for the GG with $\alpha(d\theta) = d\theta$, $\zeta_0 = 0.1$, $\sigma = 0.3, 0.7$ and sample size 100k. Dots represent samples from the model (n=100k) and lines represent the asymptotic proportions as $n \to \infty$, as given by Equation (4.3).*

REMARK 8. *The number of points sampled by the point process up to time t is distributed as the number of points sampled up to time t by a Cox process with mean measure $\mathbb{1}_{\theta \leq L(\tau)\tau^{\xi}} W(d\theta)d\tau$, where the CRM W has base measure $\alpha(d\theta) = d\theta$. The same asymptotic result would hold and the proof are analogous and follow by a change of measure.*

REMARK 9. *The counts K_n and $K_{n,r}$ are asymptotically equivalent to deterministic quantities, contrary to what happens in normalised completely random measures, and more generally in Poisson-Kingman random partitions, where the asymptotics include a random component (see section 6.1 about α-diversity in [60]). The reason lies in the fact that our construction involves a CRM with a growing support as the number n (or t) grows, whose mass $W([0,t)) \to \infty$ almost surely as $t \to \infty$. Informally, a law of large numbers is therefore at play, which implies $W([0,t))/\mathbb{E}(W([0,t)) \to 1$ almost surely as $t \to \infty$, resulting in the appearance of the term $\kappa(1,0) = \mathbb{E}[W([0,t))]/\overline{\alpha}(t)$ in the depoissonisation step (see Equations (4.1) and (B.2)). We refer to the Appendix for a formal proof. Our results hold in the case where $\overline{\alpha}$ is regularly varying with index $\xi > 0$. It would be interesting to investigate the slowly varying case ($\xi = 0$) as a random component may arise in this situation.*

4.2. *Inference.*

4.2.1. *Posterior characterization.*

Assuming we observe the first n time-ordered points $(\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}$ from the Cox process Q, we want to charac-

terise the conditional distribution of the CRM W given the time-ordered observations. The following posterior characterization follows from [33, Proposition 3.1 page 18].

PROPOSITION 10. *Given the first n time-ordered observations $(\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}$ from the Cox process Q, with unique cluster labels $\theta_1^*, \dots, \theta_{k_n}^*$, the conditional distribution of the CRM W is given by*

$$W' + \sum_{j=1}^{k_n} \omega_j^* \delta_{\theta_j^*}$$

where the random positive weights $(\omega_1^*, \dots, \omega_{k_n}^*)$ are independent of the random measure W'. W' is an inhomogeneous CRM with mean measure $\nu'(d\omega, d\theta) = e^{-\omega(\tau_{(n)}-\theta)_+}\rho(d\omega)\alpha(d\theta)$. The masses of the fixed atoms are conditionally independent given the data with density

$$\mathbb{P}(\omega_j^* \in dw_j \mid (\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}) = \frac{\rho(dw_j)\, w_j^{m_{n,j}} e^{-w_j(\tau_{(n)}-\theta_j^*)}}{\kappa(m_{n,j}, \tau_{(n)} - \theta_j^*)}.$$

In particular, when W is a generalised gamma process the masses are conditionally gamma distributed

$$\omega_j^* \mid (\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n} \sim \text{Gamma}(m_{n,j} - \sigma_0, \zeta_0 + \tau_{(n)} - \theta_j^*).$$

4.2.2. *Parameter estimation and prediction.* We consider the CRM with base measure $\alpha(d\theta) = \xi\, \theta^{\xi-1} d\theta$ and generalised gamma Lévy measure with parameters σ_0 and ζ_0. The set of parameters is therefore $\eta = (\xi, \sigma_0, \zeta_0)$. Having observed a partition Π_n, we aim at estimating the parameters η and predict Π_{n+m} for $m \geq 1$. However, unlike in the CRP, the marginal likelihood $\mathbb{P}(\Pi_n|\eta)$ is intractable. We use a sequential Monte Carlo algorithm [23, 22] with target distribution $\mathbb{P}(d\theta_{(1:n)}, d\tau_{(1:n)}|\Pi_n, \eta)$ in order to get unbiased estimators of the marginal likelihoods $\mathbb{P}(\Pi_n|\eta)$ for a grid of values of η, and compute the maximum likelihood estimate $\widehat{\eta}$ (see Appendix C for details). Since the model is not Markovian, the time complexity of the SMC algorithm is of order $\mathcal{O}(nK_n)$. The proposal distribution for the arrival times $\tau_{(n)}$ is a truncated normal on $[\tau_{(n-1)}, \infty)$, while the proposal for the cluster location of a new cluster is uniform on $[0, \tau_{(n)}]$. We also use a sequential Monte Carlo algorithm in order to sample from the predictive $\mathbb{P}(\Pi_{n+m}|\Pi_n, \widehat{\eta})$ using Proposition 10.

We also consider experiments where the partition is not observed. In this setting, we consider a classical Bayesian hierarchical construction, see

e.g. [43]. We assume that the observations $(y_1, \ldots, y_n)$ are conditionally independent, with

$$y_i|\Pi_n, U_1, \ldots, U_{k_n} \sim F_{U_{c_i}} \tag{4.4}$$

where F_U is a known probability distribution parameterized by U, with pdf f_U, and the $(U_1, \ldots, U_{k_n})$ are the cluster parameters, iid from some distribution G_0 on Θ. This leads to the following factorization of the likelihood

$$p(y_1, \ldots, y_n|\Pi_n) = \prod_{j=1}^{k_n} g(y_{A_{n,j}}) \tag{4.5}$$

where $y_A = \{y_i | i \in A\}$ and

$$g(y_A) = \left[\int_\Theta \prod_{i \in A} f_U(y_i)\right] dG_0(U) \tag{4.6}$$

We will assume that G_0 is a conjugate prior for f_U, so that g, and therefore the likelihood (4.5) can be evaluated analytically. We use a sequential Monte Carlo algorithm with target distribution $\mathbb{P}(d\theta_{(1:n)}, d\tau_{(1:n)}, \Pi_n|\eta, y_{1:n})$ in order to get unbiased estimators of the marginal likelihoods $\mathbb{P}(y_{1:n}|\eta)$ for a grid of values of η, and compute the maximum likelihood estimate $\widehat{\eta}$. More details about the algorithm are contained in the Appendix C.

5. Random partitions and random multigraphs. The non-exchangeable random partition model proposed can be used to derive models for random multigraphs, see [7]. Recall that $\Pi_n = (A_{n,1}, \ldots, A_{n,K_n})$, where the blocks are sorted in order of appearance. For each $i = 1, 2, \ldots$, let c_i be the index of the cluster to which item i belongs, that is $i \in A_{n,c_i}$ for all $n \geq i$. An undirected multigraph $G = \Phi(\Pi)$, possibly with self-loops and with a countably infinite number of edges, is derived from the random partition Π by

$$G = ((c_1, c_2), (c_3, c_4), \ldots)$$

where each pair (c_{2n-1}, c_{2n}) represents an undirected edge between the vertex c_{2n-1} and the vertex c_{2n}. The set of vertices is either $\{1, \ldots, K\}$ if the partition has a finite number of blocks, or the set $\mathbb{N}$. Let G_n be the restriction of G to the first n edges that is, to the first $2n$ items of Π. Then K_{2n}, the number of clusters in Π_{2n}, is also the number of vertices of G_n, $m_{2n,j}$ is the degree of vertex j, $j = 1, \ldots, K_{2n}$ and $K_{2n,j}/K_{2n}$ is the proportion of vertices of degree j.

The multigraphs G obtained by transformation of an exchangeable random partition form a subclass of the edge-exchangeable graphs [18, 10]. This subclass is called rank one edge-exchangeable graphs by Janson [36]. Of particular interest is the so-called Hollywood model [18], obtained from a two-parameter CRP random partition. In this case, inherited from the properties of the associated random partition [61], one can obtain sparse multigraphs with power-law degree distribution. A consequence of the exchangeability assumption is the fact that the degree sequence grows linearly with the number of edges: for any vertex j, its degree $m_{2n,j} \asymp n$ almost surely as the number of edges n tends to infinity. As shown in the following corollary of the results of Section 4, our construction allows to obtain sparse multigraphs with power-law degree distribution and sublinear growth rate for degree sequences.

COROLLARY 11. *Let Π be a non-exchangeable partition with parameters α and ρ verifying assumptions (A1-A3). Let $G = \Phi(\Pi)$ the associated random multigraph. For a subgraph G_n corresponding to the first n edges, let K_{2n} be the number of vertices, $m_{2n,j}$ the degree of vertex j and $K_{2n,r}$ the number of vertices of degree $r \geq 1$. Then, almost surely as the number of edges n tends to infinity*

$$\begin{aligned} m_{2n,j} &\asymp L^*_{\xi+1}(n)\, n^{1/(1+\xi)}, \quad j \geq 1 \\ K_{2n} &\sim \widetilde{\ell}(n)(2n)^{(\sigma+\xi)/(1+\sigma)} \\ \frac{K_{2n,r}}{K_{2n}} &\to \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)}, \quad r \geq 1 \end{aligned}$$

where the slowly varying functions $L^*_{\xi+1}$ and $\widetilde{\ell}$ are defined in Equations (B.1) and (B.2).

6. Discussion. To obtain random partitions with the microclustering property, one option is to give up the exchangeability assumption, as we did in this paper. An alternative approach is to drop the Kolmogorov consistency assumption discussed in the introduction. Miller et al. [55] and Betancourt et al. [68], who derived random partition models with the microclustering property, take this option, and consider a collection $(\Pi_n)_{n\geq 1}$ of finitely exchangeable random partitions of $[n]$ that do not define a (Kolmogorov-consistent) random partition of $\mathbb{N}$. Another related contribution is the work of [69] where the authors also define a collection $(\Pi_n)_{n\geq 1}$ of finitely exchangeable random partitions of $[n]$ that do not satisfy Kolmogorov consistency property; the authors emphasise that it is indeed a desirable feature for modeling frequencies of frequencies, which motivates their work. Their model is also

based on some Poissonisation idea. Other random partition models that do not satisfy Kolmogorov consistency are also described in [4].
In [24] the authors present a multitype Yule process that can be seen as a non-exchangeable generalization of the CRP, where a new customer sits at a new table with constant probability $r \in (0,1)$ and otherwise randomly picks one of the previous customers and sits at his table. This model has a very simple urn construction, and has the microclustering property, as $m_{n,j} \asymp n^{1-r}$. The asymptotic properties of K_n and $K_{n,j}$ are however very different from those obtained with our model. The number of clusters K_n increases linearly with n, and the model exhibits a power-law behaviour for the proportion of clusters of a given size within an intermediate range. That is $\sum_{r \geq S} K_{nr} \asymp nS^{-\frac{1}{1-r}}$ valid for $1 \ll S \ll n^{1-r}$. In contrast, our model induces a sublinear growth of the number of clusters K_n, as for exchangeable random partitions, and the same asymptotic power-law behaviour as the two-parameter Chinese restaurant process for the proportion of clusters of a given size.

There has been a lot of interest over the past years in the development of non-exchangeable partitions based on dependent Dirichlet processes and more generally dependent random measures [53, 32, 13, 29, 6, 16, 51, 11, 12]. The focus of these works is rather different though, as they do not aim to capture/characterise the microclustering property. The model presented here builds on a Poisson construction on an augmented space, and is therefore somewhat reminiscent of the work of [63, 52, 17].

The authors of [7] considered a general class of exchangeable and non-exchangeable random partitions of $\mathbb{N}$, motivated by preferential attachments models for random multigraphs. For certain values of the parameters, it can generate partitions with the microclustering property, but with a somewhat different asymptotic behaviour for the number of clusters. The microclustering property is obtained whenever the number of clusters grows linearly with the dataset [7, Theorem 7]. In our approach, the number of clusters always grows sublinearly, and the rate can be controlled by the properties of the Lévy measure ρ.

7. Experiments. In what follows we compare our non-exchangeable model to the two-parameter Chinese restaurant process [61]. For the non-exchangeable model, we consider a GG with mean measure $\rho(dw)\alpha(d\theta) = 1/\Gamma(1-\sigma)w^{-1-\sigma}e^{-\zeta w}dw\xi\theta^{\xi-1}d\theta$ where the parameters $\xi \in \{1,2,3\}$, $\sigma \in [0,1)$ and $\zeta > 0$ are unknown. For the two-parameter CRP, the two parameters $\sigma_2 \in [0,1)$ and $\kappa_2 > 0$ are considered unknown. We present two sets of experiments. In the first set of experiments, the partition is observed for

a training set, and we aim at predicting the size of the clusters on a test set. In the second set of experiments, we consider an application to (semi-supervised) clustering, where the partition is not completely observed and has to be inferred from data. In each set of experiments the data are partitioned into a training set of size n_{train} and a test set of size n_{test} where $n_{\text{train}} + n_{\text{test}} = n$.

7.1. *Observed partition.* In order to best assess the performance of the two random partition models we first assume that the partition is observed. The parameters of each model are estimated on the training data using maximum likelihood with a grid of values for the parameters: 25 equidistant points in $[0, 1)$ for σ, $\xi \in \{1, 2, 3\}$, and a grid obtained by dichotomic search on the interval $[0, 100]$ for ζ, and similarly for the two-parameter CRP. The EPPF $\mathbb{P}(\Pi_{n_{\text{train}}}|\sigma_2, \kappa_2)$ of the two-parameter CRP has an analytic form and is calculated directly. For our method, we approximate the likelihood $\mathbb{P}(\Pi_{n_{\text{train}}}|\xi, \sigma, \zeta)$ using sequential Monte Carlo methods with 10000 particles, as described in Section 4.2.

For each cluster $j = 1, \dots, K_{n_{\text{train}}}$ in the training set, we then aim at predicting its size $m_{k,j}$ for $k = n_{\text{train}} + 1, \dots, n$. Let $m_{k,j}^{(\text{true})}$ be the true size of cluster j in the partition of size k and consider the L2 error

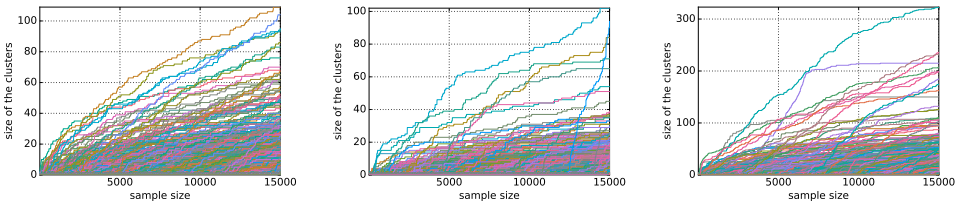
$$E = \frac{1}{K_{n_{\text{train}}}} \sum_{j=1}^{K_{n_{\text{train}}}} \frac{1}{n_{\text{test}}} \sum_{k=n_{\text{train}}+1}^{n} \left(m_{k,j} - m_{k,j}^{(\text{true})}\right)^2 \geq 0.$$

We are interested in the distribution of the predictive error

$$\mathbb{P}\left(E \in dE \mid \Pi_{n_{\text{train}}}, \widehat{\eta}\right) \tag{7.1}$$

where $\widehat{\eta}$ are the fitted parameters, under the two-parameter CRP or our model. Additionally, we want to check that the model can still capture the distribution of the cluster sizes adequately. To this aim, we also report 95% predictive credible intervals for the proportion of clusters of a given size in the test set, and compare this to the empirical distribution.

Synthetic data. In order to validate the inference procedure, we first run experiments on a simulated dataset, where the data are simulated from our model with parameters set to $(\xi, \sigma, \zeta) = (1, 0.41, 10)$. In this model, the cluster size grows at a rate of $\sqrt{n}$, as can be seen from Figure 4(a) that shows the growth of the cluster sizes with respect to the sample size n. Additionally, the proportion of clusters of a given size has an asymptotic power-law distribution, see the top row of Figure 7. As shown in Figure 5,



(a) Synthetic (b) Amazon (c) Math Overflow

FIG 4. *Evolution of the clusters' sizes $m_{j,n}$ with respect to the sample size n for the (a) synthetic, (b) Amazon and (c) Math Overflow datasets.*

the SMC estimate of the log-likelihood is rather accurate, and we recover the true parameters. The mean and quantiles of the predictive error under our model and the two-parameter CRP are reported in Table 1. As expected, the predictive under our model outperforms the two-parameter CRP, which is misspecified in that case. Posterior predictive of the proportion of clusters of a given size is reported in the first row of Figure 7.

Real data. We consider two datasets from the SNAP database [44] of the same size $n = 15000$ with $n_{train} = 5000$. The first one is the Amazon dataset of movies' reviews [54] where each movie represents a cluster containing its reviews, which are ordered. The second dataset is a time-ordered collection of answers to questions in the Math Overflow website[1] where the clusters contain answers to the same question. Evolutions of the cluster sizes are reported in Figure 4(b-c) for these datasets. We aim at predicting, based on the training set, the number of reviews to a given movie for the Amazon dataset, and the number of questions answered to a given question for the Math Overflow dataset. In both cases our non-exchangeable model provides better predictions of the cluster sizes (see Table 1 and Figure 6). Estimates and credible intervals for the parameter σ and σ_2 are reported in Table 2. Figure 7 shows that both models give reasonable predictive fit to the proportion of clusters of a given size.

7.2. *Partially observed and unobserved partition.* In this subsection we consider that observations are word counts in documents, which are modeled using a hierarchical Dirichlet-Multinomial model, and we aim at finding a partition of the documents into topics, and predicting the growth of these topics. Let V denote the size of the vocabulary, and let y_i, $i = 1, \ldots, n$,

[1] https://mathoverflow.net/

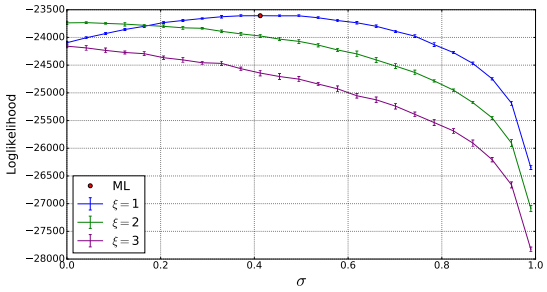


FIG 5. *Loglikelihood estimates for $\zeta = 10$ and different values of ξ and σ. For every grid point, 10 SMC estimates are obtained, and the mean and ± 1 standard deviation error bars are reported.*

TABLE 1
Mean and quantiles of the predictive error using 100 samples from the predictive distributions, when the partition is observed.

	Non-exchangeable		Two-parameter CRP	
	L2 error	90% CI	L2 error	90% CI
Synthetic	6.92	[6.37, 7.42]	14.9	[13.4, 16.2]
Amazon	6.14	[5.58, 6.96]	8.43	[7.35, 9.42]
Math Overflow	1.08×10^2	$[1.01, 1.22] \times 10^2$	1.67×10^2	$[1.58, 1.77] \times 10^2$

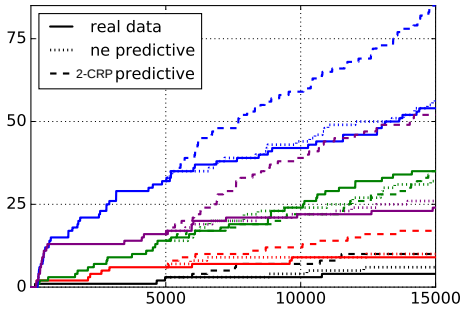


FIG 6. *Amazon dataset. Observed (plain line) and predicted sizes of some clusters (in different colours) from the non-exchangeable (dotted line) and the two-parameter CRP models (dashed line).*

be a V-dimensional vector of word counts in document i. Assume that Π_n follows our non-exchangeable random partition model with $\mathrm{GG}(\alpha, \sigma, \zeta)$ and

TABLE 2
MLE for the the parameter σ of the non-exchangeable model and the discount parameter σ_2 of the two-parameter CRP model, and 0.025 and 0.975 quantiles of their posterior distributions, when the partition is observed.

	Non-exchangeable		Two-parameter CRP	
	MLE of σ	$[q_{0.025}, q_{0.975}]$	MLE of σ_2	$[q_{0.025}, q_{0.975}]$
Synthetic	0.41	[0.38, 0.45]	0.46	[0.41, 0.51]
Amazon	0.63	[0.57, 0.66]	0.65	[0.60, 0.69]
Math Overflow	0.37	[0.34, 0.40]	0.30	[0.24, 0.36]

base measure $\alpha(d\theta) = \xi\theta^{\xi-1}d\theta$, and

$$U_j \overset{iid}{\sim} \text{Dirichlet}(h, \dots, h) \qquad j = 1, \dots, k_n$$
$$y_i \mid \Pi_n, U_1, \dots, U_{k_n} \overset{ind}{\sim} \text{Multinomial}(N_i, U_{c_i}) \qquad i = 1, \dots, n$$

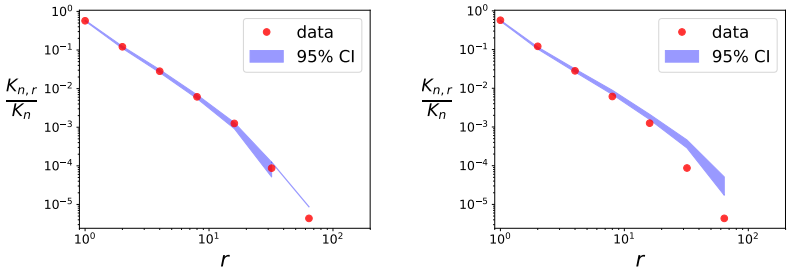
where U_j is a V-dimensional probability vector defining the word frequencies for the j's cluster/topic and $h > 0$ is the scale parameter of the Dirichlet distribution.

As before, we are interested in predicting the growth of the clusters' sizes and consider the following error which generalizes the error measure defined in the previous subsection to the case of unobserved partitions:

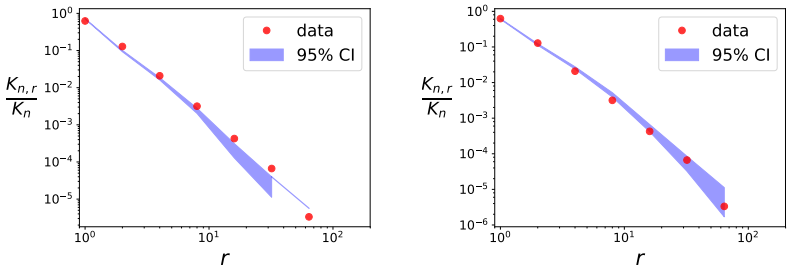
$$\widetilde{E} = \frac{1}{n_{\text{test}}} \sum_{k=1}^{K_{n_{\text{train}}}} ((m_{k,n} - m_{k,n_{\text{train}}}) - (\hat{m}_{k,n} - \hat{m}_{k,n_{\text{train}}}))^2$$

where $\hat{m}_{k,n}$ is the size of the k-th cluster at time n inferred by the model. In the following experiments we consider two datasets of size $n = 4000$, and split each dataset into three subsets. The first $n_{\text{pre-train}} = 1000$ observations are provided with labels and are used in the pre-training stage to tune the parameters of the random partition models, which are estimated with MLEs for both the non-exchangeable model and the two-parameter CRP, as done in the previous experiments. Given the estimates of the parameter, the objective is to infer and predict the partition of the remaining unlabelled observations. For this purpose, the following $n_{\text{train}} = 1000$ observations are used to infer the partition, and the last $n_{\text{test}} = 2000$ points are used to compare prediction performances.

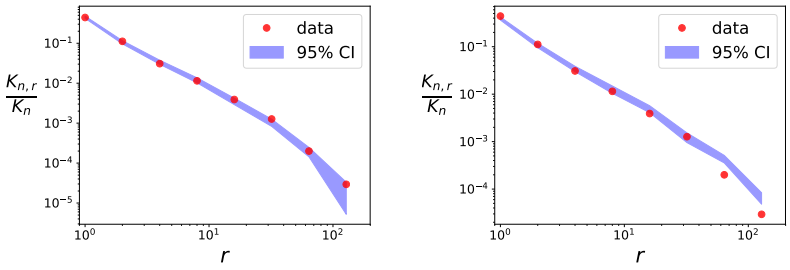
For the non-exchangeable model the partition is inferred using a SMC algorithm which gives an approximation of the distribution of the cluster labels, as described in Section 4.2.2, while a Gibbs sampler is used for the two-parameter CRP model.



(a) Synthetic



(b) Amazon



(c) Math Overflow

FIG 7. *Empirical proportions of clusters of given size (red dots) and* 95% *posterior predictive credible intervals (blue) for our non-exchangeable model (left) and the two-parameter CRP (right).*

For each dataset we consider two scenarios, corresponding to the clustering task in the semi-supervised and unsupervised settings. In the first one,

TABLE 3
Mean and quantiles of the predictive error using 100 samples from the predictive distributions when the partition of the data is partially observed. Lower is better.

	Non-exchangeable		Two-parameter CRP	
	L2 error	90% CI	L2 error	90% CI
Synthetic	**3.83**	[2.99, 4.46]	4.00	[3.47, 4, 64]
Amazon	**17.06**	[9.66, 30.29]	336.02	[303.96 366.31]

TABLE 4
Mean and quantiles of the predictive error using 100 samples from the predictive distributions when the partition of the data is unknown. Lower is better.

	Non-exchangeable		Two-parameter CRP	
	L2 error	90% CI	L2 error	90% CI
Synthetic	**3.85**	[3.12, 4.62]	4.07	[3.30 4.90]
Amazon	**117.70**	[96.85, 137.51]	605.46	[556.84 654.93]

the first 200 observations in the training set of size $n_{\text{train}} = 1000$ are provided with labels, while in the second case the partition of the training set is fully unobserved.

Synthetic data. A dataset of $n = 4000$ observations $(y_1, \dots, y_n)$ is sampled with parameters $(\xi, \sigma, \zeta) = (1, 0.5, 1)$, $V = 15$, $N_i = 1000$, $h = 1$. Table 5 reports the estimates of the power-law parameters, and both models obtain rather precise estimates. Tables 3 and 4 shows that the non-exchangeable model outperforms the two-parameter CRP in predicting the partition of the observations in the test set. It is worth noting that providing the labels for the first 200 observations in the training set does not improve the predictive performances significantly. This is due to the fact that the number of counts N_i per observation is much bigger than the dimensionality V of the Dirichlet parameters, which facilitates the clustering. This in not the case in the real data example, where predictive performances will differ between the semi-supervised and unsupervised cases.

Real data. We use again the Amazon dataset of movies' reviews, and consider the actual reviews as observations in a bag-of-words representation. We preprocess the dataset using the text-mining R package *tm* [26, 27] retaining the first time-ordered $n = 4000$ reviews. Estimates of the parameters σ and σ_2 can be found in Table 5 and the prediction errors in Tables 3 and 4, showing that the proposed model provides more accurate predictions than the two-parameter CRP for this dataset.

TABLE 5
MLE for the the parameter σ of the non-exchangeable model and posterior mean of the discount parameter σ_2 of the two-parameter CRP model, and 0.025 and 0.975 quantiles of their posterior distributions.

	Non-exchangeable		Two-parameter CRP	
	MLE of σ	$[q_{0.025}, q_{0.975}]$	Posterior mean of σ_2	$[q_{0.025}, q_{0.975}]$
Synthetic	0.54	[0.50, 0.58]	0.43	[0.37, 0.49]
Amazon	0.58	[0.54, 0.62]	0.65	[0.60, 0.70]

Supplementary material. A demo of the simulation and inference for the non-exchangeable random partition model can be found at https://github.com/giuseppedib/microclustering.

Acknowledgements. GDB acknowledges support from EPSRC under grant EP/L016710/1. FC and YWT acknowledge funding from the ERC under the European Union's 7th Frame-work programme (FP7/2007-2013) ERC grant agreement no. 617071. FC acknowledges support from EPSRC under grant EP/P026753/1.

APPENDIX A: PROOF OF THE MAIN THEOREMS

PROOF OF PROPOSITION 2. From Eq. (3.2), the joint distribution of $((\theta_{(i)}, \tau_{(i)})_{i=1,\dots,n})$ is given by

$$\begin{aligned}
&\mathbb{P}(\Pi_n = \{A_{n,1}, \dots, A_{n,k_n}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \dots, k_n, K_n = k_n, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n) \\
&= \left\{ \sum_{i=1}^{k_{n-1}} \left[\prod_{\substack{j=1 \\ j \neq i}}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \kappa(m_{n-1,i} + 1, t_n - x_i^*) \tilde{\alpha}(x_i^*)\, \delta_{x_i^*}(dx_{k_n}^*) \right. \\
&\left. + \left[\prod_{j=1}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \kappa(1, t_n - x_n^*) \alpha(dx_{k_n}^*) \right\} \\
&\times e^{-\int_0^{t_n} \psi(t_n - x) \alpha(dx)} \left[\prod_{i=1}^{n} \mathbb{1}_{x_{c_i}^* < t_i} \right] \mathbb{1}_{t_1 < t_2 < \dots < t_n} dt_1 \dots dt_n.
\end{aligned}$$

Integrating over the location of the n-th point, we obtain

$$\begin{aligned}
&\mathbb{P}(\Pi_{n-1} = \{A_{n,1}, \dots, A_{n-1,k_{n-1}}\}, \theta_j^* \in dx_j^* \text{ for } j = 1, \dots, k_{n-1}, \\
&\quad K_{n-1} = k_{n-1}, \tau_{(i)} \in dt_i \text{ for } i = 1, \dots, n) \\
&\quad = \left\{ \sum_{i=1}^{k_{n-1}} \left[\prod_{\substack{j=1 \\ j \neq i}}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \kappa(m_{n-1,i} + 1, t_n - x_i^*) \tilde{\alpha}(x_i^*) \right. \\
&\quad \left. + \left[\prod_{j=1}^{k_{n-1}} \kappa(m_{n-1,j}, t_n - x_j^*) \alpha(dx_j^*) \right] \int_0^{t_n} \kappa(1, t_n - x) \alpha(dx) \right\} \\
&\quad \times e^{-\int_0^{t_n} \psi(t_n - x) \alpha(dx)} \left[\prod_{i=1}^{n-1} \mathbb{1}_{x_{c_i}^* < t_i} \right] \mathbb{1}_{t_1 < t_2 < \dots < t_n} dt_1 \dots dt_n.
\end{aligned}$$

In the generalised gamma Process case we have

$$\kappa(m, u) = \frac{1}{\Gamma(1 - \sigma)} \frac{\Gamma(m - \sigma)}{(\zeta + u)^{m - \sigma}}.$$

Therefore $\kappa(m+1,u)=\kappa(m,u)\frac{m-\sigma}{\zeta+u}$ and

$$\sum_{i=1}^{k_{n-1}}\left[\prod_{\substack{j=1\\ j\neq i}}^{k_{n-1}}\kappa(m_{n-1,j},t_n-x_j^*)\alpha(dx_j^*)\right]\kappa(m_{n-1,i}+1,t_n-x_i^*)\alpha(dx_i^*)$$
$$=\left[\prod_{j=1}^{k_{n-1}}\kappa(m_{n-1,j},t_n-x_j^*)\alpha(dx_j^*)\right]\sum_{i=1}^{k_{n-1}}\frac{m_{n-1,i}-\sigma}{t_n-x_i^*+\zeta}$$

hence

$$\begin{aligned}&\mathbb{P}(\Pi_{n-1}=\{A_{n,1},\dots,A_{n-1,k_{n-1}}\},\theta_j^*\in dx_j^*\text{ for }j=1,\dots,k_{n-1},\\&\quad K_{n-1}=k_{n-1},\tau_{(i)}\in dt_i\text{ for }i=1,\dots,n)\\&=\frac{1}{\Gamma(1-\sigma)^{k_{n-1}}}\left[\prod_{j=1}^{k_{n-1}}\frac{\Gamma(m_{n-1,j}-\sigma)\,\alpha(dx_j^*)}{(t_n-x_j^*+\zeta)^{m_{n-1,j}-\sigma}}\right]\\&\times\left(\sum_{j=1}^{k_{n-1}}\frac{m_{n-1,j}-\sigma}{t-x_j^*+\zeta}+\int_0^{t_n}\frac{\alpha(dx)}{(t_n-x+\zeta)^{1-\sigma}}\right)\\&\times e^{-\int_0^{t_n}\psi(t_n-x)\alpha(dx)}\left[\prod_{i=1}^{n-1}\mathbb{1}_{x^*_{c_i}<t_i}\right]\mathbb{1}_{t_1<\dots<t_n}dt_1\dots dt_n.\end{aligned}$$

from which we obtain the results of the theorem. □

Proof of proposition 3 and theorem 4. Given W, $N(t)$ is a non-homogeneous Poisson process with rate $\overline{W}(t)=\sum_{j\geq1}\omega_j\mathbb{1}_{\vartheta_j\leq t}$. Hence, using Fubini's and Campbell's theorems,

$$\mathbb{E}[N(t)]=\mathbb{E}\left[\int_0^t\overline{W}(x)dx\right]=\overline{\overline{\alpha}}(t)\kappa(1,0)$$

where $\overline{\overline{\alpha}}(t)=\int_0^t\overline{\alpha}(t)$. Similarly,

$$\begin{aligned}\operatorname{var}(N(t))&=\operatorname{var}(\mathbb{E}[N(t)\,|\,W])+\mathbb{E}[\operatorname{var}(N(t)\,|\,W)]=\operatorname{var}(\overline{\overline{W}}(t))+\mathbb{E}[\overline{\overline{W}}(t)]\\&=\kappa(2,0)\int_0^t(t-\theta)^2\alpha(d\theta)+\kappa(1,0)\overline{\overline{\alpha}}(t)\\&=\kappa(2,0)\int_0^t\overline{\overline{\alpha}}(x)dx+\kappa(1,0)\overline{\overline{\alpha}}(t)\end{aligned}$$

Using Karamata's Theorem [5, Proposition 1.5.8] and Assumption (A3), we obtain that

$$\mathbb{E}[N(t)] \sim \frac{\kappa(0,1)}{\xi+1} t^{\xi+1} L(t) \text{ and } \operatorname{var}(N(t)) \sim \frac{\kappa(2,0)}{(\xi+1)(\xi+2)} t^{\xi+2} L(t).$$

Therefore, for any $0 < a < \xi$ we have $\operatorname{var}(N(t)) = O(t^{-a}\mathbb{E}[N(t)]^2)$. Using [15, Lemma B.1], we conclude that, almost surely as t tends to infinity

$$N(t) \sim \mathbb{E}[N(t)] \sim \frac{\kappa(1,0)}{\xi+1} t^{\xi+1} L(\xi).$$

Conditional on W, $X_j(t)$ is a non-homogeneous Poisson process with rate $\omega_j \mathbb{1}_{\vartheta_j > t}$ hence

$$X_j(t) \sim \omega_j t$$

almost surely as t tends to infinity. It follows similarly that $M_j(t)$, the size, at time t, of the jth cluster to appear, satisfies

$$M_j(t) \sim \omega_j^* t$$

where $\omega_j^* = W(\{\theta_j^*\})$. Additionally, in the same fashion as in [58, Theorem 4.5], Proposition 1 implies that, denoting by τ_j^* the arrival time of the j-th cluster and by ω_j^*, θ_j^* its weight and location respectively,

$$\{\tau_j^*, \theta_j^*, \omega_j^*\}_{j\geq 1} \sim \text{Poisson}\left(\omega\, e^{-\omega(\tau-\theta)_+} \mathbb{1}_{\theta<\tau} d\tau\, \alpha(d\theta)\rho(d\omega)\right)$$

from which it follows that, given a sequence of iid $e_k \sim \text{Exp}(1)$, the clusters' arrival times are defined as $\tau_j^* = \Phi^{-1}(\gamma_j)$, where $\gamma_j = \sum_{k=1}^{j} e_k$. Proposition 2 then implies the result.

□

Proof of proposition 5. Observe that $\mathbb{P}(X_j(t) > 0 \mid W) = 1 - e^{-\omega_j(t-\vartheta_j)_+}$. By the marking theorem [41, Chapter 5], for each t, $\{(\omega_j, \vartheta_j) \mid j \geq 1, X_j(t) > 0\}$ is a Poisson point process with mean measure $\rho(d\omega)\alpha(d\theta)(1 - e^{-\omega(t-\theta)_+})$. It follows that

$$\mathbb{E}[K(t)] = \operatorname{var}[K(t)] = \int_0^\infty \int_0^t \left(1 - e^{-(t-\theta)\omega}\right) \alpha(d\theta)\rho(d\omega) = \int_0^t \psi(t-\theta)\alpha(d\theta).$$

Similarly to [30, Proposition 2], it follows from the monotonicity of $K(t)$ and the Borel-Cantelli Lemma that $K(t) \sim \mathbb{E}[K(t)]$ almost surely as $t \to$

∞. Using the Tauberian theorems [28, Chapter XIII, Section 5] recalled in Lemma 13, Lemma 14 and $\alpha(t) \sim \xi t^{\xi-1} L(t)$, we obtain

$$\mathbb{E}[K(t)] \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}$$

as t tends to infinity.

We proceed similarly for $K_r(t)$. For each $t > 0$ and $r \geq 1$, $\{(\omega_j, \vartheta_j) \mid j \geq 1, X_j(t) = r\}$ is a Poisson point process with mean measure $\rho(d\omega)\alpha(d\theta)\frac{\omega^r (t-\theta)_+^r}{r!} e^{-\omega(t-\theta)_+}$. It follows that

$$\begin{aligned}
\mathbb{E}[K_r(t)] &= \operatorname{var}[K_r(t)] \\
&= \int_0^t \int_0^\infty \frac{(t-\theta)^r}{r!}\, \omega^r e^{-(t-\theta)\omega} \rho(d\omega)\alpha(d\theta) \\
&= \int_0^t \frac{(t-\theta)^r}{r!} \kappa(r, t-\theta)\alpha(d\theta).
\end{aligned}$$

Using Lemma 13 and 14, we obtain: if $\sigma = 0$, $\mathbb{E}[K_r(t)] = o(L(t)\ell(t))t^{\sigma+\xi}$; if $\sigma \in (0,1)$,

$$\mathbb{E}[K_r(t)] \sim \frac{\sigma\Gamma(r-\sigma)}{r!\Gamma(1-\sigma)} \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}.$$

If $\sigma = 1$, $\mathbb{E}[K_1(t)] \sim \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} L(t)\ell_\sigma(t)\, t^{\sigma+\xi}$ and $\mathbb{E}[K_r(t)] = o(L(t)\ell(t))t^{\sigma+\xi}$ for all $r \geq 2$. For the almost sure result, we proceed as for $K(t)$, using the monotonicity of $\sum_{r\geq s} K_r(t)$ [30, Corollary 21], the equality $\operatorname{var}\left[\sum_{r\geq s} K_r(t)\right] = \mathbb{E}\left[\sum_{r\geq s} K_r(t)\right]$ and the fact that $\mathbb{E}[K_r(t)] \asymp K(t)$ for $\sigma \in (0,1)$, $\mathbb{E}[K_r(t)] = o(K(t))$ for $\sigma = 0$ and $\mathbb{E}[K_1(t)] \sim K(t)$ for $\sigma = 1$. □

Proof of proposition 10. The result follows from known results in Poisson partition calculus. Let us define $f(\omega, \theta) = (\tau_{(n)} - \theta)_+\omega$ and denote by $\mathcal{P}(dW \mid \nu)$ the distribution of the CRM W with Lévy measure ν, then Proposition 3.1 part (ii) in [33] provides the Bayes formula for our model:

the joint distribution of the CRM W and $(\tau_{(i)}, \theta_{(i)})_{i=1,\dots,n}$ can be written as

$$
\begin{aligned}
&e^{-\int_0^{\tau_{(n)}} \overline{W}(t)dt} \left[\prod_{i=1}^{n} \mathbb{1}_{\theta_i<\tau_{(i)}}\right] \mathbb{1}_{\tau_{(1)}<\cdots<\tau_{(n)}} d\tau_{(1)} \dots d\tau_{(n)} \mathcal{P}(dW \mid \nu) \\
&= \mathcal{P}(dW \mid e^{-f}\nu) \left[\prod_{j=1}^{k_n} \kappa(m_{n,j}, t_n - x_j^*)\alpha(dx_j^*)\right] e^{-\int_0^{t_n} \psi(t_n-\theta)\alpha(d\theta)} \\
&\quad \times \left[\prod_{i=1}^{n} \mathbb{1}_{x^*_{c_i}\le t_i}\right] \mathbb{1}_{t_1<t_2<\ldots<t_n} \, dt_1 \dots dt_n
\end{aligned}
$$

where the first term $\mathcal{P}(dW \mid e^{-f}\nu)$ is the posterior distribution of W and the remaining terms describe the marginal distribution as in Proposition 1. The result then follows from formulas (33) and (34) in [33]. □

APPENDIX B: BACKGROUND ON REGULAR VARIATION AND TECHNICAL LEMMA

We recall the following definitions which can be found in [5] and [65].

DEFINITION 12 (Regularly varying function). *A measurable function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *is* regularly varying *at* ∞ *with index* $\xi \in \mathbb{R}$ *if for every* $x > 0$

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^{\xi}.$$

If $\xi = 0$ we say that the function is *slowly varying*. An important property of the regularly varying function is that they can be written as $f(x) = \ell(x)x^{\xi}$ where ξ is the exponent of variation and ℓ is a slowly varying function.

Let $L^{\#}$ be the de Brujin conjugate [5] of a slowly varying function L. Regularly varying functions $f(x) = L(x)x^{\xi}$ of index $\xi > 0$ admit asymptotic inverse $g(x) = L^*_{\xi}(x)x^{1/\xi}$ which are regularly varying of index ξ^{-1} (see [5, Proposition 1.5.15] or [30, Lemma 22]) with slowly varying part

$$L^*_{\xi}(x) = \{L^{1/\xi}(x^{1/\xi})\}^{\#}. \tag{B.1}$$

Note that if $L(t) = c$, then $L^*_{\xi}(t) = c^{1/\xi}$. From Equation (4.1) and Proposition 5, it follows that the slowly varying function appearing in Corollary 6

is

$$\widetilde{\ell}(n) = \frac{\Gamma(\sigma+1)\Gamma(\xi+1)}{\Gamma(\sigma+\xi+1)} \left(\frac{\xi+1}{\kappa(1,0)}\right)^{(\sigma+\xi)/(\xi+1)} L^*_{\xi+1}(n)^{\sigma+\xi}$$

(B.2)

$$\times L\left\{\left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} n^{1/(\xi+1)} L^*_{\xi+1}(n)\right\} \ell_\sigma \left\{\left(\frac{\xi+1}{\kappa(1,0)}\right)^{1/(\xi+1)} n^{1/(\xi+1)} L^*_{\xi+1}(n)\right\}.$$

The following lemma is a compilation of Tauberian results in Propositions 17, 18 and 19 in [30]. See also [28, Chapter XIII].

LEMMA 13. *Let ρ be a Lévy measure on $(0,\infty)$ with tail Lévy intensity $\overline{\rho}(x) = \int_x^\infty \rho(d\omega)$. Assume*

$$\overline{\rho}(x) \sim x^{-\sigma}\ell(1/x)$$

as x tends to 0, where $\sigma \in [0,1]$ and ℓ is a slowly varying function at infinity. For any $\sigma \in [0,1)$,

$$\psi(t) \sim \Gamma(1-\sigma)t^\sigma \ell(t)$$

and for $r = 1, 2, \ldots$

$$\begin{cases} \kappa(r,t) \sim t^{\sigma-r}\ell(t)\Gamma(r-\sigma) & \text{if } \sigma \in (0,1) \\ \kappa(r,t) = o(t^{\sigma-r}\ell(t)) & \text{if } \sigma = 0 \end{cases}$$

as t tends to infinity. For $\sigma = 1$,

$$\psi(t) \sim t\ell_1(t)$$
$$\kappa(1,t) \sim \ell_1(t)$$

and

$$\kappa(r,t) \sim t^{1-r}\ell(t)\Gamma(r-1)$$

for all $r \geq 2$ as t tends to infinity, where $\ell_1(t) = \int_t^\infty x^{-1}\ell(x)dx$.

LEMMA 14. *Let f and g be locally bounded, regularly varying functions with $f(x) = \ell_f(x)x^a$ and $g(x) = \ell_g(x)x^b$ where $a, b > -1$ and ℓ_f, ℓ_g are slowly varying functions. Then as t tends to infinity*

$$\int_0^t f(x)g(t-x)dx \sim \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} tf(t)g(t)$$
$$\sim \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} \ell_f(t)\ell_g(t)t^{a+b+1}.$$

PROOF. Let us split the integral in the following way

$$\int_0^t f(x)g(t-x)dx = \int_0^{\frac{t}{2}} f(x)g(t-x)dx + \int_0^{\frac{t}{2}} f(t-x)g(x)dx.$$

Let $\delta \in (0, \min(a,b)+1)$. From Potter's Theorem [5, Theorem 1.5.6], there is X such that for all $t > 2X$, $u \in [X/t, 1/2]$,

$$\frac{f(tu)}{f(t)} \leq 2u^{a-\delta}, \quad \frac{g(t(1-u))}{g(t)} \leq 2(1-u)^{b-\delta}.$$

Take $t > 2X$. We have

$$\int_X^{\frac{t}{2}} \frac{f(x)g(t-x)}{tf(t)g(t)}\,dx = \int_0^{\frac{1}{2}} \frac{f(ut)}{f(t)} \frac{g((1-u)t)}{g(t)} \mathbb{1}_{u\in[X/t,1/2]}\,du$$

where the integrand function is bounded by $4u^{a-\delta}(1-u)^{b-\delta}$ which is integrable, hence we have convergence to $\int_0^{\frac{1}{2}} u^a(1-u)^b du \in (0,\infty)$ by the dominated convergence theorem. We proceed analogously for the second integral. Since $\int_0^1 u^a(1-u)^b du = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$, we have the result. □

APPENDIX C: ALGORITHM'S DETAILS

This section presents details about the parameter estimation in the non-exchangeable model in the case the partition in not observed and the observations follow the model described in Equations (4.4-4.6). The random partition model's parameters $\eta = (\xi, \sigma_0, \zeta_0)$ are estimated by maximum likelihood; marginal likelihood estimates are obtained with a Sequential Monte Carlo algorithm. Let us denote the unobserved variables as $x_i = (c_i, \theta_{(i)}, \tau_{(i)})$ for $i = 1, \ldots, n$. Defining for each $n \in \mathbb{N}$

$$f(n) := \left[\prod_{j=1}^{k_{n-1}} \frac{\kappa(m_{n-1,j}, \tau_{(n)} - \theta_j^*)}{\kappa(m_{n-1,j}, \tau_{(n-1)} - \theta_j^*)}\right] \times e^{-\int_0^{\tau_n} \psi(\tau_{(n)}-\theta)\alpha(d\theta) + \int_0^{\tau_{n-1}} \psi(\tau_{(n-1)}-\theta)\alpha(d\theta)} \mathbb{1}_{\theta_{(n)}<\tau_{(n)}} \mathbb{1}_{\tau_{(n-1)}<\tau_{(n)}}$$

the predictive distribution can be written as

$$p(x_n \,|\, x_{1:n-1}, \eta) = \begin{cases} f(n)\frac{m_{n-1,j}-\sigma}{\tau_{(n)}+\theta_j^*-\zeta} & c_n = j \in \{1, \ldots, k_{n-1}\} \\ f(n)\frac{\alpha(\theta_{(n)})}{(\zeta+\tau_{(n)}-\theta_{(n)})^{1-\sigma}} & c_n = k_{n-1}+1 \end{cases}$$

Denoting by N_p the number of particles in the SMC sampler, the marginal likelihood estimate can be computed using the importance weights of the

particles (see [21],[2]): $\hat{p}(y_{1:n} \mid \eta) = \hat{p}(y_1 \mid \eta) \prod_{t=2}^{n} \hat{p}(y_t \mid y_{1:t-1}, \eta)$ with $\hat{p}(y_t \mid y_{1:t-1}, \eta) = \frac{1}{N_p} \sum_{j=1}^{N_p} w_t^{(j)}$ and

$$w_t^{(j)} = \frac{p(x_t^{(j)}, y_t \mid x_{1:t-1}^{(j)}, y_{1:t-1}, \eta)}{q(x_t^{(j)} \mid x_{1:t-1}^{(j)}, y_{1:t}, \eta)} \qquad \text{for all } j = 1, \dots, N_p$$

where q denotes the proposal distribution which we define as

$$\begin{aligned} q(x_n | x_{1:n-1}, y_{1:n}, \eta) = {} & q(\tau_{(n)} \mid \tau_{(n-1)})\, q(c_n \mid \tau_{(n)}, x_{1:n-1}, y_n, \eta) \mathbb{1}_{c_n \in \{1,\dots,k_{n-1}\}} \\ & + q(\tau_{(n)} \mid \tau_{(n-1)})\, q(c_n \mid \tau_{(n)}, x_{1:n-1}, y_{1:n}, \eta)\, q(\theta_n \mid \tau_{(n)}) \mathbb{1}_{c_n = k_{n-1}+1} \end{aligned}$$

with $q(\tau_{(n)} \mid \tau_{(n-1)})$ being the density of a truncated Gaussian distributions centered at $\tau_{(n-1)}$, $q(\theta_n \mid \tau_{(n)}) = 1/\tau_{(n)} \mathbb{1}_{\theta_n \leq \tau_{(n)}}$, and

$$q(c_n \mid \tau_{(n)}, x_{1:n-1}, y_{1:n}, \eta) \propto \begin{cases} \frac{m_{n-1,j} - \sigma}{\tau_{(n)} - \theta_j^* + \zeta} \frac{g(y_{A_{n-1,j} \cup \{n\}})}{g(y_{A_{n-1,j}})} & c_n = j \in \{1, \dots, k_{n-1}\} \\ H_{\tau_{(n)}}(\mathbb{R}_+)\, g(y_{\{n\}}) & c_n = k_{n-1} + 1 \end{cases}$$

where g is defined in Equation (4.5) and $H_{\tau_{(n)}}$ in Equation (3.4). Therefore we have that the importance weight of the j-th particle at time t can be written as

$$w_t^{(j)} = \begin{cases} \frac{Sf(t)}{q(\tau_{(t)}^{(j)} \mid \tau_{(t-1)}^{(j)})} & c_t = j \in \{1, \dots, k_{t-1}\} \\ \frac{Sf(t)}{q(\tau_{(t)}^{(j)} \mid \tau_{(t-1)}^{(j)}) q(\theta_{(t)}^{(j)} \mid \tau_{(t)}^{(j)}) H_{\tau_{(t)}^{(j)}}(\mathbb{R}_+)} & c_t = k_{t-1} + 1 \end{cases}$$

where S is the normalising constant.

REFERENCES

[1] Aldous, D. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983* 1–198.

[2] Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 269–342.

[3] Bacallado, S., Favaro, S. and Trippa, L. (2015). Looking-backward probabilities for Gibbs-type exchangeable random partitions. *Bernoulli* **21** 1–37.

[4] Betz, V., Ueltschi, D. and Velenik, Y. (2011). Random permutations with cycle weights. *The Annals of Applied Probability* **21** 312–331.

[5] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987). *Regular variation* **27**. Cambridge university press.

[6] Blei, D. and Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* **12** 2461–2488.

[7] Bloem-Reddy, B. and Orbanz, P. (2017). Preferential Attachment and Vertex Arrival Times. *arXiv preprint arXiv:1710.02159*.

[8] Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* 929–953.
[9] Broderick, T., Jordan, M. I. and Pitman, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian analysis* **7** 439–476.
[10] Cai, D., Campbell, T. and Broderick, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 4249–4257. Curran Associates, Inc.
[11] Camerlenghi, F., Lijoi, A., Orbanz, P., Prünster, I. et al. (2019a). Distribution theory for hierarchical processes. *The Annals of Statistics* **47** 67–92.
[12] Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., Rodríguez, A. et al. (2019b). Latent nested nonparametric priors (with discussion). *Bayesian Analysis* **14** 1303–1356.
[13] Caron, F., Davy, M. and Doucet, A. (2007). Generalized Pólya urn for time-varying Dirichlet process mixtures. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007* 33–40.
[14] Caron, F. and Fox, E. (2017). Sparse Graphs using Exchangeable Random Measures. *Journal of the Royal Statistical Society B* **79** 1295-1366. Part 5.
[15] Caron, F. and Rousseau, J. (2017). On sparsity and power-law properties of graphs based on exchangeable point processes. *arXiv preprint arXiv:1708.03120*.
[16] Caron, F., Neiswanger, W., Wood, F., Doucet, A. and Davy, M. (2017). Generalized Pólya urn for time-varying Pitman–Yor processes. *Journal of Machine Learning Research* **18** 1-32.
[17] Chen, C., Rao, V., Buntime, W. and Teh, Y. W. (2013). Dependent Normalized Random Measures. In *International Conference on Machine Learning*.
[18] Crane, H. and Dempsey, W. (2017). Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*.
[19] Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*, second ed. Springer.
[20] De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE transactions on pattern analysis and machine intelligence* **37** 212–229.
[21] Del Moral, P. (2004). Feynman-kac formulae. In *Feynman-Kac Formulae* 47–93. Springer.
[22] Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 411–436.
[23] Doucet, A., de Freitas, N. and Gordon, N., eds. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.
[24] Durrett, R., Schweinsberg, J. et al. (2005). Power laws for family sizes in a duplication model. *The Annals of Probability* **33** 2094–2126.
[25] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability* **23** 1721–1754.
[26] Feinerer, I., Hornik, K. and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* **25** 1–54.
[27] Feinerer, I. and Hornik, K. (2018). tm: Text Mining Package R package version 0.7-6.
[28] Feller, W. (1971). *An introduction to probability theory and its applications* **2**. John Wiley & Sons.
[29] Foti, N. J. and Williamson, S. A. (2015). A survey of non-exchangeable priors for

Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence* **37** 359–371.
[30] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv* **4** 88.
[31] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical sciences* **138** 5674–5685.
[32] Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association* **101** 179–194.
[33] James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*.
[34] James, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics* **33** 1771–1799.
[35] James, L. F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36** 76–97.
[36] Janson, S. (2017). On edge exchangeable random graphs. *arXiv preprint arXiv:1702.06396*.
[37] Karlin, S. (1967). Central Limit Theorems for Certain Infinite Urn Schemes. *Journal of Mathematics and Mechanics* **17** 373-401.
[38] Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21** 59–78.
[39] Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–22.
[40] Kingman, J. F. C. (1978). Random partitions in population genetics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **361** 1–20. The Royal Society.
[41] Kingman, J. F. C. (1993). *Poisson processes* **3**. Oxford University Press, USA.
[42] Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability* 705–711.
[43] Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16** 526–558.
[44] Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
[45] Lijoi, A., Mena, R. H. and Prünster, I. (2005a). Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Statistical Inference for Stochastic Processes* **8** 283–309.
[46] Lijoi, A., Mena, R. H. and Prünster, I. (2005b). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100** 1278–1291.
[47] Lijoi, A., Mena, R. and Prünster, I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786.
[48] Lijoi, A., Mena, R. H. and Prünster, I. (2007b). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 715–740.
[49] Lijoi, A. and Prünster., I. (2003). On a normalized random measure with independent increments relevant to Bayesian nonparametric inference. In *Proceedings of the 13th European Young Statisticians Meeting* 123-134. Bernoulli Society.
[50] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.) Cambridge University Press.

[51] Lijoi, A., Nipoti, B., Prünster, I. et al. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291.
[52] Lin, D., Grimson, E. and Fisher, J. W. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in neural information processing systems* 1396–1404.
[53] MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science* 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.
[54] McAuley, J., Targett, C., Shi, Q. and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* 43–52. ACM.
[55] Miller, J., Betancourt, B., Zaidi, A., Wallach, H. and Steorts, R. (2015). Microclustering: When the Cluster Sizes Grow Sublinearly with the Size of the Data Set. *arXiv:1512.00792v1*.
[56] Müller, P. and Rodriguez, A. (2013). *Chapter 8: Random Partition Models.* In *Nonparametric Bayesian Inference. NSF-CBMS Regional Conference Series in Probability and Statistics* **Volume 9** 87–92. Institute of Mathematical Statistics and American Statistical Assocation, Beachwood, Ohio, USA; and Alexandria, Virginia, USA.
[57] Nieto-Barajas, L. E., Prünster, I. and Walker, S. G. (2004). Normalized random measures driven by increasing additive processes. *The Annals of Statistics* **32** 2343–2360.
[58] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92** 21–39.
[59] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102** 145–158.
[60] Pitman, J. (2003). *Poisson-Kingman partitions.* In *Statistics and science: a Festschrift for Terry Speed. Lecture Notes–Monograph Series* **Volume 40** 1–34. Institute of Mathematical Statistics, Beachwood, OH.
[61] Pitman, J. (2006). *Combinatorial stochastic processes. Ecole d'été de Probabilité de Saint-Flour - 2002.* Springer.
[62] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900.
[63] Rao, V. and Teh, Y. W. (2009). Spatial normalized gamma processes. In *Advances in neural information processing systems* 1554–1562.
[64] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* **31** 560–585.
[65] Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling.* Springer Science & Business Media.
[66] Sudderth, E. B. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems* 1585–1592.
[67] Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* 985–992. Association for Computational Linguistics.
[68] Zanella, G., Betancourt, B., Wallach, H., Miller, J. W., Zaidi, A. and Steorts, R. C. (2016). Flexible Models for Microclustering with Application to Entity Resolution. In *Advances in Neural Information Processing Systems* 1417–1425.

[69] Zhou, M., Favaro, S. and Walker, S. G. (2016). Frequency of Frequencies Distributions and Size Dependent Exchangeable Random Partitions. *Journal of the American Statistical Association*.