Bayesian Nonparametric Methods and Applications in Statistical Network Modelling



Xenia Miscouridou

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2019

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. My personal contributions are as outlined in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text, acknowledgements, forms and papers.

Xenia Miscouridou

Michaelmas 2019

This thesis is dedicated to
my family
Chrystala, Demetris, Nicholas and Maria.

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisors, Professors François Caron and Yee Whye Teh. I thank them for their guidance and belief in me. They have supported me in multiple ways throughout my DPhil and their contribution has been crucial. I would like to say a heartfelt thank you to François for being a true mentor to me supporting me with enthusiasm from the very beginning of my DPhil. He is an inspirational academic, unique person and someone I look up to.

I am thankful to my amazing co-authors Adrien Todeschini and Juho Lee. We had a beautiful collaboration and a productive time working together. They are amongst the kindest and smartest people I have encountered in my DPhil. I am thankful to my unique colleagues Giuseppe Di Benedetto and Fadhel Ayed for being true friends. My DPhil would have not been as rich in experiences and laughter without them.

Being a student in the OxWasp CDT has been a wonderful experience which has offered me lifelong friends. The Oxford Statistics Department, OxCML community and the Royal Oak gatherings will always be unforgettable. The Bayesian Nonparametrics reading group was a great group to be part of and seminar series to attend, and I feel proud to have been organising it in my final year. I would like to thank François for being the main contributor of the group, Professor Judith Rousseau for being an inspirational person to talk to and learn from - a role model to me, Francesca Panero for being a unique and trusted friend, Konstantina Palla for her precious advice, Frauke Harms for her warm encouragement and Cian Naik for his so generous help. All the weekly reading groups and seminars I have attended in the Department have been crucial in supporting my knowledge and understanding of Statistics and Machine Learning and I am thankful to the organisers. I would like to thank my trustworthy colleagues Chris Maddison and Benjamin Bloem-Reddy for their

sincere advice and incredible support in multiple occasions. I am grateful to my amazing friend Yannis Assael for believing so much in me and for all the beautiful discussions we had, my deskmate code magician Adam Kosiorek for being a great person to share thoughts and ideas with, my close friends Qinyi Zhang and Leon Law with whom I feel lucky to have shared the same academic journey the last eight years.

My experience would have not been the same without two wonderful administrators; my favourite and extremely efficient Joanna Stoneham and super lovely Beverley Lane for making the Statistics Department feel more like a home than a workplace. I am grateful to my caring flatmates and friends, Philippos Demetriou and Danae Varda for making my life in Oxford so much brighter and the sweet couple of Giuseppe Di Benedetto and Jonida Popi for being the most welcoming people ever. Another beautiful part of my years in Oxford was singing at the Nostos choir which has offered me enjoyable and relaxing music evenings as well as faithful friends.

Spending a year at the Alan Turing Institute has been an unforgettable experience. I had the luck to meet two great people, Alexis Bellot and Leonidas Aristodemou who were an important part of my daily life. Our daily enlightening discussions on a variety of topics have influenced my way of thinking. I am grateful to Professor Chris Holmes for his unique mentorship, constant support and understanding, and I feel extremely fortunate to have had the opportunity to collaborate with his awesome students, Matthew Willetts and Alexander Camuto; my 'Turing lab'.

I am truly grateful and deeply thankful to some amazing people from the community of Statistics and Machine Learning, who have offered me their precious advice, their warm guidance and wise mentorship; Senior Research Scientist Lawrence Murrey, Associate Professors Clara Grazien and Lamiae Azizi and Professors Kerrie Mengersen and Christian Robert.

I could have not completed my research without the financial means. I thank the EPSRC body for providing my PhD funding as well as the A.G. Leventis Foundation in Cyprus for awarding me with academic scholarships to further support my research.

I feel extremely lucky and happy to have had the support of my friends, mentors and big family back home in Cyprus. A heartfelt thank you goes to my two faithful friends Marina and Elena Loizidou. They have always been there for me, lifting me up, deeply believing in me and showing me encouragement that words cannot describe. I thank my lovely and wise grandmother Xenia for giving me all that courage and unconditional love during our lunch time phone calls and my morning walk phone calls to the office. I am grateful to Father Michalis Pegasiou for his constant guidance and the unfeigned advice he has always been giving to me. My choice to study Mathematics and my desire for an academic career roots back in my high-school years and I would like to thank all those teachers that really did inspire me. Specifically I would like to thank my favourite Maths teacher Pakis Pyrgos for believing in me more than any other teacher ever did and much more than I did. I hope I can live up to his expectations.

The contribution of several people in proofreading my thesis has been crucial. I am grateful to my supervisor François Caron for his constructive feedback and valuable input, to my other half Alexis Santis for offering his corrections and suggestions, to my loyal friend Michelle Kokkinou for so kindly offering to proofread my thesis and my incredible sister Maria for her valuable additions and corrections.

I am forever grateful to the most important people in my life, my family. The people to whom I owe who I am today and what I have achieved are my wonderful and infinitely supportive parents Demetris and Chrystala. I thank my superb and inspirational older brother Nicholas for all the support, courage, care and advice he has given to me - including his unique input on surviving the PhD. I thank my amazing and irreplaceable younger sister Maria for always being by my side, taking care of me, cherishing me, and making my life more beautiful. I am thankful to my exceptionally unique Alexis Santis, my special person, for being so heartwarming, indescribably supportive, showing me uncountable love and making my days brighter.

I could not be more grateful to each and every single one of you.

Abstract

Bayesian Statistics provide us with a powerful approach to model real-world phenomena and quantify the uncertainty therein by treating unknown factors as random. The Bayesian paradigm prescribes quantifying the prior knowledge about some state of the world, and after having obtained new information updating that knowledge in order to update the prior beliefs and propose posterior knowledge. A central approach to Bayesian Statistics is modelling, i.e. to represent a data generating process using statistical models equipped with some parameters which are to be estimated via Bayesian inference.

Bayesian Nonparametric modelling comes with great flexibility as it provides infinitely many parameters. Bayesian Nonparametric models are data adaptive because when given a finite set of data, the size of the finite set of parameters to be used, adapts to the complexity of the observed data. Bayesian Nonparametric models have been used in many applications of machine learning such as density estimation, clustering, latent feature models, survival analysis and function approximation. One of the contributions of this thesis to Bayesian Nonparametrics is the formulation of priors on random graphs to model networks. Our motivation for networks sources from the fact that networks are found in numerous areas of modern society reflecting the patterns of connection across a wide range of physical and social phenomena.

Our research focus is on modelling two types of networks; undirected static networks and directed networks with temporal connections. In both cases our objective is to overcome the limitations of the current literature and propose network methodology that captures real-world network properties. Precisely, the property of graph sparsity is particularly important. However, traditional Bayesian network models that assume the desirable property of exchangeability are necessarily not sparse. The first random

graph to allow sparsity as well as exchangeability was recently introduced by Caron and Fox [2017] whose framework we follow. Our objective is to propose models with communities in their latent structure, and we therefore propose a generalisation of existing undirected network models with community structure to the sparse regime. Regarding temporal interaction network data, our goal is to capture reciprocation in interactions and therefore define a family of network models that generalises existing classes of reciprocating models to the sparse graph case.

The key building blocks of our models as well as for various other popular Bayesian Nonparametric constructions, are infinite-activity completely random measures. The use of these random measures offers flexibility, but is also accompanied by computational complexity. The third objective of this thesis is a methodological and theoretical contribution to Bayesian Nonparametrics consisting of a novel framework to form approximations of infinite-activity completely random measures. By providing finite approximations to these infinite data structures we offer practicality and efficiency. Our constructions can be used to develop efficient posterior inference algorithms and shed light on current Bayesian Nonparametric issues of computational complexity.

Contents

1 Introduction **1**

1.1 Motivation 1

1.1.1 Networks 1

1.1.2 Network Properties 3

1.1.3 Limitations and the Need for a New Framework 4

1.1.4 The Bayesian Nonparametric Framework 7

1.2 Background 8

1.2.1 Poisson Point Processes 8

1.2.2 Completely Random Measures 9

1.2.3 Caron-Fox model [Caron and Fox, 2017] 10

1.2.4 Generalised Gamma Process 12

1.2.5 Regular Variation and Power-Law 12

1.2.6 Vectors of Completely Random Measures 13

1.2.7 Hawkes Processes 13

1.3 Contributions and Thesis Outline 15

1.3.1 Modular Graphs with Overlapping Communities 15

1.3.2 Dynamic Interactions with Reciprocating Relationships 17

1.3.3 Useful Representations of Completely Random Measures . . . 19

2 Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities **21**

3 Modeling Sparsity, Heterogeneity, Reciprocity and Community Structure in Temporal Interaction Data **71**

4 A Unified Construction for Series Representations and Finite Approximations of Completely Random Measures **89**

5 Discussion **138**

5.1 Summary 138

5.2 Extensions 139

5.2.1 Sparse Networks with Core-Periphery Structure 139

5.2.2 Topological Covariates in the Link Probability 140

5.2.3 Sociability Parameters' Evolution in Time 140

5.2.4 Statistical Consistency 141

5.2.5 Assortativity and Disassortativity in Link Probability 142

5.2.6 Number of Communities 142

5.2.7 Generic Reciprocation Patterns 143

5.2.8 Other Applications 144

Bibliography **145**

Chapter 1

Introduction

This thesis follows an integrated format and contains five chapters, with the first chapter presenting the introduction, and the last chapter the discussion. The three intermediate chapters correspond to individual papers. The first paper has been accepted and will appear in the Journal of Royal Statistical Society: Series B (Statistical Methodology), the second paper has been published at the International Conference on Neural Information Processing Systems and the third paper has been submitted to the Bernoulli Journal and is under review. Each paper contains a sufficient review of background literature specific to the topics covered therein, whereas a more general introduction and overview is given in Chapter 1.

1.1 Motivation

1.1.1 Networks

Networks are complex data formulations that characterise multiple areas of modern society. They are ubiquitous in social sciences, the economy, medicine and engineering. Examples include friendship networks, collaboration or citation networks, business relation networks, biological networks or transportation networks. The large availability of network data and the widespread prevalence of networks that are usually very large in scale are key components of the 'Big Data' era we live in. Understanding and explaining these networks to address key questions impacting society has driven a growing interest in network analysis and methodology. Work in this field has flagged the need to develop new tools that reflect the characteristics and features of these 21^{st} century networks.

In network science, a network is defined by a set of nodes with connecting links

between them. Similarly, in graph theory a graph is defined by a set of vertices that connect to form edges. It is common to use the terms graph and network interchangeably as well as the terms links and edges and this convention also applies to this thesis.

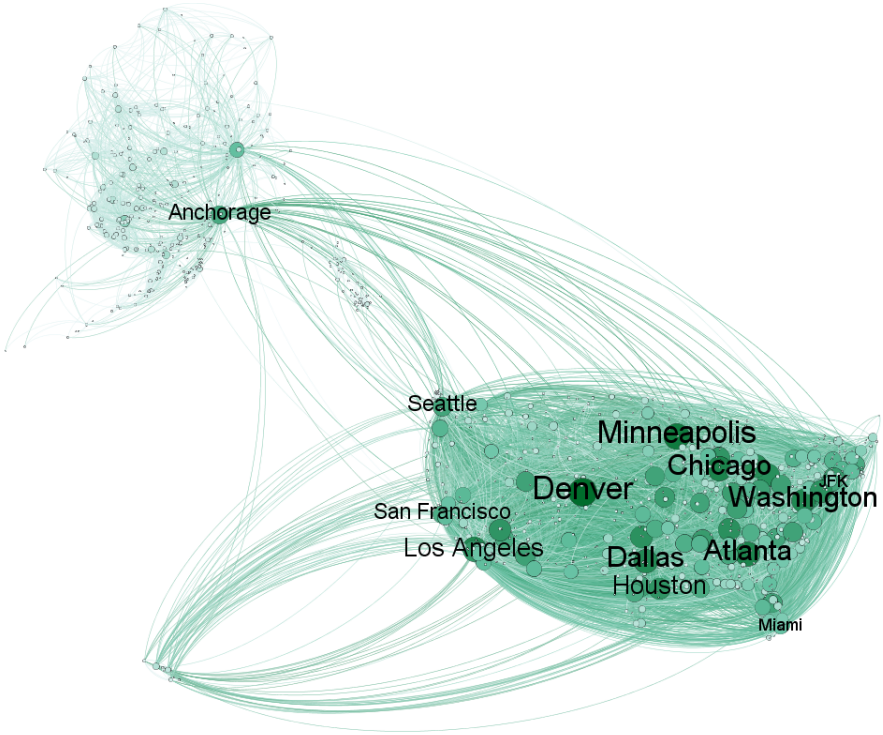


Figure 1.1: Network of connections between airports in the US in 2010 based on the USairport dataset provided by the Department of Transportation [U.S., 2010]. Figure is illustrated using the software Gephi [Bastian et al., 2009].

An example of a real-world network is shown in Figure 1.1. This network corresponds to connections between airports in the US in 2010 [U.S., 2010] and we will be referring to it as the USairport dataset. The network's nodes are the airports and the edges correspond to connections between the airports. Precisely, in this network there are 1574 nodes and 17215 edges. For two airports to be considered connected there should be at least one flight between them in either direction. This is an example of an undirected network, which precisely means that we do not take into account the direction of the edge, being in this case the direction of the flight. The airports are named according to the city they exist in, but note that for illustration purposes not all airport names are shown in Figure 1.1. Networks like the USairport network that are formed of undirected static connections will be referred to as static undirected

networks.

Another type of graph that will concern us is a graph consisting of temporal directed interaction data. In this graph type data come in the form of triples (i, j, t) meaning that there is an action from node i towards node j at time t. Some examples of such networks are social networks of online messages or people commenting on online blogs. We will consider a real example of a temporal network, named MathOverflow, taken from the Stanford Large Network Dataset Collection [Leskovec and Krevl, 2014]. This dataset consists of interactions on the stack exchange website MathOverflow over 2350 days. Its corresponding temporal directed graph consists of 24818 people, 239978 edges and 506550 interactions. A triple (i, j, t) means that at time t, a user i has responded or commented to a question or response of user j. We are interested in understanding the factors that underpin these interactions. For example, we know that a user is likely to imminently post online if they have recently received a comment from another user. The networks formed of temporal directed interactions will be referred to as temporal (or dynamic) directed networks.

1.1.2 Network Properties

In both types of networks presented above it is often observed that they exhibit some type of latent (or hidden) structure, as for example community structure. By network structure, we refer to the existence of latent blocks (also referred to as latent groups or communities) in the graph. In particular, we are interested in overlapping blocks in which case members belong to multiple latent groups in the network. In the example of the USairport dataset, the communities' formulation is based on geographical location. This means that the airports that are closer to each other are highly likely to be connected. What is often observed is that one of the groups corresponds to a hub; i.e. a central group consisting of important or major nodes. In the case of the USairport, the hub corresponds to a group of some major airports. Similarly, latent structure is observed in the temporal dataset of MathOverflow. There is ground truth information that there are three different types of user actions in the dataset which gives rise to three latent communities in the corresponding graph.

Another network property that has a leading role in the motivation and focus of this thesis, is network sparsity. An informal definition is that a graph is said to be sparse if only very few edges are observed with respect to the maximum possible

number of edges. This is observed in both examples; the USairport and the MathOverflow datasets. To formalise mathematically the concept of sparsity, we consider the rate of growth of the number of edges with respect to the number of nodes. Precisely, a graph is said to be dense if the rate of growth of the number of edges with respect to the number of nodes is quadratic, whereas a graph is sparse if this rate is subquadratic. We are interested in sparsity as it is often observed in large real-world networks [Newman, 2010]. However, it is highly non-trivial to have a statistical graph model that handles graphs across the whole range of dense to sparse.

Related to the graph sparsity property is the degree distribution behaviour. The degree of a node is the number of edges attached to it and the degree distribution is the distribution of all the nodes' degrees. A common feature of real-world networks mentioned above is the presence of hubs which means that some nodes have a lot of connections (making them major nodes), while most nodes have very few connections. The presence of hubs gives the degree distribution a long tail indicating the presence of nodes with a much higher degree than most other nodes. This behaviour is called degree heterogeneity. It is the scenario explained above in which in the same graph, some nodes have high degree whereas many others have degree one. This case is observed in both examples of the USairport and the MathOverflow datasets. In the former, we observe that some airports have a large number of flights whereas many others have very few. Similarly, in the MathOverflow example, some users respond many times whereas some others only once. The degree distribution of the two networks is shown in Figure 1.2 on a log-log scale. In both cases we observe that the logarithm of the degree distribution is roughly linear suggesting that informally both empirical degree distributions are 'close' to being power-law at least in some interval. Degree heterogeneity is closely related to power-law behaviour in the degree distribution of networks [Barabási and Albert, 1999; Bollobás, 2001; Newman, 2010; Berger et al., 2014]. Power-law behaviour is a useful property to assume as it comes with scale invariance. Note however that in our constructions power-law behaviour is only shown empirically.

1.1.3 Limitations and the Need for a New Framework

Traditionally, a network is represented by a random matrix (or random 2-array) known as the adjacency matrix. An appealing property to assume on random matrices is the so called exchangeability. A random graph represented by an adjacency matrix is exchangeable if the adjacency matrix is exchangeable. The notion of exchangeability

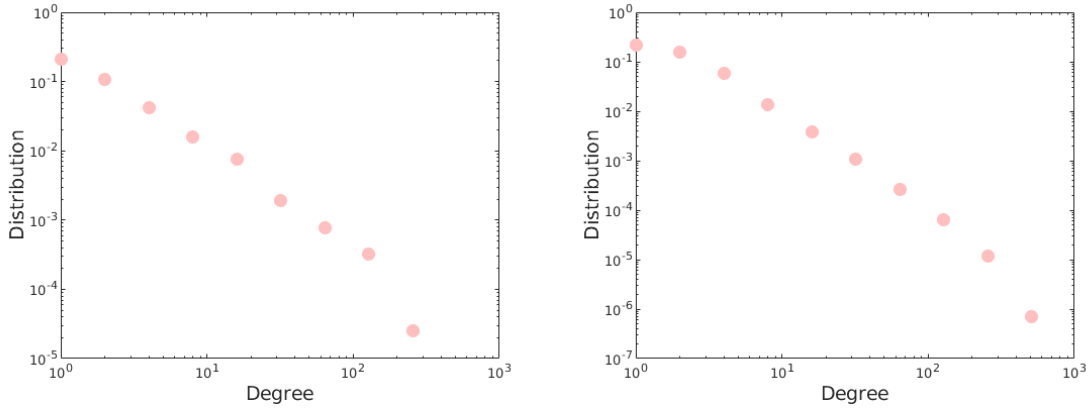


Figure 1.2: Empirical degree distribution on a log-log scale for the (Left): USairport dataset (Right): MathOverflow dataset.

on matrices was introduced by Aldous [1981, 1985] and Hoover [1979] and is given in Definition 1.

Definition 1 (Random 2-array Joint Exchangeability) *An infinite random binary matrix* $(A_{i,j})_{i,j\in\mathbb{N}}$ *is called jointly exchangeable if and only if for any* $n > 0$ *and any permutation* π *of* S_n

$$(A_{i,j})_{1\leq i,j\leq n} \stackrel{d}{=} \left(A_{\pi(i),\pi(j)}\right)_{1\leq i,j\leq n}$$

where S_n *is the symmetric group*[1] *of all permutations on the set* $\{1,\ldots,n\}$.

A graphical way to illustrate joint exchangeability on a binary symmetric adjacency matrix is shown in Figure 1.3. The interpretation of exchangeability is that it offers distributional invariance. Applying a permutation on the rows and the columns of the matrix, does not change its distribution. In other words, if we relabel every member of the network, the distribution of the random graph does not change which also means that the order with which the nodes appear is not relevant. The graph that is being represented by this adjacency matrix is structurally invariant. Structural invariance of a graph means that its realisation depends only on some structural features such as the number of edges or triangles, but not on their specific vertex locations.

Exchangeability of a binary random matrix, as defined in Definition 1, is accompanied by a representation theorem known as the Aldous-Hoover theorem [Aldous,

[1]The symmetric group on a finite set X is the group whose elements are all bijective functions from X to X and whose group operation is that of function composition.

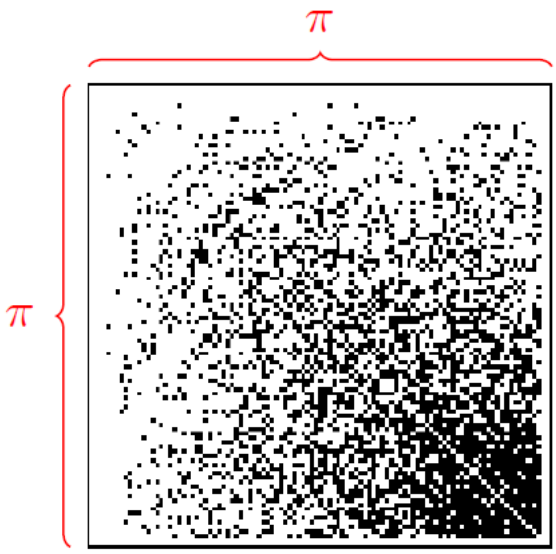


Figure 1.3: Joint exchangeability of a symmetric adjacency matrix by applying the same permutation π on rows and columns. Figure taken from Orbanz [2015b].

1981, 1985; Hoover, 1979] which provides a characterisation of such matrices. Precisely, the Aldous-Hoover theorem states that for any exchangeable matrix there exists a mixture model representation that is only based on transformed uniform random variables. A detailed exposition of exchangeability defined on several random structures is given by Orbanz and Roy [2015]. For the case of undirected acyclic[2] graphs the representation theorem simplifies and can be expressed in terms of a random measurable function $w : [0, 1]^2 \to [0, 1]$ called a graphon.

The Aldous-Hoover theorem encouraged the formulation of Bayesian models for arrays [Lloyd et al., 2012], a framework that consists of many statistical models for networks [Hoff et al., 2002; Nowicki and Snijders, 2001; Airoldi et al., 2008; Kim and Leskovec, 2012; Miller et al., 2009]. A crucial and undesirable consequence of the exchangeability property on a graph model, is that the Bayesian probabilistic version of that model is either trivially empty or dense. This is undesirable since real-world networks are often found to be sparse [Newman, 2010]. This limitation therefore raises the need for formulating new network methodology. Note here that one can sacrifice infinite exchangeability and consider one of the following two alternatives. The first is to resort to non-exchangeable generative models as for example the preferential attachment model [Barabási and Albert, 1999] that prescribes the generation of random scale-free networks using a preferential attachment mechanism. This model handles

[2]An acyclic graph is a graph with no cycles.

graphs in which there exist some nodes with an unusually high degree. The second alternative is to consider sequences of finitely exchangeable models [Chung and Lu, 2002], [Bickel and Chen, 2009], [Wolfe and Olhede, 2013].

It is evident that traditional network models do not allow the coexistence of infinite exchangeability and graph sparsity. The first model to capture both of these two properties is the Caron-Fox model, proposed by Caron and Fox [2017] who use an alternative construction with a different notion of exchangeability. Instead of representing the graph by a discrete adjacency matrix, they move from discrete space to continuous space and represent the graph by an exchangeable point process on the plane. This representation builds on completely random measures (CRMs) with random locations and random weights or equivalently, Poisson processes. A detailed explanation of CRMs and Poisson processes is given in Section 1.2. The Caron-Fox graph inherits structural invariance from the exchangeability of the underlying random measures that represent it. In a similar way that the Aldous-Hoover theorem characterises infinitely exchangeable matrices, a theorem provided by Kallenberg [1990] characterises the distribution of the infinitely exchangeable random measures. The resulting graphs represented by infinitely exchangeable random measures vary across the range of dense to sparse, exhibit power-law behaviour and degree heterogeneity. The definition of random measure infinite exchageability and more details on the Caron-Fox model are given in Section 1.2.3.

The Caron-Fox framework is flexible but relies on sampling CRMs which is computationally expensive. This thesis proposes principled methods to sample CRMs efficiently making the relevant models more practical to implement. Particularly, infinite-activity CRMs are appealing tools to use as they allow the representation of graphs that simultaneously have the desirable properties of sparsity and exchangeability. This is further explained in Section 1.2.3. The area of Statistics which deals with CRMs is called Bayesian Nonparametrics (BNP) which we motivate and explain in Section 1.1.4.

1.1.4 The Bayesian Nonparametric Framework

BNP is a branch of Statistics and Machine Learning that follows the Bayesian paradigm and assumes infinitely many parameters. The first component of BNP is the Bayesian paradigm which prescribes that all sources of uncertainty are treated probabilistically. We place some prior beliefs on the unknowns and after having seen new observations,

we update our beliefs to form posterior beliefs. In Bayesian modelling we place prior distributions on the model parameters and subsequently aim to approximate the posterior distributions using Bayesian inference methods. One of the benefits of Bayesian inference concerns uncertainty quantification. We represent our uncertainty about the unknown parameters by placing priors beliefs on them and then Bayesian inference gives us posterior beliefs on those unknown parameters in the form of probability distributions.

The second defining assumption of BNP is to be nonparametric which means that such models have infinitely many parameters. To understand the flexibility of the nonparametric models, first consider the opposite, namely parametric models. Parametric models consist of a fixed and finite set of parameters. Having a finite parameter set hinders parametric models from being arbitrarily flexible because the complexity and capacity of the model is limited. In contrast, nonparametric models, have an infinite in size parameter set which then adapts its effective complexity based on the complexity of the observed data.

1.2 Background

In this section we present some background material on the main tools and concepts that the following chapters build on. The key ingredients in all the papers presented in the thesis are CRMs which are presented and defined first. CRMs are directly connected to Poisson point processes which are subsequently explained. Next, we present the Caron-Fox model [Caron and Fox, 2017] on which the graph models proposed in this thesis build on and expand on. In our proposed multivariate extensions in Chapter 2 [Todeschini et al., 2019], we use a specific CRM, the Generalised Gamma Process (GGP) which we define in this Section. Moreover, we present material on inhomogeneous point processes and precisely on Hawkes processes which construct the model of temporal directed interactions presented in Chapter 3 [Miscouridou et al., 2018a]. For more details on random measures and stochastic processes we refer the reader to Daley and Vere-Jones [2008].

1.2.1 Poisson Point Processes

Definition 2 (Poisson Point Process) *[Kingman, 1967, Chapter 2] Let $\mathcal{B}$ denote the Borel σ-algebra of $\mathbb{R}_+^d$. A Poisson process on $\mathbb{R}_+^d$ is defined as the random countable subset Π of $\mathbb{R}_+^d$ such that*

1. *for $A \in \mathcal{B}$, the count measure defined as $N(A) := \#\{\Pi \cap A\}$ follows a Poisson distribution with rate $\nu(A)$ where ν is a non atomic measure on $\mathcal{B}$ and*

2. *for any disjoint subsets $A_1, \ldots, A_n \in \mathcal{B}$, the random variables $N(A_1), \ldots, N(A_n)$ are independent.*

We call ν the mean measure of the Poisson process Π.

1.2.2 Completely Random Measures

Definition 3 (Completely Random Measure) *[Daley and Vere-Jones, 2008, Chapter 10] Let $\mathcal{B}$ denote the Borel σ-algebra of $\mathbb{R}_+$. A random measure W on $\mathbb{R}_+$ is said to be completely random if for any countable number of disjoint measurable sets $A_1, A_2, \cdots \in \mathcal{B}$, the random variables $W(A_1), W(A_2), \ldots$ are independent.*

Note that there is a direct relationship between CRMs and Poisson processes. In particular, a CRM admits a representation theorem (see Theorem 4) which states that any CRM can be decomposed into a sum of three independent components, one of which describes a Poisson point process.

Theorem 4 *(Kingman [1967]) Any completely random measure W on $\mathbb{R}_+$ can be uniquely represented in the form*

$$W(A) = \sum_{k=1}^{\infty} U_k \delta_{\theta_k}(A) + \xi(A) + \int_0^{\infty} wN(A \times dw), \tag{1.1}$$

for every Borel measurable set A, where the sequence $\{\theta_k\}$ enumerates a countable set of fixed atom locations of the measure W the sequence $\{U_k\}$ is a sequence of mutually independent nonnegative random variables determining the masses at these locations. The measure $\xi(\cdot)$ is a fixed non-atomic boundedly finite measure on $\mathbb{R}_+$, and $N(\cdot)$ is a Poisson process on $\mathbb{R}_+^2$, independent of $\{U_k\}$. The mean measure of the Poisson process, denoted by ν may be unbounded on sets of the form $A \times (0, \infty)$ but satisfies

$$\int_{w>\epsilon} \nu(A \times dw) < \infty \quad \text{and} \quad \int_{0<w<\epsilon} w\, \nu(A \times dw) < \infty$$

for any Borel measurable set A, any $0 < \epsilon < \infty$ and for all $x \in \mathbb{R}_+$ such that

$$\nu\left(\{x\} \times (0, \infty)\right) = 0.$$

We refer the reader to Kingman [1967] and Daley and Vere-Jones [2008] for more details. Following Caron and Fox [2017] we consider in our constructions a realisation of a CRM that admits a representation equal to the third term of the sum in Eq. (1.1) and can be seen as a realisation of a Poisson process on $\mathbb{R}_+^2$ with mean measure (or Lévy measure) denoted by ν. Then, such a CRM denoted by W can be written in the following form

$$W = \sum_{i=1}^{\infty} w_i \delta_{\theta_i},$$

and its distribution is characterised by the distribution of the random point set $(w_i, \theta_i)_{i=1,2,\ldots}$ of atoms at locations θ_i with weights w_i. Assuming a homogeneous process with independent increments we write the measure ν as

$$\nu(dw, d\theta) = \rho(dw)\lambda(d\theta)$$

where ρ is a measure on $\mathbb{R}_+$ and λ is the Lebesgue measure on $\mathbb{R}_+$. We write $W \sim CRM(\rho, \lambda)$. For ρ the assumption we assume throughout is that it satisfies the following condition

$$\int_0^{\infty} \left(1 - e^{-w}\right) \rho(dw) < \infty.$$

1.2.3 Caron-Fox model [Caron and Fox, 2017]

Caron and Fox [2017] propose a novel approach to network modelling. They represent a graph using a point process which means that the graph construction is based on CRMs (see definition in Section 1.2.2). They consider a CRM as defined in Section 1.2.2 and define the mean measure of the CRM as $\nu(dw, d\theta) = \rho(dw)\lambda(d\theta)$, where λ is taken again to be the Lebesgue measure on $\mathbb{R}_+$.

To represent the network, they draw a realisation $W \sim \text{CRM}(\rho, \lambda)$ which admits the form $W = \sum_{i=1}^{\infty} w_i \delta_{\theta_i}$. The θ_i's correspond to the nodes' locations and the parameters w_i are associated to the nodes' weight parameters. For undirected binary graphs, a connection exists between two nodes i and j if there is a point at locations (θ_i, θ_j) and (θ_j, θ_i). An undirected binary graph is therefore represented by a symmetric point process Z on the plane

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

where $z_{ij} = z_{ji} = 1$ if i and j are connected and 0 otherwise. The novelty of the Caron-Fox model is their suggestion to replace the traditional approach of using a

discrete random 2-array (adjacency matrix) to represent the graph by a symmetric point process on the plane. In Section 1.1.3 we have shown the definition of infinite exchangeability in the discrete framework of random matrices. In this Section we move to the continuous framework and define exchangeability on point processes. A point process Z is jointly exchangeable if and only if

$$\left(Z\left(B_i \times B_j\right)\right) \stackrel{d}{=} \left(Z\left(B_{\pi(i)} \times B_{\pi(j)}\right)\right), \text{ for } (i,j) \in \mathbb{N}^2, \tag{1.2}$$

for any permutation π of $\mathbb{N}$ and intervals $B_i = [h(i-1), hi], i \in \mathbb{N}, h > 0$. Kallenberg [2005], provides a characterisation theorem for exchangeable random measures on the plane and in this way gives direct analogies to the classical representation of graphs based on exchangeability of the adjacency matrix. A point process that satisfies Eq. (1.2) admits a representation theorem due to Kallenberg [1990] which is presented in Caron and Fox [2017]. The representation theorem provides a general construction for exchangeable random measures and therefore random graphs that are represented by these measures.

Having specified a class of priors that determines the graph (and the graph parameters) the next step is to define the link probability of connection of two nodes based on their parameters. Caron and Fox [2017] assume that two nodes $i \neq j$ connect with probability

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

where the $(w_i, \theta_i)_{i=1,2,\ldots}$ are the points of a Poisson point process on $\mathbb{R}_+^2$ and $z_{ij} \in \{0,1\}$ denotes the existence of a link (or not) between nodes i and j. The parameters $w_i > 0$ can be interpreted as sociability parameters since the higher the values of two sociability parameters w_i and w_j are, the higher the probability is that they are connected. In Caron and Fox [2017], the authors explain that the sparsity of the resulting graph depends on the type of the mean measure of the Poisson process and precisely on the jump part ρ of the mean measure. If ρ satisfies the condition

$$\int_0^\infty \rho(dw) = \infty, \tag{1.3}$$

then the CRM is called a CRM of infinite-activity. On the other hand, if the integral in Eq. (1.3) is finite, then the CRM is called of finite-activity. Caron and Fox [2017] show that if the CRM is finite-activity then the graph is dense (i.e. number of edges scales quadratically with the number of nodes asymptotically), whereas if the CRM is infinite-activity then the graph is sparse (i.e. number of edges scales subquadratically with the number of nodes asymptotically).

1.2.4 Generalised Gamma Process

In all networks considered below, we take our CRM to be the GGP which has Lévy intensity ρ defined as below,

$$\rho(dw) = \frac{w^{-1-\sigma}}{\Gamma(1-\sigma)} e^{-\tau w} dw$$

where the two parameters (σ, τ) verify

$$(\sigma, \tau) \in (-\infty, 0] \times (0, \infty) \text{ or } (\sigma, \tau) \in (0, 1) \times [0, \infty).$$

The GGP measure is appealing because of its flexibility, directly interpretable parameters with respect to the graph properties and simple form of the log density. In Caron and Fox [2017] it is shown that the infinite or finite-activity property of the CRM which determines the global graph sparsity is simply tuned by the parameter σ. If $\sigma < 0$ the CRM is of finite-activity and the resulting graph is dense, whereas if $\sigma \geq 0$, the corresponding CRM is of infinite-activity and the resulting graph is sparse with precise rate of $N_e = \Theta\left(N^{\frac{2}{1+\sigma}}\right)$ for $\sigma \in (0, 1)$ as suggested in Caron and Rousseau [2017]. The degree distribution of the resulting graph has a power-law behaviour with the parameter τ tuning the exponential decay of the tails of the power-law degree distribution.

1.2.5 Regular Variation and Power-Law

The GGP measure belongs to the family of regularly varying CRMs which is the family of CRMs of interest capturing the desirable graph properties. Regular variation is a property that concerns the tail of the Lévy intensity ρ. Consider $W \sim \text{CRM}(\rho, \lambda)$. Then the tail Lévy intensity denoted by $\bar{\rho}(x)$ is defined by

$$\bar{\rho}(x) = \int_x^\infty \rho(w) dw$$

for any $x > 0$.

Definition 5 *(Regular Variation) Let $W \sim$ CRM(ρ, λ). The CRM is said to be regularly varying if the tail Lévy intensity verifies*

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

for some $\sigma \in [0, 1]$ where ℓ is a slowly varying function verifying $\lim_{t\to\infty} \ell(ct)/\ell(t) = 1$ for any $c > 0$.

The precise way that regular variation links to the properties of sparsity is shown in Theorem 5 in Caron and Fox [2017].

1.2.6 Vectors of Completely Random Measures

In order to achieve the first thesis objective which is to add structure to the Caron and Fox [2017] model we consider communities in the network's latent structure. These are captured by vectors of affiliations as opposed to scalar sociability parameters used in Caron and Fox [2017]. The vectors of affiliations require multivariate extensions of CRMs, namely vectors of CRMs. Various approaches for vectors of CRMs have recently been proposed by various authors [Epifani and Lijoi, 2010; Leisen and Lijoi, 2011; Leisen et al., 2013; Griffin et al., 2013; Lijoi et al., 2014]. The definition of a vector of CRMs is given below.

Definition 6 *Let $\mathcal{B}$ denote the Borel σ-algebra of $\mathbb{R}_+$. A vector of CRMs $(W_1, \ldots, W_p)$ on $\mathbb{R}_+$ is a collection of random measures W_k, $k = 1, \ldots, p$ such that for any collection of disjoint measurable subsets $A_1, \ldots, A_n$ of $\mathcal{B}$ the vectors*

$$(W_1(A_1), \ldots, W_p(A_1)), (W_1(A_2), \ldots, W_p(A_2)), \ldots, (W_1(A_n), \ldots, W_p(A_n))$$

are mutually independent.

In all our constructions the vectors of CRMs we use are the so called compound CRMs proposed in Griffin and Leisen [2017]. Compound CRMs are dependent random measures which can be constructed from a GGP and are appealing because of their computational tractability. For a detailed exposition on compound CRMs and their properties we refer the reader to Griffin and Leisen [2017].

1.2.7 Hawkes Processes

Important building blocks of our network constructions in Chapter 3 are inhomogeneous point processes. Precisely we are interested in Hawkes processes introduced by Hawkes as self-exciting [Hawkes, 1971b] and mutually-exciting processes [Hawkes, 1971a]. Considering a single Hawkes process only it can only be self-exciting and its inhomogeneity emerges from the fact that the current rate of events depends on the whole history of the process.

Let $(t_k)_{k\geq 1}$ be a sequence of event times with $t_k \geq 0$, and let $\mathcal{H}_t = (t_k | t_k \leq t)$ be the subset of event times between time 0 and time t. Let $N(t) = \sum_{k\geq 1} \mathbf{1}_{t_k \leq t}$ denote the number of events between time 0 and time t, where $\mathbf{1}_A = 1$ if A is true, and 0 otherwise. Assume that $N(t)$ is a counting process with conditional intensity

function $\lambda(t)$, which means that for any $t \geq 0$ and any infinitesimal interval dt the following equation holds.

$$\Pr\left(N\left(t+dt\right)-N\left(t\right)=1|\mathcal{H}_t\right)=\lambda\left(t\right)dt. \tag{1.4}$$

Eq. (1.4) defines a point process with inhomogeneous intensity which is self-exciting. By self-exciting we mean that the probability of another point occurring in the interval $(t, t+dt)$, depends on the past events of the same process. This probability is captured by the intensity function $\lambda(t)$ shown in Eq. (1.5) below. There are various forms the intensity can take but in our model we focus on the following

$$\lambda(t) = \mu + \int_0^t g(t-u)\, dN(u), \tag{1.5}$$

where $\mu = \lambda(0) > 0$ is a constant scalar and g is a nonnegative function, referred to as the kernel which satisfies $g(s) \geq 0$ for $s > 0$ and $g(s) = 0$ for $s < 0$. In this thesis however, we do not use self-exciting processes but mutually-exciting processes that come in the form of pairs. Each pair consists of two dependent point processes which mutually excite each other through the form of their intensities. Note that for the purposes of this thesis, by mutual-excitation we mean that each process only affects the other one and there is no additional self-excitation effect. In all cases, the pattern of excitation is captured through the intensity since this is the part of the process that determines the occurrence of a future event in that process. For the mutually-exciting process of interest, the intensity of each process depends on historic events of the other process.

Consider a counting process denoted by $N(t)$ with intensity $\lambda(t)$ and similarly another counting process $\tilde{N}(t)$ with corresponding intensity $\tilde{\lambda}(t)$. Then, $N(t)$ and $\tilde{N}(t)$ form a pair of mutually-exciting Hawkes processes [Hawkes, 1971a] if the conditional intensity functions $\lambda(t)$ and $\tilde{\lambda}(t)$ take the form

$$\lambda(t) = \mu + \int_0^t g(t-u)\, d\tilde{N}(u) \qquad \tilde{\lambda}(t) = \tilde{\mu} + \int_0^t \tilde{g}(t-u)\, dN(u) \tag{1.6}$$

where $\mu = \lambda(0) > 0$, $\tilde{\mu} = \tilde{\lambda}(0) > 0$ are non time dependent terms and $g, \tilde{g}$ are non-negative time dependent kernel functions. Eq. (1.6) defines the rates of a pair of mutually-exciting processes in which the current rate of events of each process depends on the occurrence of past events of the opposite process.

1.3 Contributions and Thesis Outline

The contributions of the thesis are given in Chapters 2, 3 and 4. These chapters include a paper each in its exact complete form, including the main paper, the bibliography and the supplementary material. This Section includes the objective, the methodology, significance and my personal contributions for each chapter.

1.3.1 Modular Graphs with Overlapping Communities

Problem and Objective

Chapter 2 consists of the paper Todeschini et al. [2019] in which we generalise the novel graph model of Caron and Fox [2017] and expand on it. The graph model of Caron and Fox [2017] captures graphs that are sparse and whose degree distribution is heterogeneous and power-law. However, their graph models have no network structure. In Todeschini et al. [2019] we address this by adding to the graph structure in the form of overlapping blocks. These are latent network groups to which the members of the network can have multiple affiliations.

Methodology

The way to achieve this objective is through the use of vectors of CRMs (see Section 1.2.6) instead of scalar CRMs and in this way we generalise the Caron-Fox model from the univariate case to the multivariate case. Our construction has interpretable parameters and each node is assigned a vector representing the levels of its affiliations to the latent communities. These vectors are taken as compound CRMs following Griffin and Leisen [2017]. For inference we implement an MCMC sampler.

Significance

The first contribution of our model is that it offers an expansion of the novel Caron-Fox model to the multivariate setting. Furthermore, our model is a sparse version of the stochastic block model and in this way generalises existing overlapping block graph models to the sparse regime. The contribution of this work to the relevant methodology is significant and is reflected in its venue of publication. The graph extension is accompanied by the computational complexity of dealing with vectors of CRMs and particularly compound CRMs. We propose an inference scheme on compound CRMs that relies on data augmentation and posterior characterisation. We implement an MCMC sampling algorithm whose complexity is linear per community,

node and edge. Our work has practical significance as well since the model and methods can be used to answer specific questions of interest arising from network datasets. For instance, we use our methods to uncover the latent structure of real-world networks in a political context and a social transportation context. We perform community detection and in the case of USairport we identify the existence of a hub. Aiming in reproducible research and maximum impact, we develop a software package with the code to perform graph generation and posterior inference which is publicly available.

Contribution

This is joint work of Adrien Todeschini, François Caron and myself. I have contributed in theoretical, methodological and experimental tasks. Regarding theory I have contributed in deriving the exact and asymptotic properties of the graph. Regarding methodology I have contributed in assessing the methodology and in particular the Gaussian approximation component. On experiments, I implemented extensive comparisons of our proposed approach to two alternative network models [Hoff et al., 2002; Airoldi et al., 2008]. The experiments were performed on two different real-world networks and the objective was to carry out a fair comparison between the models. The aspects of comparison were the degree distribution fit and diagnostics, assessment of sparsity and recovery of communities compared to the ground truth. Finally, I have contributed in the development of the software package that accompanies the model which is available online.

The paper and the software are referenced as below.

- A. Todeschini, **X. Miscouridou**, and F. Caron. Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019, (to appear).

- A. Todeschini, **X. Miscouridou**, and F. Caron. SNetOC: Matlab software package for Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities, *https://github.com/misxenia/SNetOC*, 2017.

1.3.2 Dynamic Interactions with Reciprocating Relationships

Problem and Objective

The second major contribution of the thesis, presented in Chapter 3, is a novel model for understanding and predicting temporal interaction data. Such data appear as time-stamped interactions represented in the form (i, j, t) indicating that an event was initiated from a sender i towards a receiver j at time t. Our objective is to understand the underlying factors that underpin these interactions. The properties which are observed in real social interaction data are community structure, reciprocity, sparsity and degree heterogeneity. Community structure affects the interactions in the sense that individuals with similar community affiliations are highly likely to exchange interactions. Reciprocity (or reciprocation) concerns the probability of the return of an action, in the sense that many events are responses to recent events of the opposite direction. Some common examples are the MathOverflow dataset of online posts presented before and message or email datasets. In an email dataset if i sends an email to j, then this increases the chances of another email to be sent from j to i. By collecting all triples of the form (i, j, t), we can formulate the corresponding network of connections in which exhibits the properties of sparsity and degree heterogeneity explained before.

Methodology

The way we achieved our objective is by developing a novel class of statistical models for temporal interaction data that can capture all the desired properties mentioned above. The construction builds on mutually-exciting Hawkes processes [Hawkes, 1971a] and the static undirected model for sparse and modular graphs with overlapping community structure [Todeschini et al., 2019]. The interactions between each directed pair of individuals are modelled using mutually-exciting Hawkes processes in which the component that captures reciprocation is the intensity function. The intensity has interpretable parameters that directly tune the properties of interest.

Significance

Combining all the properties in a single model is non-trivial and there is no such construction to our knowledge. The proposed model generalises existing reciprocating relationship models [Blundell et al., 2012] to the sparse and power-law regime. Additionally, our model unifies literature as it is a natural extension of two recently

introduced graph model classes that capture sparsity and power-law degree distribution in network data. The first class is the class of vertex-exchangeable models whose construction is based on exchangeable random measures [Caron and Fox, 2017; Veitch and Roy, 2015; Herlau et al., 2016; Borgs et al., 2018; Todeschini et al., 2019; Palla et al., 2016; Janson, 2017b]. The second class consists of edge-exchangeable models [Crane and Dempsey, 2015, 2016; Cai et al., 2016; Williamson, 2016; Janson, 2017a; Ng and Silva, 2017]. We carry out experiments on a variety of temporal real network datasets for interaction data and show that the proposed model outperforms competing models in link prediction, whilst importantly keeping interpretable parameters.

Contribution

This is joint work of François Caron, Yee Whye Teh and myself. I have been the main contributor of the paper. Regarding theory I have formulated the model construction based on mutually-exciting processes and derived the exact and asymptotic properties of the graph class across two axes of directions. The first axis is static and concerns the graph growth in size, i.e. the number of nodes and edges of the undirected graph. The second direction is temporal and concerns new interactions between existing pairs of nodes formed in time. On inference, I have proposed and implemented a two-stage approximate Bayesian inference, a type of modular Bayesian inference which directly reflects the two natures of the model; static and dynamic. I have performed simulation experiments to verify the validity of the inference procedure in recovering the same parameter values used in simulation. I then implemented experiments to compare our model to competing machine learning models. The experiments concerned uncovering the underlying graph structure of several real temporal datasets and predicting interactions in a future time window.

The paper is published as below.

- **X. Miscouridou**, F. Caron and Y. W. Teh, Modelling Sparsity, Heterogeneity, Reciprocity and Community Structure. *Proceedings of the International Conference on Neural Information Processing Systems*, 2018.

Our code is available and can be found at `https://github.com/OxCSML-BayesNP/HawkesNetOC`.

1.3.3 Useful Representations of Completely Random Measures

Problem and Objective

Our network model contributions outlined in Sections 1.3.1 and 1.3.2 build on CRMs of infinite-activity. The infinite-activity property is the key property allowing for sparsity and degree heterogeneity in both model classes outlined above. Beyond network applications, models with infinite-dimensional parameters have been used in various other BNP applications such as density estimation or clustering. One of the most popular infinite-activity CRMs which is also heavily used in Chapters 2 and 3 is the GGP. Another popular infinite-activity measure is the (stable) beta process. Apart from special cases, exact simulation and scalable inference with these infinite capacity models is challenging. Several research papers have overcome this limitation by considering finite truncations which are often crude and lack error quantification. In this work, we propose a generic, unified and principled framework to derive finite-dimensional approximations of CRMs.

Methodology

The finite-dimensional approximations we propose are obtained by defining a Poisson random measure on an extended space. We propose two forms of truncations. The first one is a sequential construction that considers the truncation of an (infinite) series representation of the CRM with stochastically decreasing weights. The second one is an iid construction and considers a measure with $n < \infty$ atoms and iid weights, converging in distribution to the CRM as n tends to infinity. Very importantly, we quantify the error associated to the truncation.

Significance

The contribution of our work to the relevant field is highly significant. Our framework is generic and expands on the constructions based on size-biased sampling of Poisson point processes [Perman et al., 1992]. Our framework contributes to literature in a unifying way, since we draw connections between our suggestions and existing series and iid representations. We propose novel examples and also include several known series representations and existing finite approximations as special cases of our framework. Additionally, these new constructions can be used to derive novel algorithms for posterior inference based either on MCMC or variational inference. In our work

we focus on MCMC inference and generalise the slice sampler for normalised CRMs mixture models of Griffin and Walker [2011].

Contribution

This is joint work of Juho Lee, François Caron and myself. I have contributed in the conception of the idea and formulating the motivation of this work. I have worked in the formulation of our constructions of finite approximations of CRMs; namely sequential and iid representations. I have proposed and derived specific examples of the framework and I have worked on the properties of the methods and in quantifying the error of the truncation.

The paper is submitted to the Bernoulli Journal and its current version on arXiv is given below.

- J. Lee, **X. Miscouridou**, and F. Caron. A unified construction for series representations and finite approximations of completely random measures. *arXiv preprint arXiv:1905.10733*, 2019.

Aside Note

Another paper (Miscouridou et al. [2018b]) that has been carried out as part of the research exchange during my DPhil but is not included in the thesis is shown below. The paper has been accepted in the Machine Learning for Healthcare Conference, 2018 and is published in the Proceedings of Machine Learning Research.

- **X. Miscouridou**, A. Perotte, N. Elhadad and R. Ranganath, Deep Survival Analysis: Nonparametrics and Missingness, *Third Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages: 244-256, 2018.

Chapter 2

Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities

Accepted for publication in the Journal of Royal Statistical Society: Series B (Statistical Methodology)

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

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

where the $(w_i, \theta_i)_{i=1,2,\dots}$ 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 $c_i \in \{1, \dots, p\}$, two nodes $i \neq j$ connect with probability

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

where the $(w_i, c_i, \theta_i)_{i=1,2,\dots}$ are the points of a (marked) Poisson point process on $\mathbb{R}_+ \times \{1, \dots, p\} \times \mathbb{R}_+$ 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}, $k = 1, \dots, p$, and that the probability that two nodes $i \neq j$ connect is given by

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

These non-negative weights can be interpreted as measuring the level of affiliation of node i to the latent communities $k = 1, \dots, p$. 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 $(w_{i1}, \dots, w_{ip}, \theta_i)_{i=1,2,\dots}$ 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:

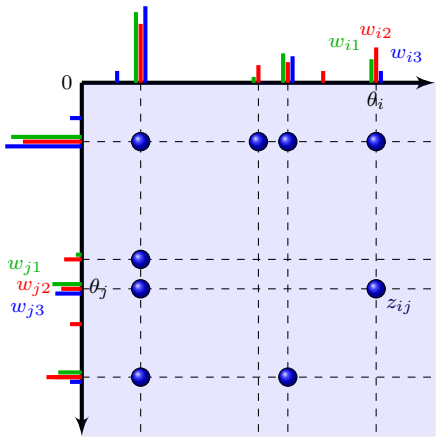


FIG 1. *Representation of a undirected graph via a point process Z. Each node i is embedded in $\mathbb{R}_+$ at some location θ_i and is associated with a set of positive attributes $(w_{i1}, \ldots, w_{ip})$. An edge between nodes θ_i and θ_j is represented by a point at locations (θ_i, θ_j) and (θ_j,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

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

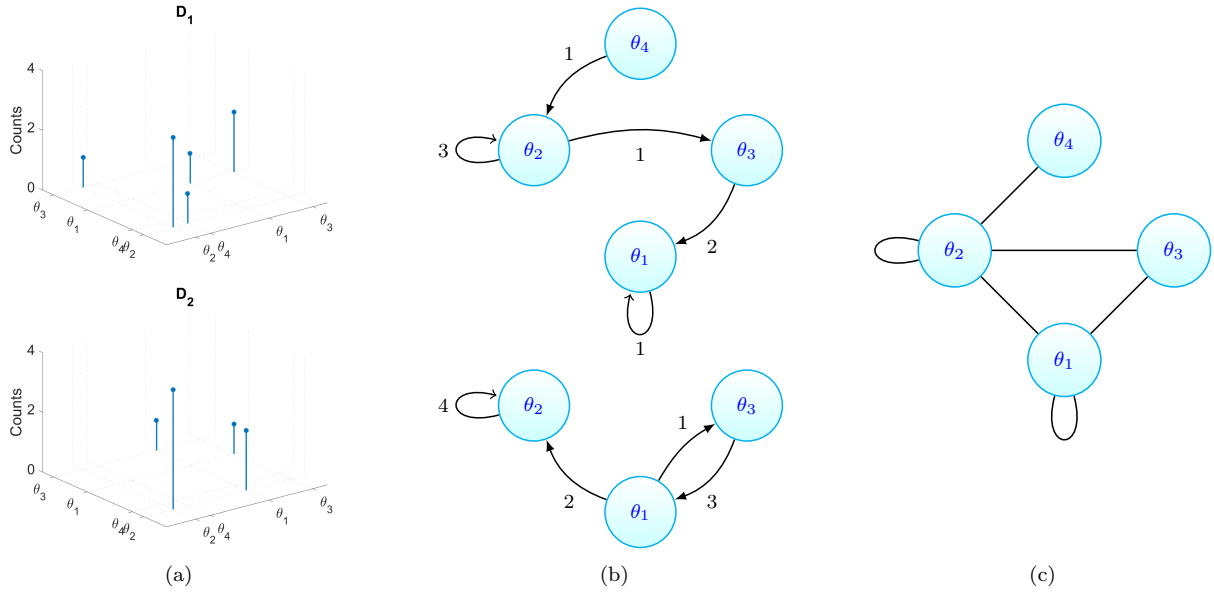


FIG 2. *An example of (a) the restriction on* $[0,1]^2$ *of the two atomic measures* D_1 *and* D_2*, (b) the corresponding multiview directed multigraphs (top: view 1; bottom: view 2) and (c) corresponding undirected graph.*

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 $\mathbb{R}_+^2$, $k = 1, \ldots, p$,

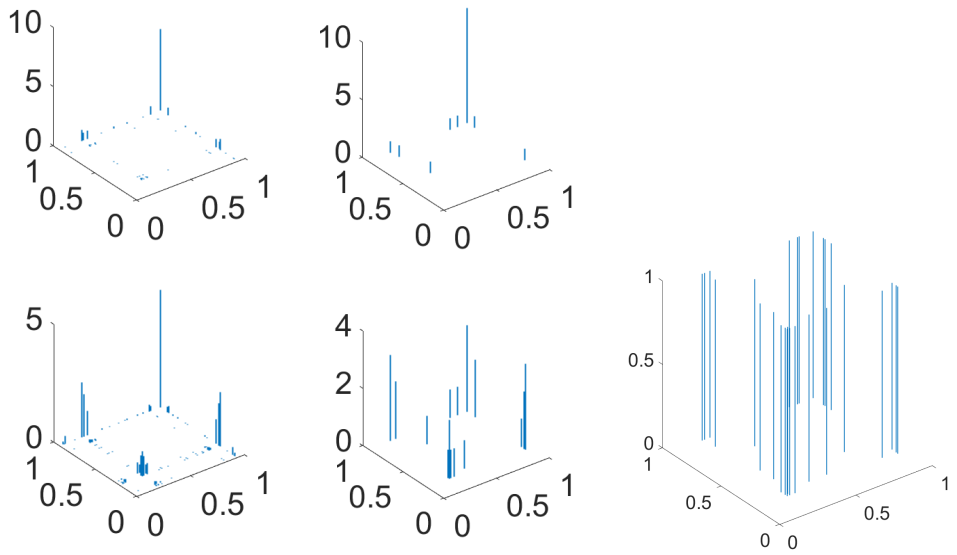
$$D_k = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} n_{ijk} \delta_{(\theta_i, \theta_j)}, \tag{9}$$

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 $W_k \times W_k$

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

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).

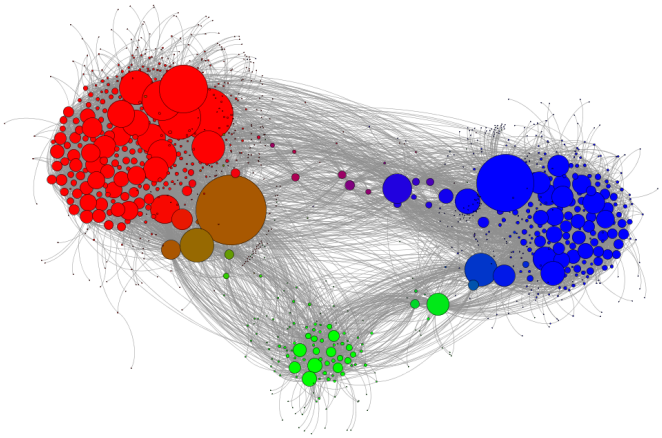


FIG 4. *Graph sampled from the model with three latent communities, identified by colors red, green, blue. For each node, the intensity of each color is proportional to the value of the associated weight in that community. Pure red/green/blue color indicates the node is only strongly affiliated to a 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 CRMs (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 GGP 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

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

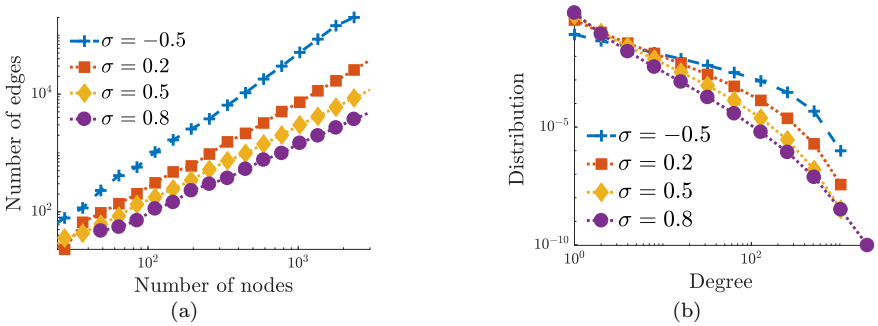


FIG 5. *Empirical analysis of the properties of CCRM based graphs generated with parameters* $p = 2$, $\tau = 1$, $a_k = 0.2$, $b_k = \frac{1}{p}$ *and averaging over various* α. *(a) Number of edges versus the number of nodes and (b) degree distributions on a log-log scale for various* σ: *one finite-activity 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 $D_{k,\alpha}^\varepsilon = \sum_{\ell=1}^{D_{k,\alpha}^*} \delta_{U_{k\ell 1}, U_{k\ell 2}}$.
5. Obtain Z from $(D_1, \ldots, D_p)$ as in (11).

Construction based on CCRMs. The hierarchical construction of CCRMs 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 GGP 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 CCRMs.

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: CCRMs

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

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

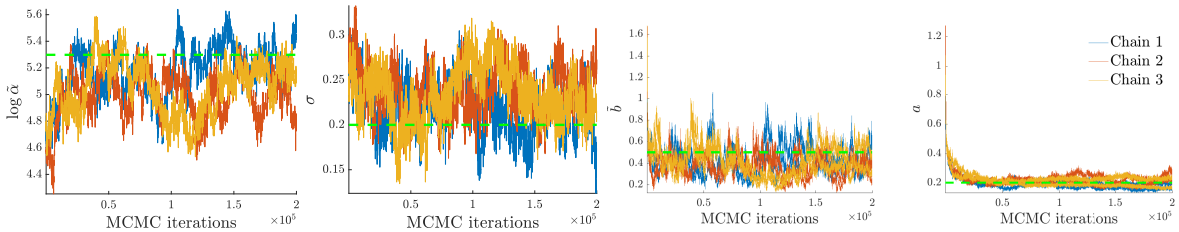
$$\Sigma_\varepsilon = \alpha \int_{\mathbb{R}_+^p} w_{1:p} w_{1:p}^T \rho_\varepsilon(dw_1, \ldots, dw_p)$$

where

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

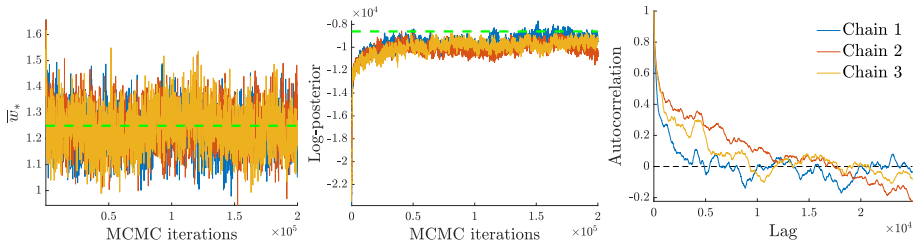
Note that μ_ε and Σ_ε can both be expressed as one-dimensional integrals using the gradient and Hessian of the moment generating function M of F. Theorem S1 in Section H of the supplementary material, which is an adaptation of the results of Cohen and Rosinski (2007) to CCRMs, gives the conditions on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

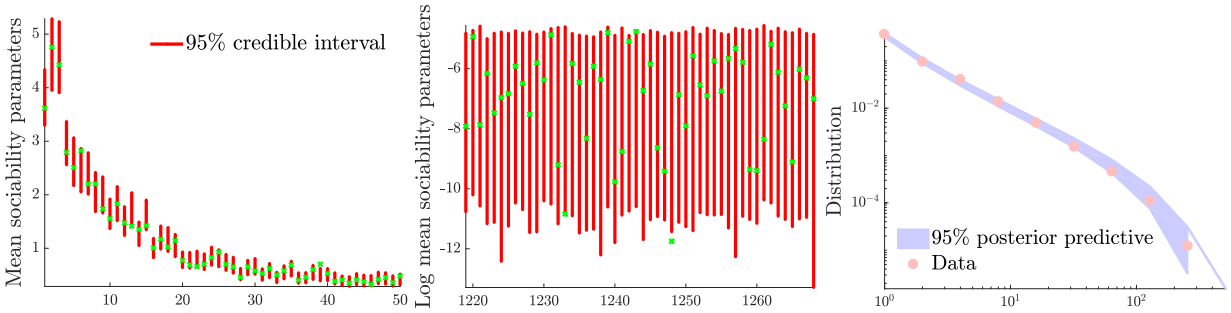
and thus the approximation is asymptotically valid. The Gaussian approximation is asymptotically valid for the CCRM defined by (17) and (18) when $\sigma \in (0, 1)$, hence is valid for all infinite-activity cases except for $\sigma = 0$.

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



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

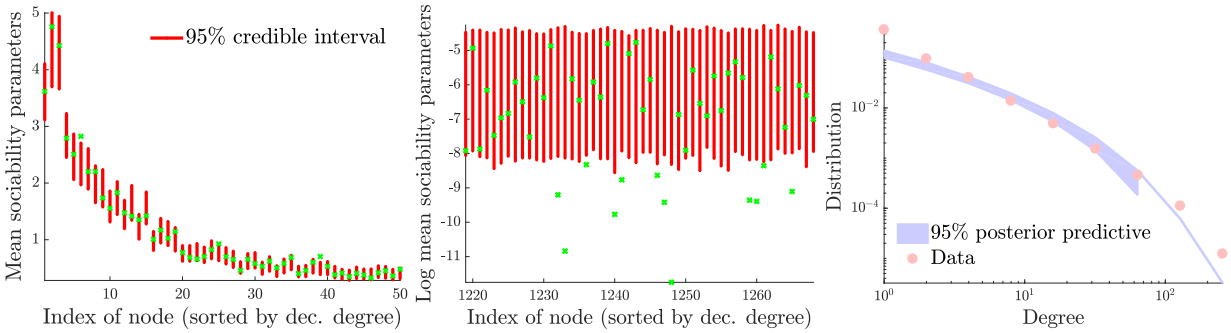
FIG 8. *95% posterior credible intervals (red) and true values (green) of (a) the mean parameters* $\overline{w}_i = \frac{1}{p}\sum_{k=1}^{p} w_{ik}$ *of the 50 nodes with highest degrees and (b) the log mean parameters* $\log \overline{w}_i$ *of the 50 nodes with lowest degrees. (c) Empirical degree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

for generating the graph. Figure 7(b) and (c) show the trace and autocorrelation plots of the log-posterior 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

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

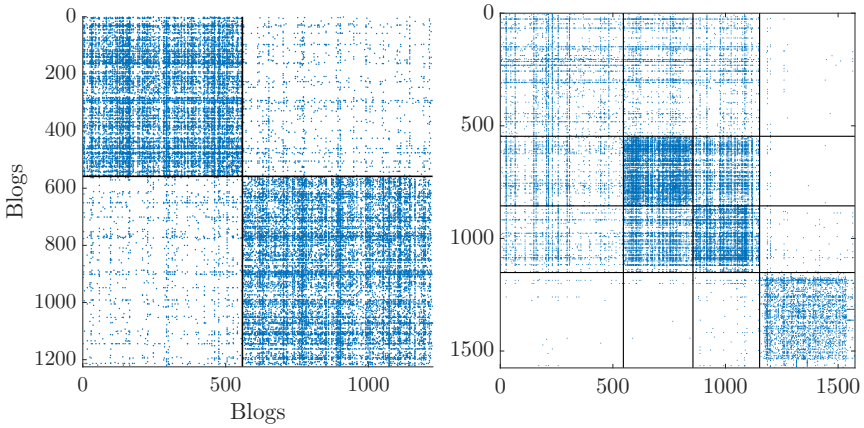
Its evaluation requires solving a combinatorial optimization problem in $O(p^3)$ using the Hungarian method. We therefore want to solve

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

[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) shows 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.*

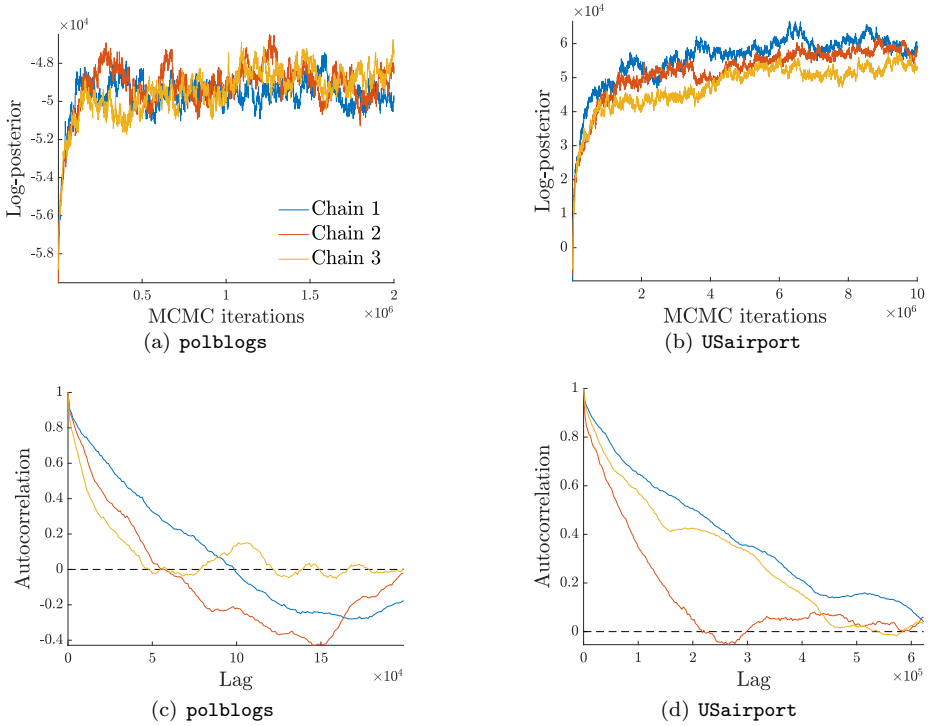
Community 1: "Hub"	**Community 2:** "East"	**Community 3:** "West"	**Community 4:** "Alaska"
New York, NY	Detroit, MI	Denver, CO	Anchorage, AK
Miami, FL	Chicago, IL	Las Vegas, NV	Fairbanks, AK
Boston, MA	Atlanta, GA	Burbank, CA	Bethel, AK
Washington, DC	Cleveland, OH	Salt Lake City, UT	Unalakleet, AK
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`.

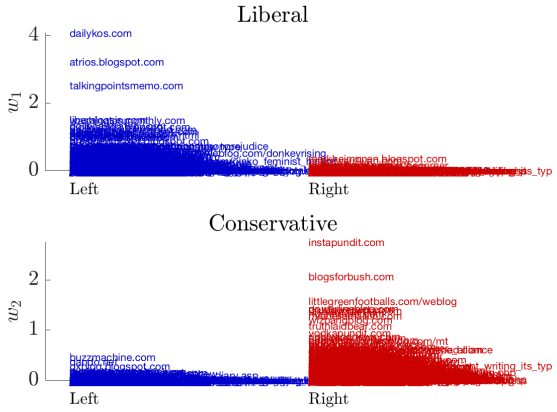


FIG 12. *Level of affiliation of each blog of the* `polblogs` *network to the communities identified as "Liberal" (top) and "Conservative" (bottom). The names of the blogs are grouped according to the left-right wing ground truth. Left-wing blogs are represented in blue on the left, right-wing blogs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

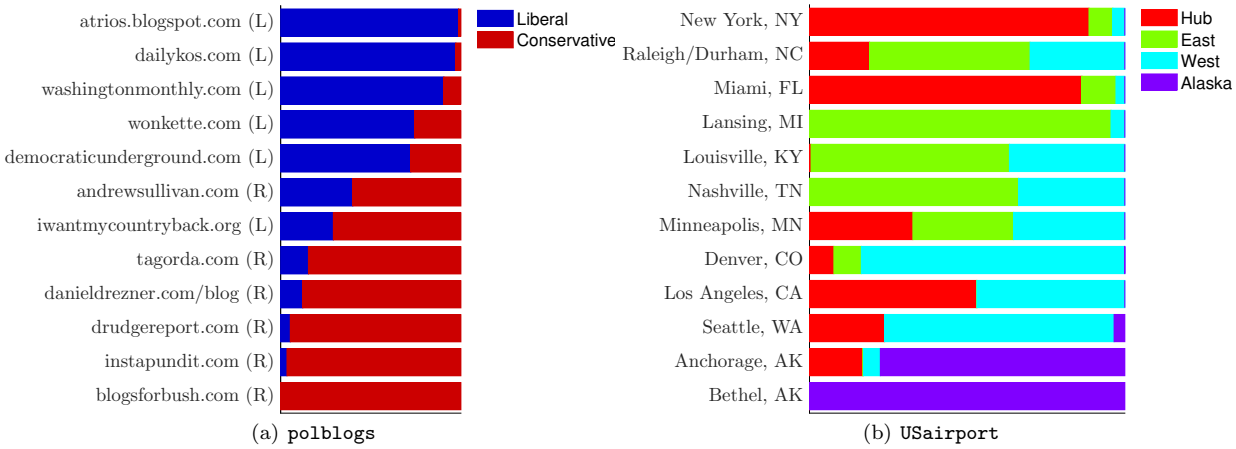


FIG 13. *Relative values of the weights in each community for a subset of the nodes of the (a)* `polblogs` *and (b)* `USairport` *networks.*

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

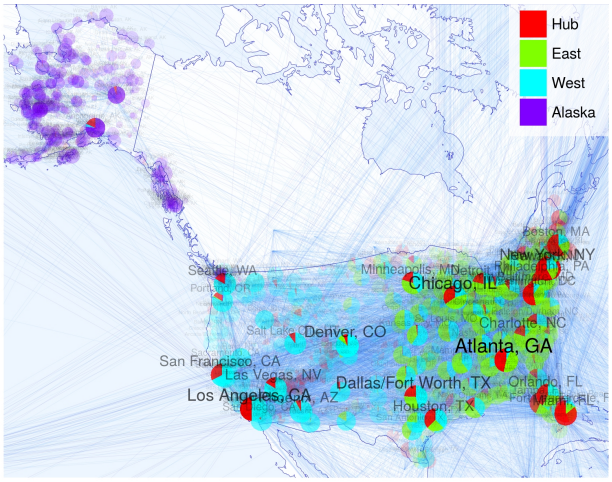


FIG 14. *Usairport network. Each pie chart represents an airport (node) and the size of the pie chart is proportional to the number of connections of that airport. The estimated weights of each airport are normalized so that to sum to 1 and their colors correspond to each uncovered community respectively.*

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

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

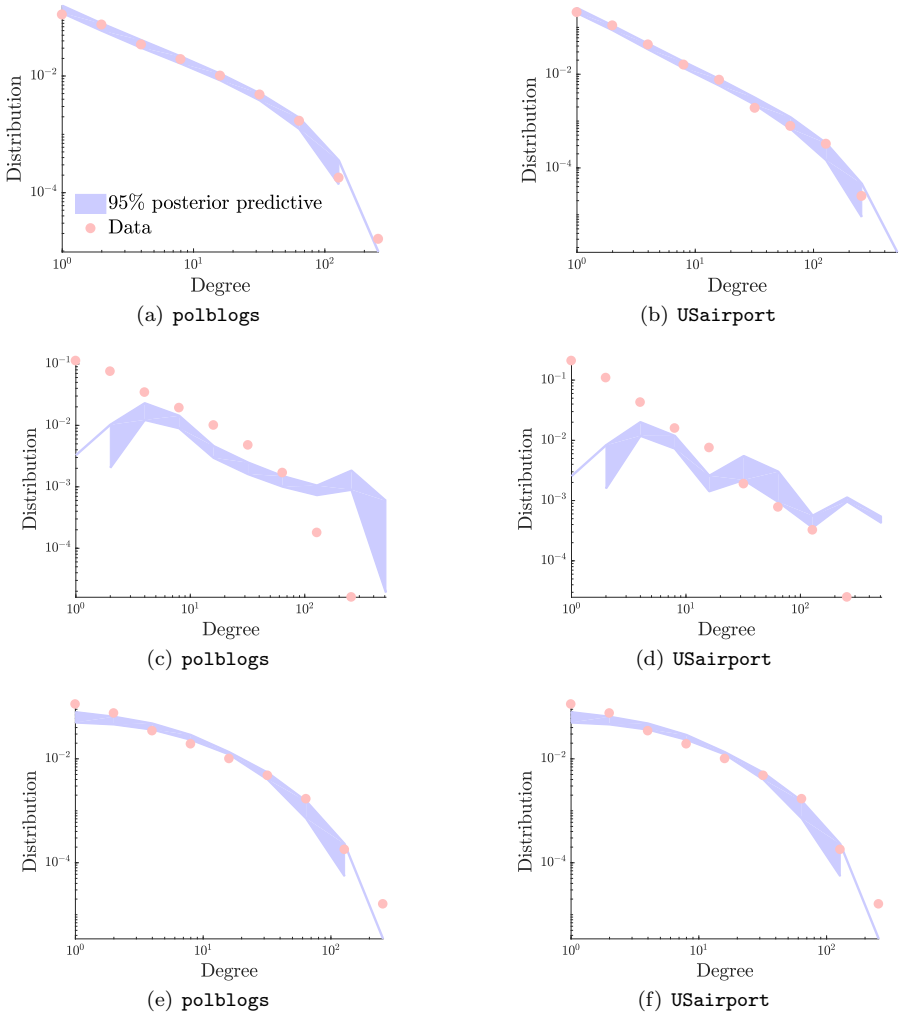


FIG 15. *Empirical degree distribution (red) and posterior predictive (blue) of the (left)* `polblogs` *and (right)* `USairport` *networks under our (top row) CCRM model, (middle row) MMSB and the (bottom row) MLFM.*

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

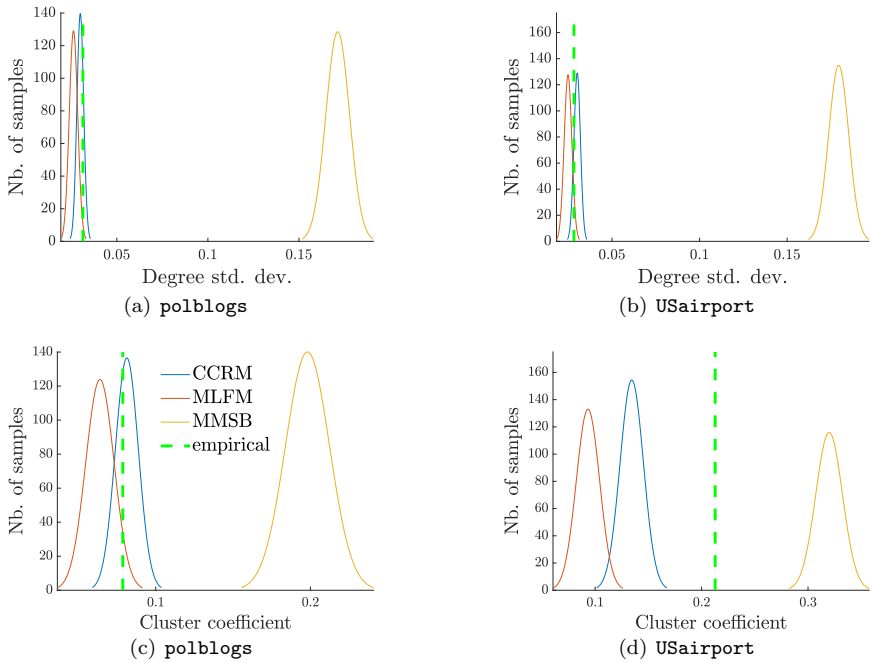


FIG 16. *Posterior predictive plots for the* `polblogs` *and* `USairport` *networks under the CCRM, MLFM and MMSB models. Top row: standard deviation of degree; bottom row: cluster coefficient (or triadic dependence).*

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 closest to it whereas the other two are further away.

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/OxCSML-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 U.S. 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*, 20:1455–1484, 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 (Statistical Methodology)*, 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 Nonnegative 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, no 622, Department of Statistics, University of Washington, 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.

Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities Supplementary material

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

A. Background on completely random measures

A.1. Completely random measures

Completely random measures (CRMs) were introduced by Kingman (1967, 1993) and are now standard tools for constructing flexible Bayesian nonparametric (BNP) models; see for example the surveys of Lijoi and Prünster (2010) or Daley and Vere-Jones (2008, Section 10.1).

A CRM W on $\mathbb{R}_+$ is a random measure such that, for any collection of disjoint measurable subsets $A_1, \ldots, A_n$ of $\mathbb{R}_+$, $W(A_1), \ldots, W(A_n)$ are independent. A CRM can be decomposed into a sum of three independent parts: a non-random measure, a countable collection of atoms with random weights (or masses) at fixed locations, and a countable collection of atoms with random masses and random locations. Here, we will only consider CRMs with random masses and random locations, which take the form

$$W = \sum_{i=1}^{\infty} w_i \delta_{\theta_i} \tag{S1}$$

where the $w_i \in \mathbb{R}_+$ are the masses and $\theta_i \in \mathbb{R}_+$ are the locations. The law of W can actually be characterized by a Poisson point process $N = \{(w_i, \theta_i)_{i=1,2,\ldots}\}$ on $\mathbb{R}_+^2$ with mean measure $\nu(dw, d\theta)$. We focus here on the case where the CRM is homogeneous with independent increments. This implies that the locations θ_i are independent of the weights w_i and the mean measure decomposes as $\nu(dw, d\theta) = \rho(dw)\lambda(d\theta)$ where λ is the Lebesgue measure and ρ is a measure on $\mathbb{R}_+$ that satisfies the condition

$$\int_0^{\infty} (1 - e^{-w})\rho(dw) < \infty. \tag{S2}$$

We write $W \sim \text{CRM}(\rho, \lambda)$. Note that $W([0, T]) < \infty$ a.s. for any $T > 0$ while $W(\mathbb{R}_+) = \infty$ a.s. if ρ is not degenerate at 0. If

$$\int_0^{\infty} \rho(dw) = \infty, \tag{S3}$$

then there will be a.s. an infinite number of jumps in any interval $[0, T]$ and we refer to the CRM as infinite-activity. Otherwise, it is called finite-activity. Let $\overline{\rho}$ be the tail Lévy intensity defined by

$$\overline{\rho}(x) = \int_x^{\infty} \rho(dw) \tag{S4}$$

for $x > 0$. This function corresponds to the expected number of points (w_i, θ_i) such that $w_i > x$ and $\theta_i \in [0, 1]$, and its asymptotic properties play an important role in the characterization of the graph properties.

A.2. Vectors of CRMs

Multivariate extensions of CRMs have been proposed by various authors (Epifani and Lijoi, 2010; Leisen and Lijoi, 2011; Leisen et al., 2013; Griffin et al., 2013; Lijoi et al., 2014). These models are closely related to Lévy copulas (Tankov, 2003; Cont and Tankov, 2003; Kallsen and Tankov, 2006) and multivariate subordinators on cones (Barndorff-Nielsen et al., 2001; Skorohod, 1991). A vector of CRMs $(W_1, \ldots, W_p)$ on $\mathbb{R}_+$ is a collection of random measures W_k, $k = 1, \ldots, p$, such that, for any collection of disjoint measurable subsets $A_1, \ldots, A_n$ of $\mathbb{R}_+$, the vectors $(W_1(A_1), \ldots, W_p(A_1))$, $(W_1(A_2), \ldots, W_p(A_2))$, $\ldots$, $(W_1(A_n), \ldots, W_p(A_n))$ are mutually independent. We only consider here vectors of CRMs with both random weights and random locations. In this case, the measures W_k, $k = 1, \ldots, p$, are a.s. discrete and take the form

$$W_k = \sum_{i=1}^{\infty} w_{ik} \delta_{\theta_i}. \tag{S5}$$

The law of the vector of CRMs can be characterized by a Poisson point process on $\mathbb{R}_+^{p+1}$ with mean measure $\nu(dw_1, \ldots, dw_p, d\theta)$. We focus again on vectors of homogeneous CRMs with independent increments where the mean measure of the Poisson process can be written as

$$\nu(dw_1, \ldots, dw_p, d\theta) = \rho(dw_1, \ldots, dw_p)\lambda(d\theta), \tag{S6}$$

where ρ is a measure on $\mathbb{R}_+^p$, concentrated on $\mathbb{R}_+^p \backslash \{\mathbf{0}\}$, which satisfies

$$\int_{\mathbb{R}_+^p} \min\left(1, \sum_{k=1}^{p} w_k\right) \rho(dw_1, \ldots, dw_p) < \infty. \tag{S7}$$

We use the same notation as for univariate (scalar) CRMs and write simply $(W_1, \ldots, W_p) \sim \mathrm{CRM}(\rho, \lambda)$. A key quantity is the multivariate Laplace exponent. For $(t_1, \ldots, t_p) \in (0, \infty)^p$, this is defined by

$$\psi(t_1, \ldots, t_p) := -\log \mathbb{E}\left[e^{-\sum_{k=1}^{p} t_k W_k([0,1])}\right] \tag{S8}$$

$$= \int_{\mathbb{R}_+^p} \left(1 - e^{-\sum_{k=1}^{p} t_k w_k}\right) \rho(dw_1, \ldots, dw_p). \tag{S9}$$

Note that this quantity involves a p-dimensional integral which may not have an analytic form, therefore not easily computable, and may be expensive to evaluate numerically. As for CRMs, if

$$\int_{\mathbb{R}_+^p} \rho(dw_1, \ldots, dw_p) = \infty \tag{S10}$$

then there will be an infinite number of $\theta_i \in [0, T]$ for which $\sum_k w_{ik} > 0$ and the vector of CRMs is called infinite-activity. Otherwise, it is called finite-activity. Note that some (but not all) CRMs may still be marginally finite-activity.

B. Proof of Propositions 4 and 5

The proof of (Caron and Fox, 2017, Appendix C) can be directly adapted to the model proposed in this paper which is the multivariate generalization of their framework in a multivariate setting. We provide a sketch of the proof. As Z is a jointly exchangeable point process verifying the expression in Eq. (21) of the main paper and the moment condition (24), it then follows from the law of large numbers that

$$N_\alpha^{(e)} = \Theta(\alpha^2) \text{ a.s. as } \alpha \to \infty.$$

Finite-activity case. If the vector of CRMs is finite-activity, the locations of the weights arise from a homogeneous Poisson process with finite rate, and $N_\alpha = \Theta(\alpha)$ a.s. It follows that

$$N_\alpha^{(e)} = \Theta(N_\alpha^2) \text{ a.s. as } \alpha \to \infty.$$

Infinite-activity case. Following Caron and Fox (2017), one can lower bound the counting process of the nodes N_α, by a counting process $\widetilde{N}_\alpha$ which is conditionally Poisson, and the same proof applies. For infinite-activity CCRMs, we use the fact that $\psi\left(W_1([0,\alpha]),\ldots,W_p([0,\alpha])\right)\to\infty$ a.s., it follows that $N_\alpha/\alpha\to\infty$ a.s., and therefore

$$N_\alpha^{(e)} = o(N_\alpha^2) \text{ a.s. as } \alpha\to\infty.$$

Finally, for compound CRMs with regularly varying ρ_0 with exponent σ and slowly varying function ℓ such that $\liminf_{t\to\infty}\ell(t)>0$, Proposition S2 in Appendix I implies that $\liminf N_\alpha/\alpha^{1+\sigma}>0$ a.s. and

$$N_\alpha^{(e)} = O(N_\alpha^{2/(1+\sigma)}) \text{ a.s. as } \alpha\to\infty.$$

C. Details on some derivations

C.1. Proof of Equation (20)

Note that for $p\in\mathbb{R}, a, b>0$ we have the identity

$$\int_0^\infty x^{p-1}e^{-ax-b/x}dx = \frac{2K_p(2\sqrt{ab})}{(a/b)^{p/2}}.$$

Additionally, $K_\nu(z)=K_{-\nu}(z)$. We have

$$\begin{aligned}
\rho(dw_1,\ldots,dw_p) &= \int_0^\infty w_0^{-p}F\left(\frac{dw_1}{w_0},\ldots,\frac{dw_p}{w_0}\right)\rho_0(dw_0)\\
&= \frac{1}{\Gamma(1-\sigma)}\int_0^\infty w_0^{-p-\sigma-1}e^{-\tau w_0}\prod_{k=1}^p\left(\frac{w_k}{w_0}\right)^{a_k-1}e^{-b_kw_k/w_0}\frac{b_k^{a_k}}{\Gamma(a_k)}dw_0dw_{1:p}\\
&= \frac{1}{\Gamma(1-\sigma)}\left[\prod_{k=1}^p w_k^{a_k-1}\frac{b_k^{a_k}}{\Gamma(a_k)}\right]\int_0^\infty w_0^{-\sum_k a_k-\sigma-1}e^{-\tau w_0-(\sum_k b_kw_k)/w_o}dw_0dw_{1:p}\\
&= \frac{1}{\Gamma(1-\sigma)}\left[\prod_{k=1}^p w_k^{a_k-1}\frac{b_k^{a_k}}{\Gamma(a_k)}\right]\frac{2K_{-\sigma-\sum_k a_k}(2\sqrt{\tau\sum_k b_kw_k})}{(\frac{\tau}{\sum_k b_kw_k})^{-(\sigma+\sum_k a_k)/2}}dw_{1:p}\\
&= \frac{2}{\Gamma(1-\sigma)}\left[\prod_{k=1}^p w_k^{a_k-1}\frac{b_k^{a_k}}{\Gamma(a_k)}\right]\left(\frac{\tau}{\sum_k b_kw_k}\right)^{(\sigma+\sum_k a_k)/2}K_{\sigma+\sum_k a_k}\left(2\sqrt{\tau\sum_k b_kw_k}\right)dw_{1:p}.
\end{aligned}$$

D. Proof of Theorem 3

Let $D_\alpha^*=\sum_{k=1}^p D_{k\alpha}([0,\alpha])$ be the number of edges in the directed graph of size α, $W_{k,\alpha}^*=W_k([0,\alpha])$ and $W_\alpha^*=(W_{1,\alpha}^*,\ldots,W_{p,\alpha}^*)^T$. Using Campbell's theorem,

$$\begin{aligned}
\mathbb{E}\left[D_\alpha^*\right] &= \mathbb{E}\left[\mathbb{E}\left[D_\alpha^*|W_\alpha^*\right]\right] = \mathbb{E}\left[(W_\alpha^*)^TW_\alpha^*\right]\\
&= \mathbb{E}\left[W_\alpha^*\right]^T\mathbb{E}\left[W_\alpha^*\right] + \mathrm{tr}(\mathrm{cov}\left(W_\alpha^*\right))\\
&= \alpha^2\mu^T\mu + \alpha\mathrm{tr}(\Sigma)
\end{aligned} \tag{S11}$$

where we define

$$\mu=\int_{\mathbb{R}_+^p} w\rho(dw_1,\ldots,dw_p),\quad \Sigma=\int_{\mathbb{R}_+^p} ww^T\rho(dw_1,\ldots,dw_p).$$

Let $w_i=(w_{i1},\ldots,w_{ip})$. Then, using the extended Slivnyak-Mecke theorem (Møller and Waagepetersen,

2003, Theorem 3.3) we obtain

$$\begin{aligned}\mathbb{E}[N_\alpha^{(e)}] &= \mathbb{E}\left[\mathbb{E}[N_\alpha^{(e)}|W_1,\dots,W_p]\right] \\ &= \mathbb{E}\left[\sum_i \mathbb{1}_{\theta_i\leq\alpha}\left[\left(1-e^{-w_i^T w_i}\right)+\frac{1}{2}\sum_{j\neq i}\mathbb{1}_{\theta_j\leq\alpha}\left(1-e^{-2w_i^T w_j}\right)\right]\right] \\ &= \alpha\int_{\mathbb{R}_+^p}\left(1-e^{-w^T w}\right)\rho(dw_1,\dots,dw_p)+\frac{\alpha^2}{2}\int_{\mathbb{R}_+^p}\psi(2w_1,\dots,2w_p)\rho(dw_1,\dots,dw_p).\end{aligned} \tag{S12}$$

Using the extended Slivnyak-Mecke theorem and then Campbell's theorem,

$$\begin{aligned}\mathbb{E}[N_\alpha] &= \mathbb{E}\left[\mathbb{E}\left[N_\alpha|W_1,\dots,W_p\right]\right] \\ &= \mathbb{E}\left[\sum_i\left(1-e^{-2w_i^T(\sum_{j\neq i}w_j 1_{\theta_j\leq\alpha})-w_i^T w_i}\right)\mathbb{1}_{\theta_i\leq\alpha}\right] \\ &= \alpha\int_{\mathbb{R}_+^p}\mathbb{E}\left(1-e^{-2w^T(\sum_j w_j\mathbb{1}_{\theta_j\leq\alpha})-w^T w}\right)\rho(dw_1,\dots,dw_p) \\ &= \alpha\int_{\mathbb{R}_+^p}\left(1-e^{-w^T w-\alpha\psi(2w_1,\dots,2w_p)}\right)\rho(dw_1,\dots,dw_p).\end{aligned} \tag{S13}$$

By monotone convergence, we have, as α tends to infinity,

$$\mathbb{E}[N_\alpha]\sim\alpha\int_{\mathbb{R}_+^p}\rho(dw_1,\dots,w_p)$$

if the CRM is finite-activity and $\mathbb{E}[N_\alpha]=\omega(\alpha)$ otherwise.

E. Approximation of the log-posterior density

The posterior probability density function, up to a normalizing constant, takes the form

$$\begin{aligned}&p\left((w_{1k},\dots,w_{N_\alpha k})_{k=1,\dots,p},\phi,\alpha\mid(z_{ij})_{1\leq i,j\leq N_\alpha}\right) \\ &\propto\left[\prod_{i=1}^{N_\alpha}\prod_{j=1}^{N_\alpha}\left(\frac{1-e^{-\sum_k w_{ik}w_{jk}}}{e^{-\sum_k w_{ik}w_{jk}}}\right)^{z_{ij}}\right]e^{-\sum_{k=1}^p(w_{*k}+\sum_{i=1}^{N_\alpha}w_{ik})^2} \\ &\quad\times\left[\prod_{i=1}^{N_\alpha}\rho(w_{i1},\dots,w_{ip};\phi)\right]\alpha^{N_\alpha}p(\phi,\alpha)g_{*\alpha}(w_{*1},\dots,w_{*p};\phi)\end{aligned} \tag{S14}$$

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 $(W_1([0,\alpha]),\dots,W_p([0,\alpha]))$.

This is intractable due to the lack of an analytic expression for $g_{*\alpha}$. We can however approximate the log-posterior. Noting that $w_{*k}=o(\sum_{i=1}^{N_\alpha}w_{ik})$ as $\alpha\to\infty$, we have

$$(w_{*k}+\sum_{i=1}^{N_\alpha}w_{ik})^2\simeq(\sum_{i=1}^{N_\alpha}w_{ik})^2+2w_{*k}\sum_{i=1}^{N_\alpha}w_{ik}.$$

Using this approximation, one can now integrate out w_{*k} and we then obtain the approximation

$$\begin{aligned}&p((w_{1k},\dots,w_{N_\alpha k})_{k=1,\dots,p},\phi,\alpha\mid(z_{ij})_{1\leq i,j\leq N_\alpha}) \\ &\simeq\propto\left[\prod_{i=1}^{N_\alpha}\prod_{j=1}^{N_\alpha}\left(\frac{1-e^{-\sum_k w_{ik}w_{jk}}}{e^{-\sum_k w_{ik}w_{jk}}}\right)^{z_{ij}}\right]e^{-\sum_{k=1}^p(\sum_{i=1}^{N_\alpha}w_{ik})^2} \\ &\quad\times\left[\prod_{i=1}^{N_\alpha}\rho(w_{i1},\dots,w_{ip};\phi)\right]\alpha^{N_\alpha}\exp\left[-\alpha\psi\left(2\sum_{i=1}^{N_\alpha}w_{i1},\dots,2\sum_{i=1}^{N_\alpha}w_{ip};\phi\right)\right]p(\phi,\alpha)\end{aligned} \tag{S15}$$

where $\psi(t_1, \ldots, t_p)$ is the multivariate Laplace exponent, which can be evaluated numerically.

F. Simulation from the tilted truncated random measure

We aim at sampling points of a Poisson point process with (finite) mean measure

$$\begin{aligned}&\alpha e^{-w_0 \sum_{k=1}^p \lambda_k \beta_k} f(\beta_1, \ldots, \beta_p) \rho_0(w_0) \mathbb{1}_{w_0 > \varepsilon} \\ &= \frac{e^{-w_0 \sum_{k=1}^p \lambda_k \beta_k} f(\beta_1, \ldots, \beta_p)}{M(-\lambda_1 w_0, \ldots, -\lambda_p w_0)} \times \alpha M(-\lambda_1 w_0, \ldots, -\lambda_p w_0) \rho_0(w_0) \mathbb{1}_{w_0 > \varepsilon}\end{aligned}$$

where we recall that M is the Laplace transform of the scores. One can first sample points (w_{0i}) from a Poisson point process with mean measure

$$\alpha M(-\lambda_1 w_0, \ldots, -\lambda_p w_0) \rho_0(w_0) \mathbb{1}_{w_0 > \varepsilon}, \tag{S16}$$

and then sample the scores

$$(\beta_{i1}, \ldots, \beta_{ip}) | w_{i0} \overset{\text{ind}}{\sim} H(\cdot | w_{i0}) \tag{S17}$$

with H an exponentially tilted version of F

$$H(d\beta_1, \ldots, d\beta_p | w_0) = \frac{e^{-w_0 \sum_{k=1}^p \gamma_k \beta_k} F(d\beta_1, \ldots, d\beta_p)}{\int_{\mathbb{R}_+^p} e^{-w_0 \sum_{k=1}^p \gamma_k \widetilde{\beta}_k} F\left(d\widetilde{\beta}_1, \ldots, d\widetilde{\beta}_p\right)}. \tag{S18}$$

In the case where F is a product of gamma distributions, H is also a product of gamma distributions. The mean measure (S16) is a tilted version of ρ_0, where the tilting function is monotone decreasing. We explain in the next section how to sample from this point process using adaptive thinning. More generally, the mean measure may be of the form of a tilted stable measure (with potentially heavier tails than exponential). One could use in this case the generic series representation developed by Rosiński (2007) for tempered/tilted stable processes.

Simulation from a tilted truncated generalized gamma process We want to sample points from a Poisson process with truncated mean measure

$$\rho^{\varepsilon}(dw) = h(w) w^{-1-\sigma} e^{-\tau w} \mathbb{1}_{w > \varepsilon} dw \tag{S19}$$

where h is a monotone decreasing and bounded function, and (τ, σ) verify either $\tau \geq 0$ and $\sigma \in (0, 1)$, or $\tau > 0$ and $\sigma \in (-1, 0]$. We will resort to the method of adaptive thinning (Lewis and Shedler, 1979; Ogata, 1981; Favaro and Teh, 2013).

For $\tau > 0$, consider the family of adaptive bounds

$$g_t(s) = h(t) t^{-1-\sigma} \exp(-\tau s)$$

with $g_t(s) > \rho(s)$ for $s > t$. We have,

$$\begin{aligned}G_t(s) &= \int_t^s g_t(s') ds' \\ &= \frac{h(t)}{\tau} t^{-1-\sigma} (\exp(-\tau t) - \exp(-\tau s))\end{aligned}$$

and

$$G_t^{-1}(r) = -\frac{1}{\tau} \log\left(\exp(-\tau t) - \frac{r\tau}{t^{-1-\sigma} h(t)}\right).$$

For $\tau = 0$, we consider bounds

$$g_t(s) = h(t) s^{-1-\sigma}$$

and we obtain

$$G_t(s) = \frac{h(t)}{\sigma}(t^{-\sigma} - s^{-\sigma},$$
$$G_t^{-1}(r) = \left[t^{-\sigma} - \frac{r\sigma}{h(t)}\right]^{-1/\sigma}.$$

The adaptive thinning sampling scheme is as follows:

1. Set $N = \emptyset, t = \varepsilon$.
2. Iterate until termination:
   (a) Draw $r \sim \text{Exp}(1)$.
   (b) If $r > G_t(\infty)$, terminate; else set $t' = G_t^{-1}(r)$.
   (c) With probability $\rho^\varepsilon(t')/g_t(t')$ accept sample t' and set $N = N \cup \{t'\}$.
   (d) Set $t = t'$ and continue.
3. Return N.

N obtained by the sampling scheme above is a draw from the Poisson random measure with intensity ρ^ε on $[\varepsilon, +\infty)$. The efficiency of this approach depends on the acceptance probability. For $\tau > 0$ and $s > t$ this is given by the ratio

$$\frac{\rho^\varepsilon(s)}{g_t(s)} = \frac{h(s)s^{-1-\sigma}}{h(t)t^{-1-\sigma}} < 1.$$

G. Bipartite networks

It is possible to use a construction similar to that of Section 2 to model bipartite graphs, and extend the model of Caron (2012). A bipartite graph is a graph with two types of nodes, where only connections between nodes of different types are allowed. Nodes of the first type are embedded at locations $\theta_i \in \mathbb{R}_+$, and nodes of the second type at locations $\theta'_j \in \mathbb{R}_+$. The bipartite graph is represented by a (non-symmetric) point process

$$Z = \sum_{i,j} z_{ij}\delta_{(\theta_i,\theta'_j)}, \tag{S20}$$

where $z_{ij} = 1$ if there is an edge between node i of type 1 and node j of type 2.

Statistical Model. We consider the model

$$W_1, \ldots, W_p \sim \text{CRM}(\rho, \lambda)$$
$$W'_1, \ldots, W'_p \sim \text{CRM}(\rho', \lambda)$$

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

$$D_k | W_k, W'_k \sim \text{Poisson}\left(W_k \times W'_k\right)$$
$$D_k = \sum_{i,j} n_{ijk}\delta_{(\theta_i,\theta'_j)}.$$

In (S20), z_{ij} is given by $z_{ij} = \min(1, \sum_{k=1}^p n_{ijk})$.

Posterior inference. We derive here the inference algorithm when $(W_1, \ldots, W_p)$ and $(W'_1, \ldots, W'_p)$ are compound CRMs with F and ρ_0 taking the form (17) and (18) respectively.

Assume that we observe a set of connections $z = (z_{ij})_{i=1,\ldots,N_\alpha;j=1,\ldots N'_\alpha}$. For $k = 1, \ldots, p$, we introduce latent variables n_{ijk}, for $1 \leq i \leq N_\alpha$, $1 \leq j \leq N'_\alpha$,

$$(n_{ij1}, \ldots, n_{ijp}) | w, w', z \sim \begin{cases} \delta_{(0,\ldots,0)} & \text{if } z_{ij} = 0 \\ \text{tPoisson}(w_{i1}w'_{j1}, \ldots, w_{ip}w'_{jp}) & \text{if } z_{ij} = 1. \end{cases}$$

The objective is to approximate the posterior density

$$p\left((w_{10},\dots w_{N_\alpha 0}),(\beta_{1k},\dots,\beta_{N_\alpha k},w_{*k})_{k=1,\dots,p},(w'_{10},\dots,w'_{N'_\alpha 0}),(\beta'_{1k},\dots,\beta'_{N'_\alpha k},w'_{*k})_{k=1,\dots,p},\phi,\alpha,\phi',\alpha'|z\right).$$

Let $m_{ik}=\sum_{j=1}^{N'_\alpha} n_{ijk}$ and $m_i=\sum_{k=1}^{p} m_{ik}$. The MCMC algorithm iterates as follows:

1. Update $(\alpha,\phi)|$rest using a Metropolis-Hastings step.
2. Update
$$w_{i0}|\text{rest} \sim \text{Gamma}\left(m_i-\sigma,\tau+\sum_{k=1}^{p}\beta_{ik}\left[\gamma_k+\left(\sum_{j=1}^{N'_\alpha} w'_{jk}\right)+w'_{*k}\right]\right).$$
3. Update
$$\beta_{ik}|\text{rest} \sim \text{Gamma}\left(a_k+m_{ik},b_k+w_{i0}\left[\gamma_k+\left(\sum_{j=1}^{N'_\alpha} w'_{jk}\right)+w'_{*k}\right]\right).$$
4. Update $(w_{*1},\dots,w_{*p})|$rest.
5. Update the latent variables $n_{ijk}|$rest.
6. Repeat steps 1-4 to update (α',ϕ'), $(w'_{10},\dots,w'_{N'_\alpha 0})$, $(\beta'_{1k},\dots,\beta'_{N'_\alpha k})_{k=1,\dots,p}$ and $(w'_{*1},\dots,w'_{*p})$.

H. Gaussian approximation of the sum of small jumps

Theorem S1 *Consider the multivariate random variable $X_\varepsilon \in \mathbb{R}_+^p$ with moment generating function*

$$\mathbb{E}[e^{-t^T X_\varepsilon}]=\exp\left[-\alpha\int_{\mathbb{R}_+^p}\left(1-e^{-\sum_{k=1}^p t_k w_k}\right)\rho_\varepsilon(dw_1,\dots,dw_p)\right],$$

where $\alpha>0$ and

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

with $\varepsilon>0$, ρ_0 a Lévy measure on $\mathbb{R}_+$ and F a probability distribution on $\mathbb{R}_+^p$. F admits the density f verifying the conditions

$$\int_0^\infty f(zu_1,\dots,zu_p)dz>0 \ U\text{-almost everywhere}$$

and

$$\int_{\mathbb{R}_+^p}\|\beta_{1:p}\|^2 f(\beta_1,\dots,\beta_p)d\beta_{1:p}<\infty,$$

where U is the uniform distribution on the unit sphere S^{p-1}. If ρ_0 is a regularly varying Lévy measure with exponent $\sigma\in(0,1)$, i.e.

$$\int_x^\infty \rho_0(dw_0)\overset{x\downarrow 0}{\sim} x^{-\sigma}\ell(1/x)$$

where $\ell:(0,\infty)\to(0,\infty)$ is a slowly varying function, then

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

as $\varepsilon\to 0$, where

$$\mu_\varepsilon=\alpha\int_{\mathbb{R}_+^p} w\rho_\varepsilon(dw_1,\dots,dw_p)$$
$$\Sigma_\varepsilon=\alpha\int_{\mathbb{R}_+^p} ww^T\rho_\varepsilon(dw_1,\dots,dw_p)$$

with

$$\mu_\varepsilon \sim \alpha \mathbb{E}[\beta] \frac{\sigma}{1-\sigma} \varepsilon^{1-\sigma} \ell(1/\varepsilon) \text{ as } \varepsilon \to 0$$
$$\Sigma_\varepsilon \sim \alpha \mathbb{E}[\beta\beta^T] \frac{\sigma}{2-\sigma} \varepsilon^{2-\sigma} \ell(1/\varepsilon) \text{ as } \varepsilon \to 0$$

where the distribution of β is F.

Proof. We write the model in spherical form. Let $r = \sqrt{\sum w_k^2}$ and $u_k = \frac{w_k}{r}$, for $k = 1, \ldots, p-1$. The determinant of the Jacobian is $\frac{r^{p-1}}{\sqrt{1-\sum_{k=1}^{p-1} u_k^2}}$ and therefore

$$\begin{aligned}\widetilde{\rho}_\varepsilon(r, u_1, \ldots, u_{p-1}) &= \frac{r^{p-1}}{u_p} e^{-r\sum_{k=1}^p \gamma_k u_k} \int_0^\varepsilon w_0^{-p} f\left(\frac{ru_1}{w_0}, \ldots, \frac{ru_p}{w_0}\right) \rho_0(dw_0) dr du_{1:p-1} \\ &:= \mu_\varepsilon(dr|u_{1:p-1}) U(du_{1:p-1})\end{aligned}$$

where $u_p = \sqrt{1-\sum_{k=1}^{p-1} u_k^2}$, $\mu_\varepsilon(dr|u) = r^{p-1} e^{-r\sum_{k=1}^p \gamma_k u_k} \int_0^\varepsilon w_0^{-p} f\left(\frac{ru_1}{w_0}, \ldots, \frac{ru_p}{w_0}\right) \rho_0(dw_0) dr$ and $U(du) = \frac{1}{u_p} du_{1:p}$ is the uniform distribution on the unit sphere S^{p-1}.

In order to apply Theorem 2.4 in Cohen and Rosinski (2007) (see also Asmussen and Rosiński, 2001), we need to show that there exists a function $b_\varepsilon : (0,1] \to (0, +\infty)$ such that

$$\lim_{\varepsilon \to 0} \frac{\sigma_\varepsilon(u)}{b_\varepsilon} > 0, \; U\text{-almost everywhere} \tag{S21}$$

where

$$\sigma_\varepsilon^2(u) = \int_0^\infty r^2 \mu_\varepsilon(dr|u)$$

and for every $\kappa > \varepsilon$

$$\lim_{\varepsilon \to 0} \frac{1}{b_\varepsilon^2} \int_{\|w_{1:p}\| > \kappa b_\varepsilon} \|w_{1:p}\|^2 \rho_\varepsilon(dw_1, \ldots, dw_p) = 0. \tag{S22}$$

Assume that $\int_0^\infty f(zu_1, \ldots, zu_p)\, dz > 0$ U-almost everywhere. With the change of variable $z = \frac{r}{w_0}$, and the dominated convergence theorem we obtain

$$\begin{aligned}\sigma_\varepsilon^2(u) &= \int_0^\infty z^{p+1} f(zu_1, \ldots, zu_p) \left[\int_0^\varepsilon e^{-zw_0\sum_{k=1}^p \gamma_k u_k} w_0^2 \rho_0(dw_0)\right] dz \\ &\sim \left(\int_0^\infty z^{p+1} f(zu_1, \ldots, zu_p)\, dz\right) \left(\int_0^\varepsilon w_0^2 \rho_0(dw_0)\right) \text{ as } \varepsilon \to 0 \\ &\sim \left(\int_0^\infty z^{p+1} f(zu_1, \ldots, zu_p)\, dz\right) \frac{\sigma}{2-\sigma} \varepsilon^{2-\sigma} \ell(1/\varepsilon) \text{ as } \varepsilon \to 0.\end{aligned}$$

Take $b_\varepsilon = \varepsilon^{1-\sigma/2}\sqrt{\ell(1/\varepsilon)}$, we then obtain

$$\lim_{\varepsilon \to 0} \frac{\sigma_\varepsilon^2(u)}{b_\varepsilon^2} = \left(\int_0^\infty z^{p+1} f(zu_1, \ldots, zu_p)\, dz\right) \frac{\sigma}{2-\sigma} > 0, \; U\text{-almost everywhere.} \tag{S23}$$

Now consider for any $\kappa > 0$,

$$\begin{aligned}I_\varepsilon &= \int_{\|w_{1:p}\| > \kappa b_\varepsilon} \|w_{1:p}\|^2 \nu_\varepsilon(dw_1, \ldots, dw_p) \\ &= \int_0^\varepsilon \int_{\|\beta_{1:p}\| > \frac{\kappa b_\varepsilon}{w_0}} w_0^2 \|\beta_{1:p}\|^2 e^{-w_0\sum_{k=1}^p \gamma_k \beta_k} f(\beta_1, \ldots, \beta_k) \rho_0(dw_0) d\beta_{1:p}.\end{aligned}$$

For $w_0 \in (0,\varepsilon)$, we have $\frac{\kappa b_\varepsilon}{w_0} \geq \frac{\kappa b_\varepsilon}{\varepsilon} = \varepsilon^{-\sigma/2}\ell(1/\varepsilon) > \kappa_2\varepsilon^{-\sigma/4}$ for ε small enough as $t^\delta\ell(t) \to 0$ for any $\delta > 0$ as $t \to \infty$. So for ε small enough

$$\begin{aligned} I_\varepsilon &< \int_0^\varepsilon \int_{\|\beta_{1:p}\|>\kappa_2\varepsilon^{-\sigma/4}} w_0^2 \left\|\beta_{1:p}\right\|^2 e^{-w_0\sum_{k=1}^p \gamma_k\beta_k} f\left(\beta_1,\ldots,\beta_k\right)\rho_0(dw_0)d\beta_{1:p} \\ &< \left[\int_{\|\beta_{1:p}\|>\kappa_2\varepsilon^{-\sigma/4}} \left\|\beta_{1:p}\right\|^2 f\left(\beta_1,\ldots,\beta_k\right)d\beta_{1:p}\right]\left[\int_0^\varepsilon w_0^2\rho_0(dw_0)\right]. \end{aligned}$$

As $\left[\int_0^\varepsilon w_0^2\rho_0(dw_0)\right] \sim \frac{\sigma}{2-\sigma}b_\varepsilon^2$ when $\varepsilon \to 0$, we conclude that

$$\lim_{\varepsilon\to 0}\frac{I_\varepsilon}{b_\varepsilon^2} \leq \lim_{\varepsilon\to 0}\frac{\sigma}{2-\sigma}\int_{\|\beta_{1:p}\|>\kappa_2\varepsilon^{-\sigma/4}} \left\|\beta_{1:p}\right\|^2 f\left(\beta_1,\ldots,\beta_k\right)d\beta_{1:p} = 0. \tag{S24}$$

Eq. (S23) and (S24) with Theorem 2.4 of Cohen and Rosinski (2007) yield

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

as $\varepsilon \to 0$, where

$$\begin{aligned} \mu_\varepsilon &= \alpha\int_{\mathbb{R}_+^p} w_{1:p}\ \rho_\varepsilon(dw_1,\ldots,dw_p) \\ &= \alpha\int_{\mathbb{R}_+^p}\int_0^\varepsilon w_0\beta_{1:p}\ e^{-w_0\sum_{k=1}^p\gamma_k\beta_k}\rho_0(dw_0)f(\beta_1,\ldots,\beta_p)d\beta_{1:p} \\ &\sim \alpha\mathbb{E}[\beta_{1:p}]\frac{\sigma}{1-\sigma}\varepsilon^{1-\sigma}\ell(1/\varepsilon) \text{ as } \varepsilon\to 0 \end{aligned}$$

and

$$\begin{aligned} \Sigma_\varepsilon &= \alpha\int_{\mathbb{R}_+^p} w_{1:p}w_{1:p}^T\ \rho_\varepsilon(dw_1,\ldots,dw_p) \\ &\sim \alpha\mathbb{E}[\beta_{1:p}\beta_{1:p}^T]\frac{\sigma}{2-\sigma}\varepsilon^{2-\sigma}\ell(1/\varepsilon) \text{ as } \varepsilon\to 0 \end{aligned}$$

using the dominated convergence theorem and lemmas S3 and S4. ■

I. Technical lemmas

Proposition S2 *Let ν be a Lévy measure defined by Eq. (5) and (13) and let ψ be its multivariate Laplace exponent. Assume that $\overline{\rho}_0$ is a regularly varying function with exponent $\sigma \in (0,1)$*

$$\overline{\rho}_0 \overset{x\downarrow 0}{\sim} x^{-\sigma}\ell(1/x). \tag{S25}$$

Then ψ is (multivariate) regularly varying (Resnick, 2013), with exponent σ. More precisely, for any $(x_1,\ldots x_p) \in (0,\infty)^p$, we have

$$\begin{aligned} \psi(tx_1,\ldots,tx_p) &= \int_{\mathbb{R}_+^p}\left(1-e^{-t\sum_{k=1}^p x_kw_k}\right)\nu(dw_1,\ldots,dw_p) \\ &\overset{t\uparrow\infty}{\sim} t^\sigma\Gamma(1-\sigma)\ell(t)\mathbb{E}\left[\left(\sum_{k=1}^p x_k\beta_k\right)^\sigma\right]. \end{aligned}$$

Proof.

$$\begin{aligned} \psi(tx_1,\ldots,tx_p) &= \int_{\mathbb{R}_+^p}\left(1-e^{-t\sum_{k=1}^p x_kw_k}\right)\nu(dw_1,\ldots,dw_p) \\ &= \int_{\mathbb{R}_+^p}\left(1-e^{-t\sum_{k=1}^p x_kw_k}\right)\nu(dw_1,\ldots,dw_p) \\ &= \int_{\mathbb{R}_+^p} f(\beta_1,\ldots,\beta_p)\left[\int_0^\infty\left(1-e^{-w_0t\sum_{k=1}^p x_k\beta_k}\right)e^{-w_0\sum_{k=1}^p\gamma_k\beta_k}\rho_0(dw_0)\right]d\beta_{1:p}. \end{aligned}$$

Using Lemmas S3, S4, and the dominated convergence theorem we get

$$\psi(tx_1,\ldots,tx_p) \overset{t\uparrow\infty}{\sim} t^{\sigma}\Gamma(1-\sigma)\ell(t)\int_{(0,\infty)^p}\left(\sum_{k=1}^{p}x_k\beta_k\right)^{\sigma} f(\beta_1,\ldots,\beta_p)d\beta_{1:p}.$$

■

Lemma S3 *If*

$$\int_x^{\infty}\rho(dw) \overset{x\downarrow 0}{\sim} x^{-\sigma}\ell(1/x)$$

where $\sigma\in(0,1)$ and ℓ is a slowly varying function (at infinity), then

$$\int_x^{\infty}e^{-cw}\rho(dw) \overset{x\downarrow 0}{\sim} x^{-\sigma}\ell(1/x).$$

Proof.

$$\int_x^{\infty}e^{-cw}\rho(dw) = \int_x^{\infty}\rho(dw) - \int_x^{\infty}(1-e^{-cw})\rho(dw)$$
$$\overset{x\downarrow 0}{\sim} x^{-\sigma}\ell(1/x)$$

as $\int_0^{\infty}(1-e^{-cw})\rho(dw)<\infty$ for any $c>0$. ■

Lemma S4 *(Gnedin et al., 2007; Bingham et al., 1989). Let ρ be a Lévy measure with regularly varying tail Lévy intensity*

$$\int_x^{\infty}\rho(dw) \overset{x\downarrow 0}{\sim} x^{-\sigma}\ell(1/x) \tag{S26}$$

where $\sigma\in(0,1)$ and ℓ is a slowly varying function (at infinity). Then (S26) is equivalent to the following

$$\int_0^{x}w^k\rho(dw) \overset{x\downarrow 0}{\sim} \frac{\sigma}{k-\sigma}x^{k-\sigma}\ell(1/x) \tag{S27}$$

$$\int_0^{\infty}(1-e^{-tw})\rho(dw) \overset{t\uparrow\infty}{\sim} \Gamma(1-\sigma)t^{\sigma}\ell(t) \tag{S28}$$

for any $k\geq 1$.

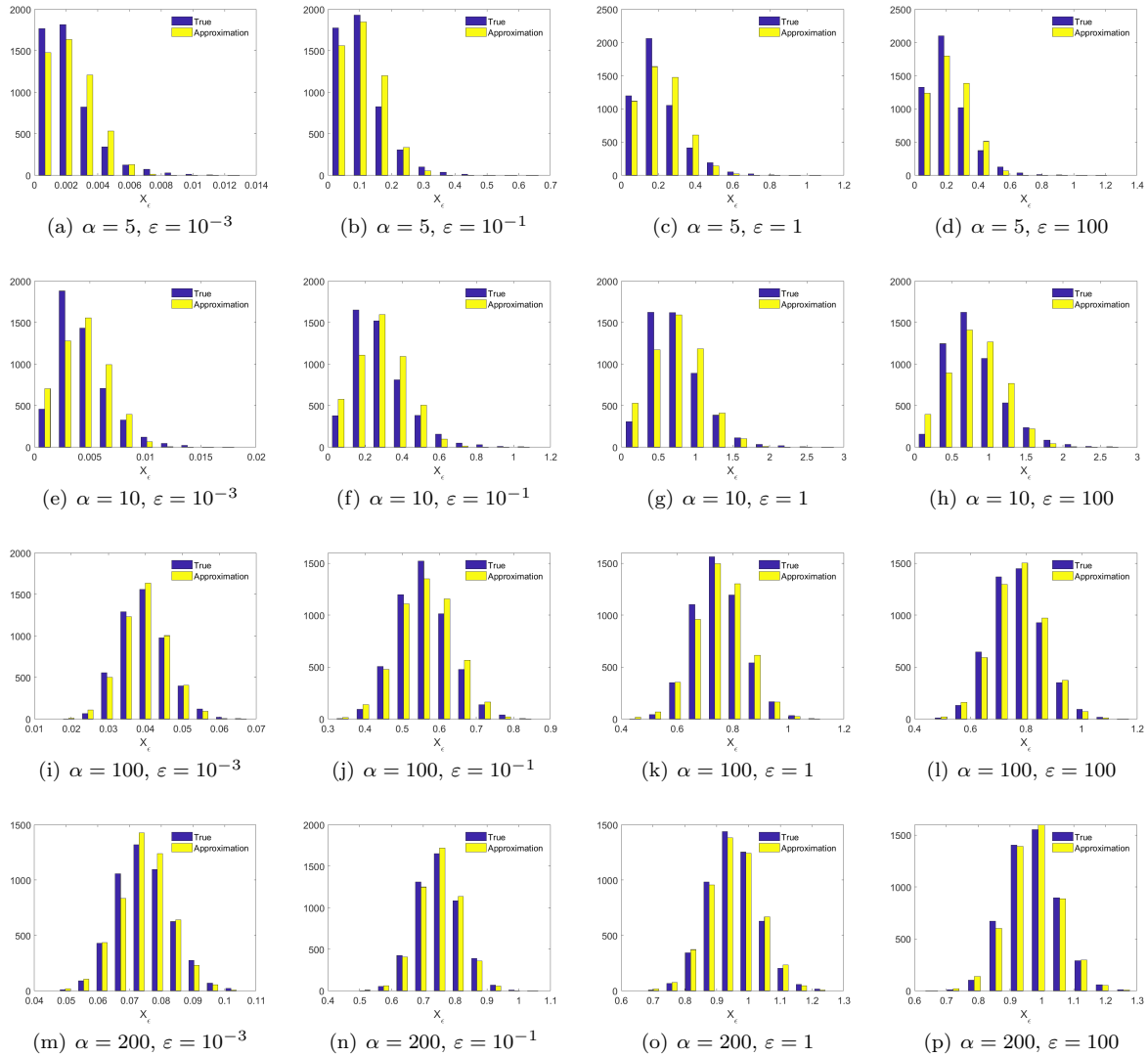
J. Empirical analysis of the Cohen and Rosinski's Gaussian approximation

We provide here some empirical analysis of the Gaussian approximation described in Section 4 in the main paper. We use the parameters $p=1, a=0.2, b=1/2, \tau=1$. We simulate graphs of different sizes $\alpha = 5, 10, 100, 200$ for $\sigma=0.2$ and $\sigma=0.5$, and set for each graph $\lambda_k = w_{*k} + 2\sum_i w_{ik}$. We then simulate the points of a Poisson process with tilted measure (42), using a small truncation threshold $\varepsilon=10^{-8}$. For different thresholds $\varepsilon = 10^{-3}, 10^{-1}, 1, 100$, we compare the true distribution of X_ε to its (truncated) Gaussian approximation. Results are reported in Figures S1-S3. While the theoretical results reported in Section 4 indicate that the Gaussian approximation holds for a fixed α as $\varepsilon\to 0$, it appears from the simulations that the distribution of X_ε is well approximated by a Gaussian distribution for large values of α, and this whatever the value of ε. This is supported by the empirical analysis of the convergence of the Markov chain for different values of ε in Section K.1.

K. Study on the accuracy of the MCMC sampler

K.1. Sensitivity on truncation level ε

This section shows that the MCMC sampler targets the same distribution under different threshold values (or truncation levels) ε. Recall that the threshold is the truncation level for simulating $w_{*1:p}$ in the infinite-activity case. We perform a sensitivity analysis on the impact of the threshold ε for which we use the values



(a) $\alpha = 5$, $\varepsilon = 10^{-3}$ (b) $\alpha = 5$, $\varepsilon = 10^{-1}$ (c) $\alpha = 5$, $\varepsilon = 1$ (d) $\alpha = 5$, $\varepsilon = 100$

(e) $\alpha = 10$, $\varepsilon = 10^{-3}$ (f) $\alpha = 10$, $\varepsilon = 10^{-1}$ (g) $\alpha = 10$, $\varepsilon = 1$ (h) $\alpha = 10$, $\varepsilon = 100$

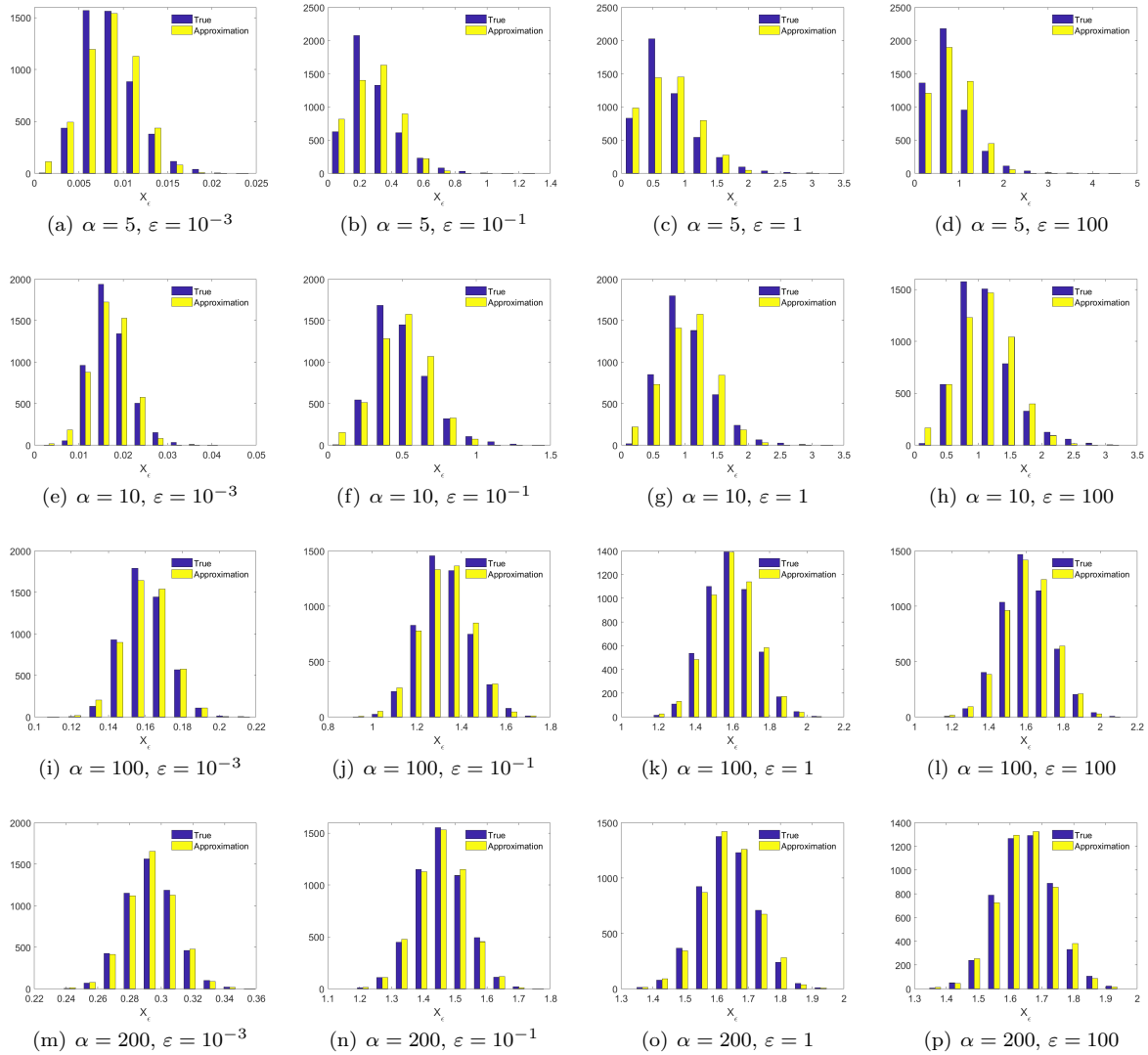
(i) $\alpha = 100$, $\varepsilon = 10^{-3}$ (j) $\alpha = 100$, $\varepsilon = 10^{-1}$ (k) $\alpha = 100$, $\varepsilon = 1$ (l) $\alpha = 100$, $\varepsilon = 100$

(m) $\alpha = 200$, $\varepsilon = 10^{-3}$ (n) $\alpha = 200$, $\varepsilon = 10^{-1}$ (o) $\alpha = 200$, $\varepsilon = 1$ (p) $\alpha = 200$, $\varepsilon = 100$

FIG S1. *Comparison of the Gaussian approximation for* $\sigma = 0.01$ *for different network sizes* $\alpha = 5, 10, 100, 200$ *and different thresholds* $\varepsilon = 10^{-3}, 10^{-1}, 1, 100$.

$100, 0.1, 10^{-4}$ and 10^{-7}. We first generate an undirected graph with $p = 2$ communities and parameters $\alpha = 50$, $\sigma = 0.5$, $\tau = 1$, $b_k = b = 0.1$, $a_k = a = 0.2$. The sampled graph has 1799 nodes and 13,416 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 and set $\mathcal{L} = 10$ leapfrog steps in 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 initialize the chains from the parameter values we use for generating the graph, i.e. from the stationary regime. This is because the objective is to assess the mixing of the chains under different thresholds and see whether chains admit the same limiting distribution. We run 200,000 iterations using our model with p communities. We repeat this procedure four times, using each time one of the truncation values $\varepsilon = 100, 0.1, 10^{-4}, 10^{-7}$.

The trace plots for the parameters $\log\tilde{\alpha}$, σ, $\tilde{b}$, a and $\overline{w}_*$ are shown in Figure S4. The dotted green line corresponds to the value of the model parameters used for generating the graph. It is clear that the chains are very close to each other and they exhibit a good mixing. The plots in Figure S5 show the trace and autocorrelation function of the approximate log-posterior density (up to a constant). There is no noticeable



(a) $\alpha = 5$, $\varepsilon = 10^{-3}$ (b) $\alpha = 5$, $\varepsilon = 10^{-1}$ (c) $\alpha = 5$, $\varepsilon = 1$ (d) $\alpha = 5$, $\varepsilon = 100$

(e) $\alpha = 10$, $\varepsilon = 10^{-3}$ (f) $\alpha = 10$, $\varepsilon = 10^{-1}$ (g) $\alpha = 10$, $\varepsilon = 1$ (h) $\alpha = 10$, $\varepsilon = 100$

(i) $\alpha = 100$, $\varepsilon = 10^{-3}$ (j) $\alpha = 100$, $\varepsilon = 10^{-1}$ (k) $\alpha = 100$, $\varepsilon = 1$ (l) $\alpha = 100$, $\varepsilon = 100$

(m) $\alpha = 200$, $\varepsilon = 10^{-3}$ (n) $\alpha = 200$, $\varepsilon = 10^{-1}$ (o) $\alpha = 200$, $\varepsilon = 1$ (p) $\alpha = 200$, $\varepsilon = 100$

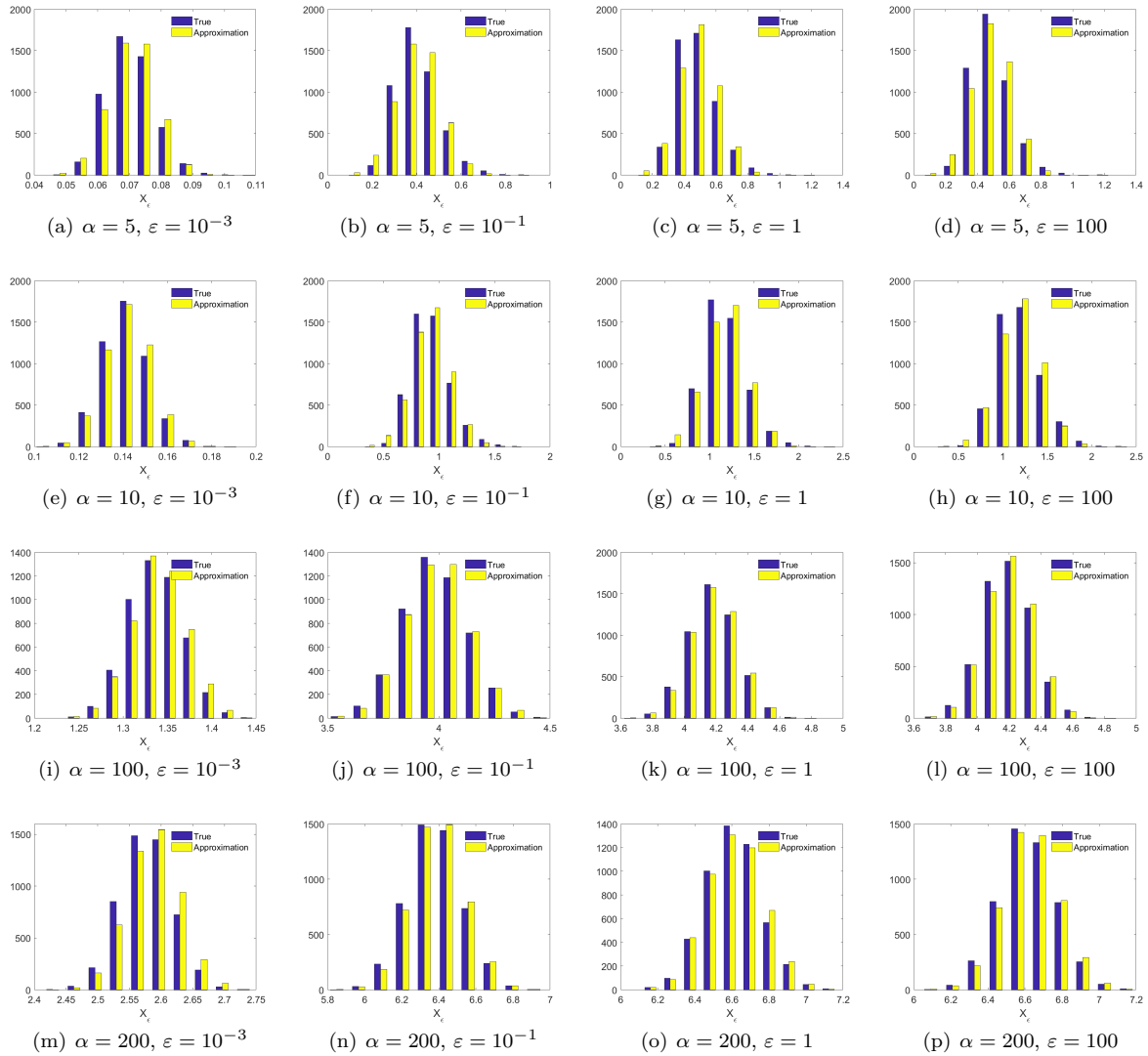
FIG S2. *Comparison of the Gaussian approximation for* $\sigma = 0.2$ *for different network sizes* $\alpha = 5, 10, 100, 200$ *and different thresholds* $\varepsilon = 10^{-3}, 10^{-1}, 1, 100$.

difference in the limiting distribution nor mixing times for different values of ε.

We repeat the sensitivity experiment on inference for another graph generated with $\sigma = 0.2$. We generate an undirected graph with $p = 2$ communities and model parameters $\alpha = 200$, $\sigma = 0.2$, $\tau = 1$, $b_k = b = 0.5$, $a_k = a = 0.2$. The sampled graph has 1268 nodes and 15,811 edges. For the inference,we place the same assumptions for the priors and run 3 MCMC chains in parallel with $\mathcal{L} = 10$ leapfrog steps in the HMC. We repeat the same procedure for each of the values of the threshold $\varepsilon = 100, 0.1, 10^{-4}, 10^{-7}$. Note that in both experiments ($\sigma = 0.5$ and $\sigma = 0.2$) we show the results for one of the 3 chains for each threshold since the behaviour was very similar across chains.

The MCMC trace plots for the parameters $\log \tilde{\alpha}$, σ, $\tilde{b}$, a and $\overline{w}_*$ are shown in Figure S6. As in the case of $\sigma = 0.2$, the MCMC trace is very similar for the different thresholds in all parameters. The trace and autocorrelation of the approximate log-posterior are shown in Figure S7.

Note that this type of sensitivity test concerns only the infinity-activity case which corresponds to the positive σ values. We have shown the plots of the MCMC output for values of $\sigma = 0.5$ and $\sigma = 0.2$ however the same behaviour was observed for other positive σ values too.



(a) $\alpha = 5$, $\varepsilon = 10^{-3}$ (b) $\alpha = 5$, $\varepsilon = 10^{-1}$ (c) $\alpha = 5$, $\varepsilon = 1$ (d) $\alpha = 5$, $\varepsilon = 100$

(e) $\alpha = 10$, $\varepsilon = 10^{-3}$ (f) $\alpha = 10$, $\varepsilon = 10^{-1}$ (g) $\alpha = 10$, $\varepsilon = 1$ (h) $\alpha = 10$, $\varepsilon = 100$

(i) $\alpha = 100$, $\varepsilon = 10^{-3}$ (j) $\alpha = 100$, $\varepsilon = 10^{-1}$ (k) $\alpha = 100$, $\varepsilon = 1$ (l) $\alpha = 100$, $\varepsilon = 100$

(m) $\alpha = 200$, $\varepsilon = 10^{-3}$ (n) $\alpha = 200$, $\varepsilon = 10^{-1}$ (o) $\alpha = 200$, $\varepsilon = 1$ (p) $\alpha = 200$, $\varepsilon = 100$

FIG S3. *Comparison of the Gaussian approximation for* $\sigma = 0.5$ *for different network sizes* $\alpha = 5, 10, 100, 200$ *and different thresholds* $\varepsilon = 10^{-3}, 10^{-1}, 1, 100$.

K.2. Sensitivity on leapfrog steps $\mathcal{L}$

This section shows a sensitivity analysis on the impact of the number of leapfrog steps $\mathcal{L}$ on the MCMC sampler which we vary $\mathcal{L} = 1, 2, 10, 20$. Recall that the leapfrog steps is the number of steps in the HMC which updates the parameters of the nodes. We generate an undirected graph with $p = 2$ communities and parameters $\alpha = 200$, $\sigma = 0.2$, $\tau = 1$, $b_k = b = 0.5$, $a_k = a = 0.2$. The sampled graph has 1799 nodes and 13,416 edges. For the inference, we assume as before 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). The chains use 1,000,000 iterations each and a truncation level $\varepsilon = 10^{-3}$. We repeat the same procedure (i.e. run the sampler with these settings and 3 parallel chains) 4 times; one for each value of leapfrog steps $\mathcal{L}$. Note that here we show the results for one chain since the behaviour was very similar cross chains. The trace plot of the parameters is shown in Figure S8. As shown, the chains exhibit a good mixing and they target the same posterior distribution.

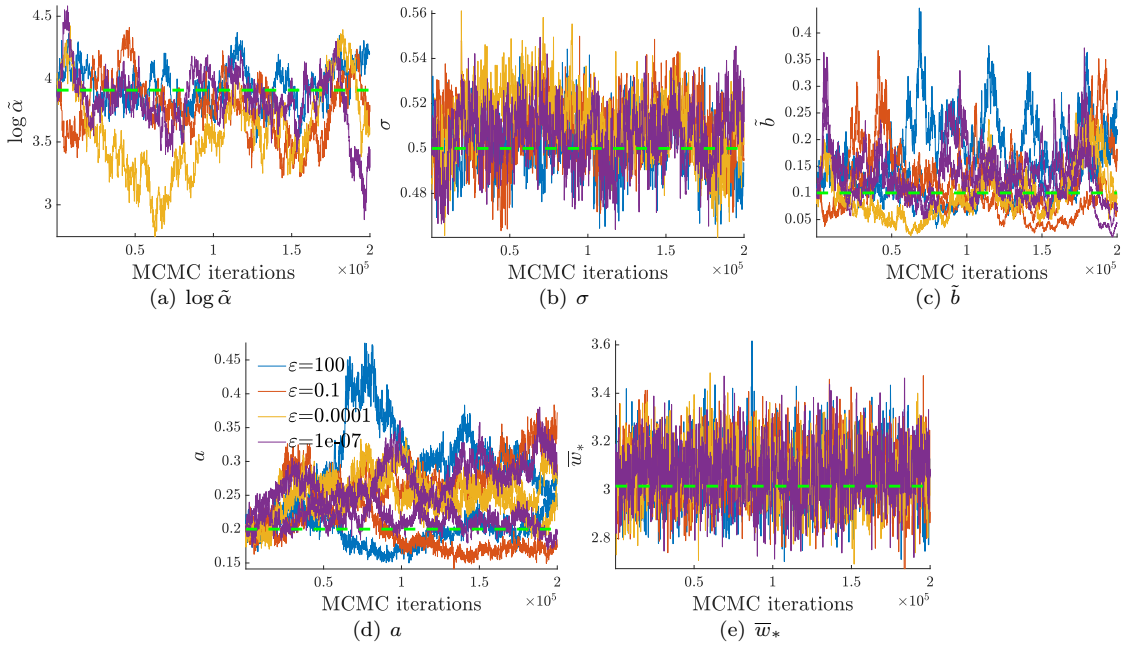


FIG S4. *MCMC trace plots for parameters (a) $\log\tilde{\alpha}$, (b) σ, (c) $\tilde{b}$ and (d) a and (e) $\overline{w}_*$ when performing inference with truncation levels $\varepsilon = 100, 0.1, 10^{-4}, 10^{-7}$. The graph is undirected and generated with $p = 2$ communities and parameters $\alpha = 50$, $\sigma = 0.5$, $\tau = 1$, $b_k = b = 0.1$, $a_k = a = 0.2$. The dotted green line, corresponds to the value of the model parameters used for generating the graph.*

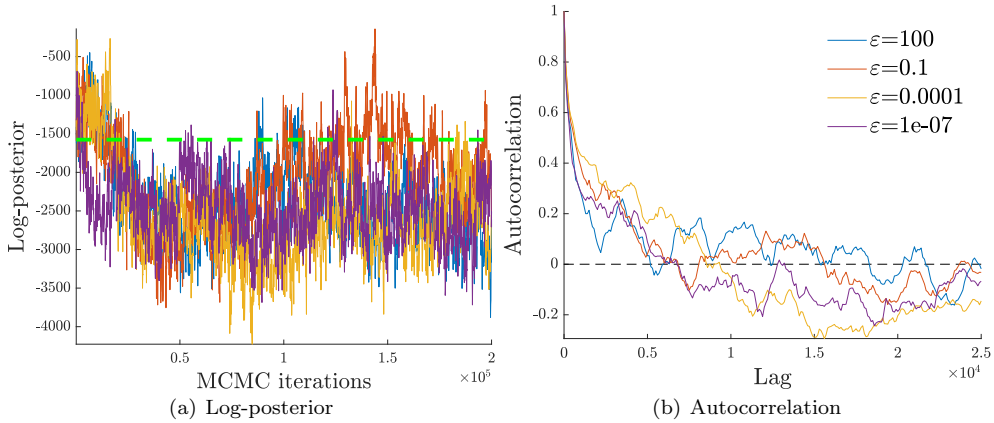


FIG S5. *Trace and autocorrelation of the approximate log-posterior probability density (up to a constant) when performing inference with truncation levels $\varepsilon = 100, 0.1, 10^{-4}, 10^{-7}$. The graph is undirected and generated with $p = 2$ communities and parameters $\alpha = 50$, $\sigma = 0.5$, $\tau = 1$, $b_k = b = 0.1$, $a_k = a = 0.2$. The dotted green line, corresponds to the value of the approximate log-posterior using the model parameters used for generating the graph.*

Clearly the parameters converge near the samples values used for generating the graph. Figure S9 shows the trace and autocorrelation plot of the log-posterior probability distribution. Unsurprisingly, chains need more time to converge when using 1 or 2 leapfrog steps compared to 10 and 20. However, as shown, there is no difference between 10 and 20, that is why we decide to use 10 as our default value and we also set it so in our experiments with real-world networks.

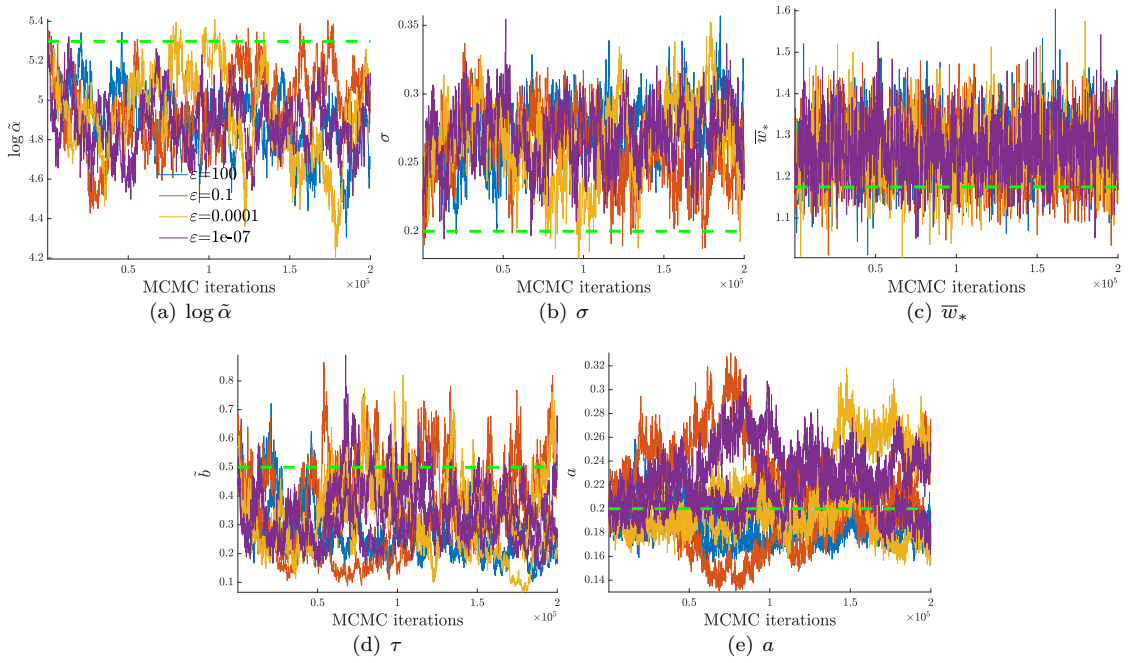


FIG S6. *MCMC trace plots for parameters (a)* $\log\tilde{\alpha}$*, (b)* σ*, (c)* $\tilde{b}$*, (d)* a *and (e)* $\overline{w}_*$ *when performing inference with truncation levels* $\varepsilon = 100, 0.1, 10^{-4}, 10^{-7}$*. The graph is undirected and generated with* $p = 2$ *communities and parameters* $\alpha = 50$*,* $\sigma = 0.2$*,* $\tau = 1$*,* $b_k = b = 0.1$*,* $a_k = a = 0.2$*. The dotted green line, corresponds to the value of the model parameters used for generating the graph.*

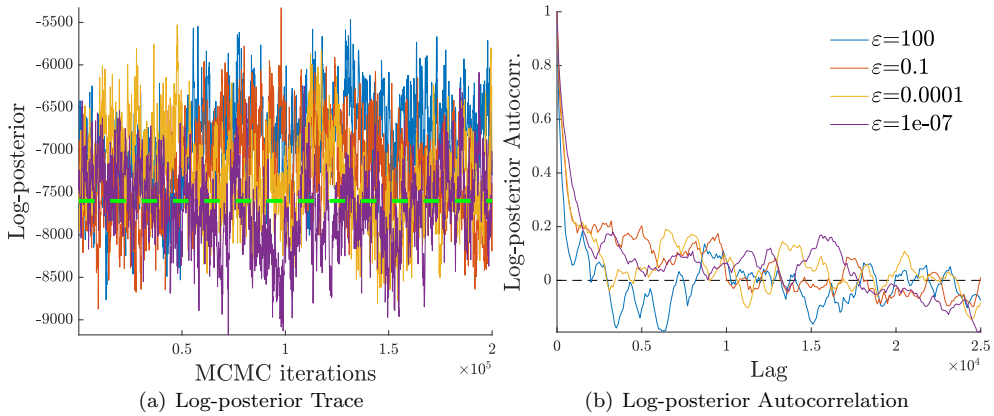


FIG S7. *Trace and autocorrelations of the approximate log-posterior probability density (up to a constant) when performing inference with truncation levels* $\varepsilon = 100, 0.1, 10^{-4}, 10^{-7}$*. The graph is undirected and generated with* $p = 2$ *communities and parameters* $\alpha = 50$*,* $\sigma = 0.2$*,* $\tau = 1$*,* $b_k = b = 0.1$*,* $a_k = a = 0.2$*. The dotted green line, corresponds to the value of the approximate log-posterior using the model parameters used for generating the graph.*

L. Experiments on finite-activity simulated data

In the main paper we only show results for the infinite-activity model, i.e. sparser graphs generated with positive values of σ. In this section we demonstrate the applicability of our methods to the case of denser graphs which correspond to finite-activity and equivalently to negative values of σ. In the dense regime

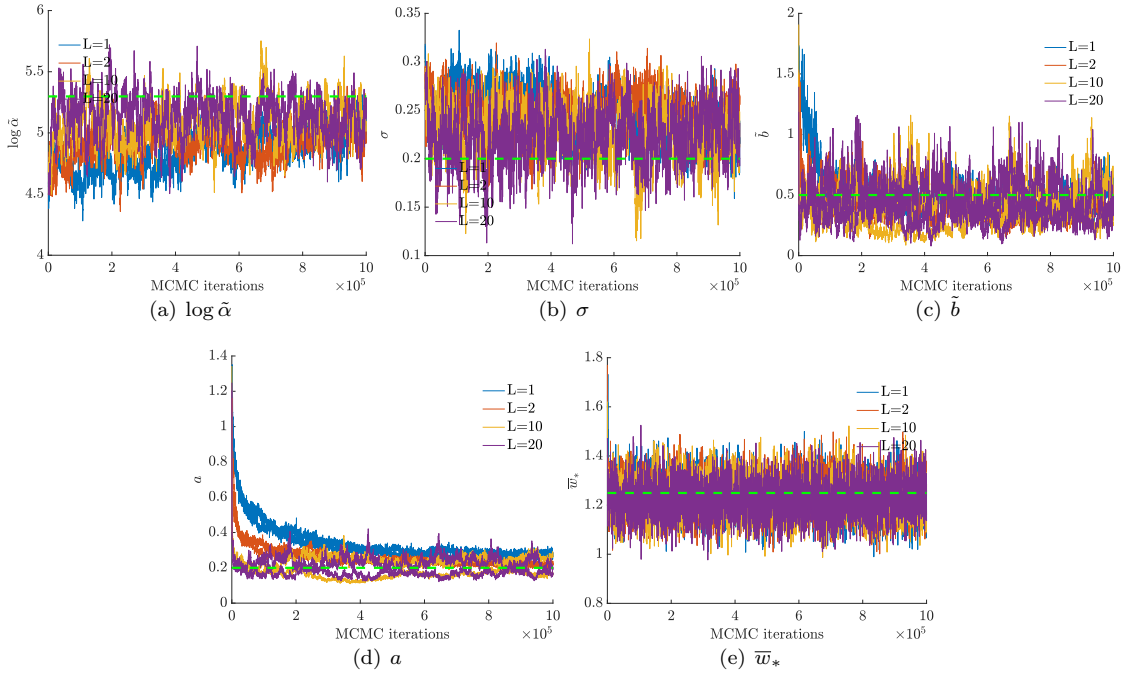


FIG S8. *MCMC trace plots for parameters (a)* $\log\tilde{\alpha}$*, (b)* σ*, (c)* $\tilde{b}$*, (d)* a *and (e)* $\overline{w}_*$ *when performing inference with number of leapfrog steps* $\mathcal{L} = 1, 2, 10, 20$*. The graph is undirected and generated with* $p = 2$ *communities and parameters* $\alpha = 200$*,* $\sigma = 0.2$*,* $\tau = 1$*,* $b_k = b = 0.5$*,* $a_k = a = 0.2$*. The dotted green line, corresponds to the value of the model parameters used for generating the graph.*

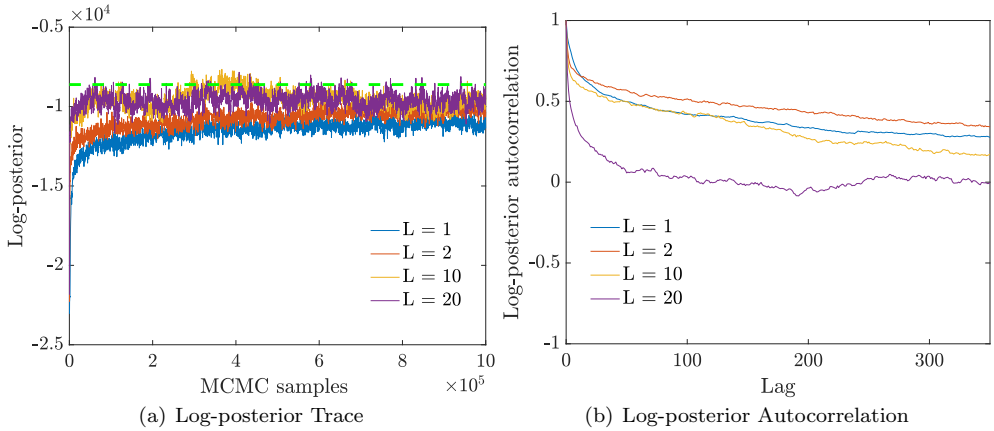


FIG S9. *Trace and autocorrelations of the approximate log-posterior probability density (up to a constant) when performing inference with number of leapfrog steps* $\mathcal{L} = 1, 2, 10, 20$*. The graph is undirected and generated with* $p = 2$ *communities and parameters* $\alpha = 200$*,* $\sigma = 0.2$*,* $\tau = 1$*,* $b_k = b = 0.5$*,* $a_k = a = 0.2$*. The dotted green line, corresponds to the value of the approximate log-posterior using the model parameters used for generating the graph.*

$(\sigma < 0)$, our model with mean measure

$$\nu(w_0, \beta_{1:p}) = \frac{\alpha}{\Gamma(1-\sigma)} w_0^{-1-\sigma} e^{-\tau} \prod_{k=1}^{p} \text{Gamma}(\beta_k; a_k, b_k),$$

has a finite number of atoms $K_\alpha \geq N_\alpha$. These atoms have iid weights $(w_{0i}, \beta_{i,1:p})_{i=1,\ldots,K_\alpha}$ given K_α, where $w_{0i} \sim \text{Gamma}(-\sigma, \tau)$ and $\beta_k \sim \text{Gamma}(a_k, b_k)$. The following parameters are more interpretable in the

dense regime. Firstly, the number of atoms K_α has mean

$$\varsigma_1 = \int \nu(w_0, \beta_{1:p}) = \frac{-\alpha\tau^\sigma}{\sigma}. \tag{S29}$$

The mean weights are, for $k = 1, \ldots, p$

$$\varsigma_{2,k} = \mathbb{E}[w_{0i}\beta_{ik}] = \frac{-\sigma}{\tau}\frac{a_k}{b_k}. \tag{S30}$$

The variance of the weights is, for $k = 1, \ldots, p$

$$\varsigma_{3,k} = \mathbb{V}ar[w_{0i}\beta_{ik}] = \frac{a_k\sigma(\sigma - a_k - 1)}{\tau^2 b_k^2}. \tag{S31}$$

In all our experiments we use the same parameter values $b = b_k$, $a = a_k$ for all $k = 1 \ldots p$, and therefore we use the simpler notation of $\varsigma_2 = \varsigma_{2,k}$ and $\varsigma_3 = \varsigma_{3,k}$.

L.1. Finite-activity graph with $\sigma = -2$

We generate an undirected graph with $p = 2$ communities and parameters $\alpha = 3000$, $\sigma = -2$, $\tau = 1$, $b_k = b = 3$, $a_k = a = 0.2$. The sampled graph has 1318 nodes and 101,110 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 and use the exact same settings as in the simulations in Section 5.1 of the main paper. Figure S10 gives the trace plots of the identifiable parameters which are (a) ς_1, (b) ς_2, (c) ς_3 and (d) $\overline{w}_*$ where ς_1 is as given in Eq. (S29), $\varsigma_2 = \varsigma_{2,k}$ as in Eq. (S30) and $\varsigma_3 = \varsigma_{3,k}$, as in Eq. (S31). The dotted green line, corresponds to the value of the model parameters used for generating the graph. Figure S10(e)-(f) shows the trace and autocorrelation plots of the approximate log-posterior distribution (up to a normalizing constant). The dotted green line corresponds to the value of the approximate log-posterior computed using the model parameters.

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 S11(a)-(b). By generating 500 graphs from the posterior predictive we assess the model fit to the degree distribution and show in Figure S11(c) that our model's posterior predictive degree distribution fits well the empirical degree distribution of the dense generated graph.

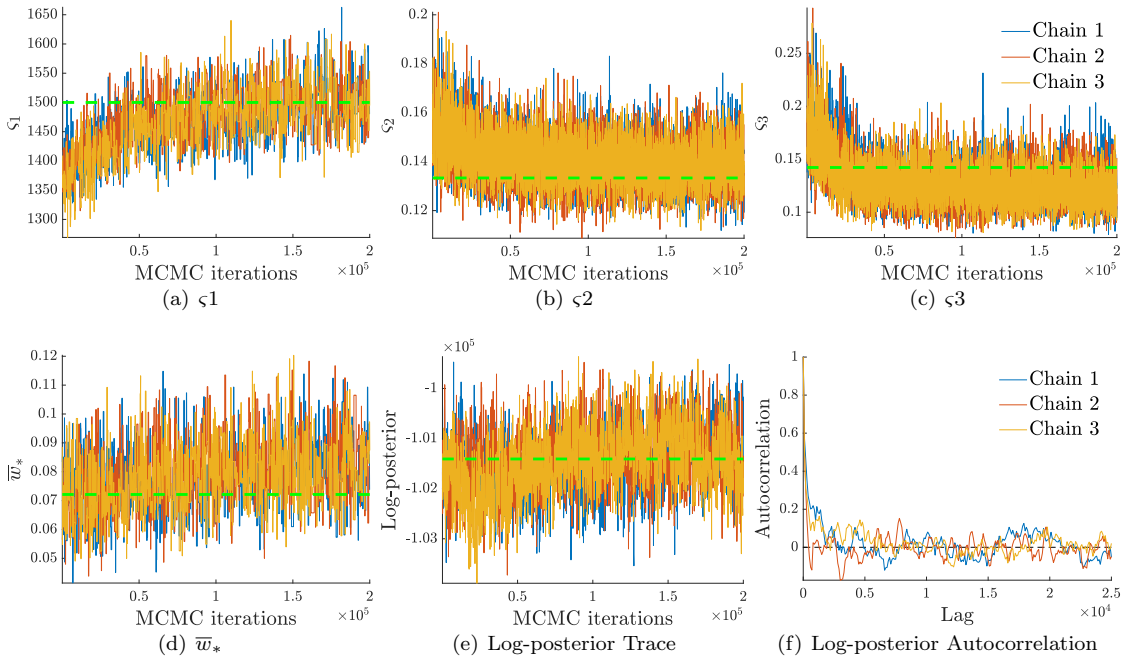
L.2. Finite-activity graph with $\sigma = -1$

This section repeats the same procedure as above but for $\sigma = -1$. We generate an undirected graph with $p = 2$ communities and parameters $\alpha = 1500$, $\sigma = -1$, $\tau = 1$, $b_k = b = 2$, $a_k = a = 0.2$. The sampled graph has 1219 nodes and 57,537 edges. We use the same priors and initialization procedure as in Section L.1 We give the trace plots of the identifiable parameters in Figure S12(a)-(d). The dotted green line as before corresponds to the value of the model parameters used for generating the graph. Posterior samples clearly converge around the sampled value. Figure S12(e)-(f) shows the trace and autocorrelation plots of the log-posterior 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. The log-posterior trace indicates convergence of the chains and the autocorrelation shows that the correlation of samples decreases quickly with increasing lag.

In Figure S13(a)-(b) we give the credible intervals of the mean parameters of both high and low degree nodes. The model is able to accurately recover the mean parameters. We also generate 500 graphs from the posterior predictive and show that our model fits the empirical degree distribution of the dense generated graph as shown in Figure S13(c).

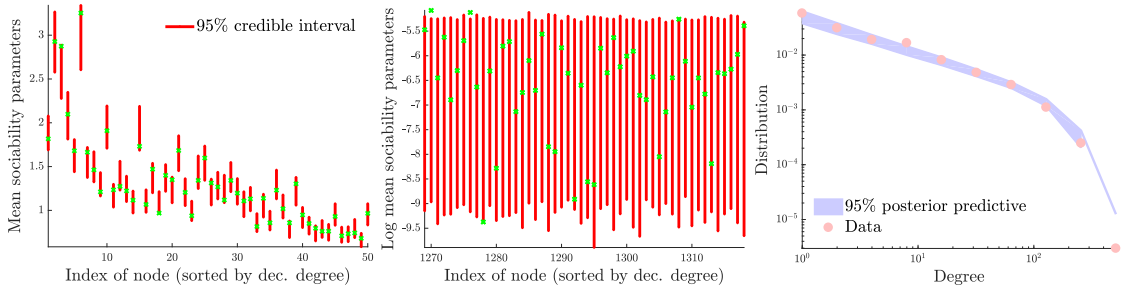
M. Additional plots for experiments on real-world data

In this section we provide additional MCMC trace plots for the two real world datasets experiments we perform in the main paper `USairport` and `polblogs`.



(a) $\varsigma 1$ (b) $\varsigma 2$ (c) $\varsigma 3$ (d) $\overline{w}_*$ (e) Log-posterior Trace (f) Log-posterior Autocorrelation

FIG S10. *MCMC trace plots of parameters(a)* $\varsigma 1$, *(b)* $\varsigma 2$, *(c)* $\varsigma 3$, *(d)* $\overline{w}_*$, *(e) trace plot of the log-posterior, (f) autocorrelation of the log-posterior for a graph generated with parameters* $p = 2$, $\alpha = 3000$, $\sigma = -2$, $\tau = 1$, $b = 3$, $a = 0.2$. *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.*

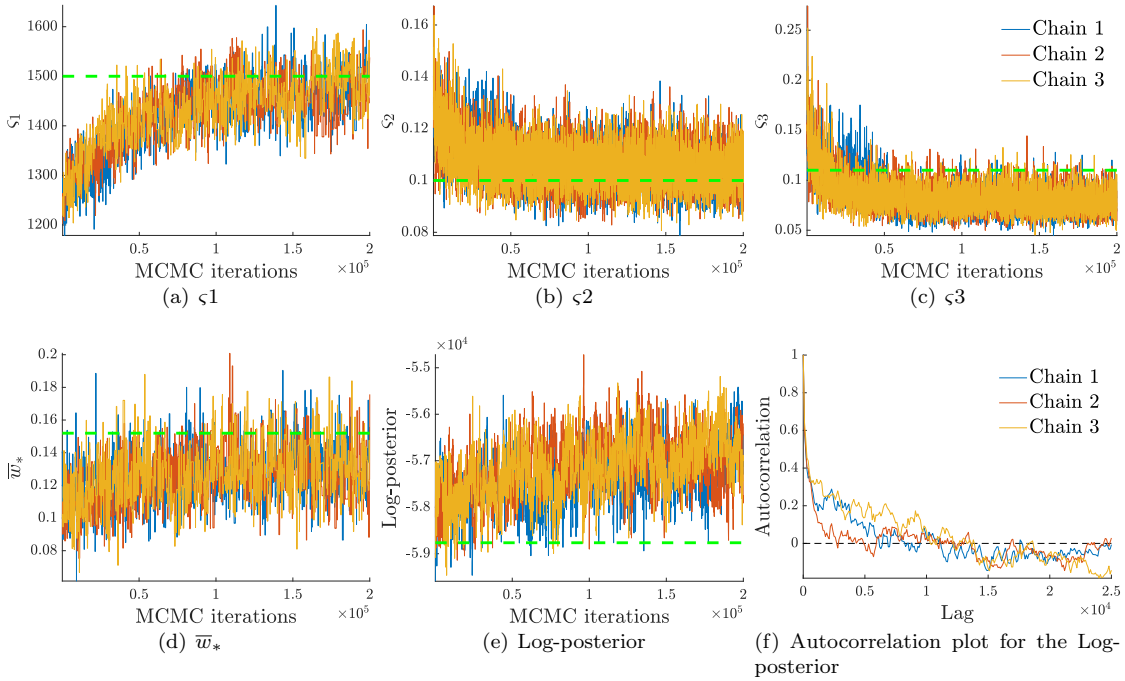


(a) 50 nodes with highest degrees (b) 50 nodes with lowest degrees (c) Degree distribution

FIG S11. *95% posterior credible intervals and true values 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 and 95% posterior predictive credible interval. Results obtained for a graph generated with parameters* $p = 2$, $\alpha = 3000$, $\sigma = -2$, $\tau = 1$, $b = 3$, $a = 0.2$.

M.1. Polblogs

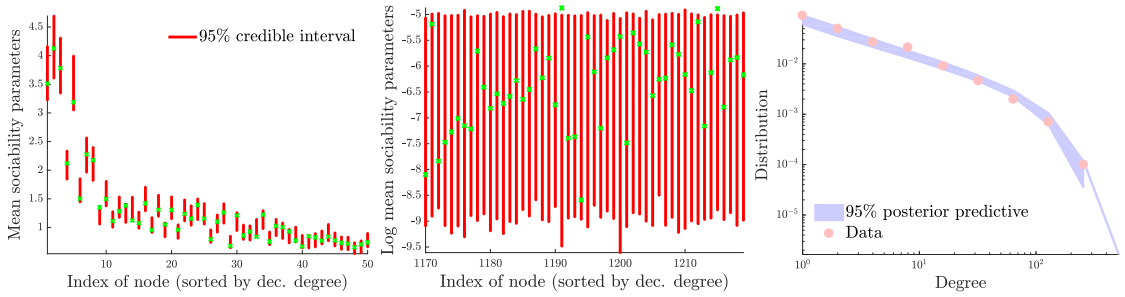
For the `polblogs` dataset we provide trace plots for the identifiable parameters $\varsigma 1$, $\varsigma 2$, $\varsigma 3$ as given above in section L in Eq. (S29), (S30) and (S31). As shown in the trace plots, the MCMC chains exhibit a good mixing and the graphs suggest convergence of the sampler. Similarly, the approximate log-posterior probability distribution, shown in Figure 11 of the main paper also demonstrates convergence of the sampler.



(a) $\varsigma 1$ (b) $\varsigma 2$ (c) $\varsigma 3$

(d) $\overline{w}_*$ (e) Log-posterior (f) Autocorrelation plot for the Log-posterior

FIG S12. *MCMC trace plots of parameters (a)* $\varsigma 1$*, (b)* $\varsigma 2$*, (c)* $\varsigma 3$*, (d)* $\overline{w}_*$*, (e) trace plot of the log-posterior, (f) autocorrelation of the log-posterior for a graph generated with parameters* $p = 2$*,* $\alpha = 1500$*,* $\sigma = -1$*,* $\tau = 1$*,* $b = 2$*,* $a = 0.2$*. 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.*



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

FIG S13. *95% posterior credible intervals and true values 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 and 95% posterior predictive credible interval. Results obtained for a graph generated with parameters* $p = 2$*,* $\alpha = 1500$*,* $\sigma = -1$*,* $\tau = 1$*,* $b = 2$*,* $a = 0.2$*.*

M.2. USairport

For the `USairport` dataset we provide trace plots for the parameters $\log \tilde{\alpha}$, σ, $\tilde{b}$, a and $\overline{w}_*$ shown in Figure S15(a)-(e). Figure S15 suggests that the posterior distribution is multimodal as one chain seems to be in a different mode than the other two. We ran longer chains, but this did not lead to improvement in the mixing for this dataset, and we acknowledge that any credible interval for the hyperparameters on this dataset should be considered with cautiousness. Note however that the two modes correspond to parameters with similar approximate log-posterior density (see Figure 11 in the main paper). Similarly, the credible intervals for the sociability parameters, community structure and posterior predictive degree distribution for each individual

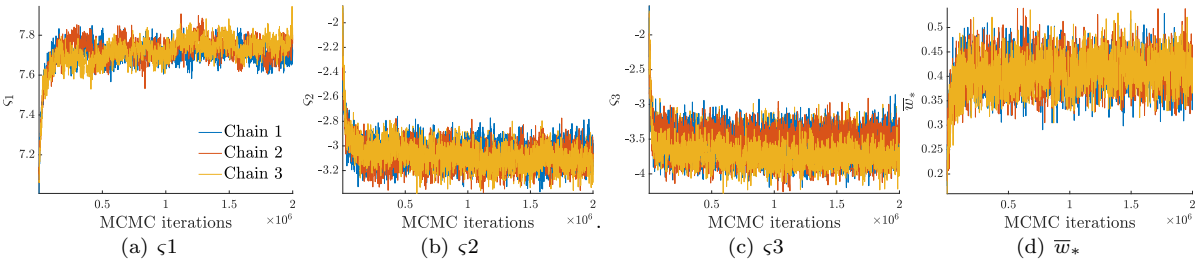


FIG S14. *MCMC trace plots of parameters (a)* ς1*, (b)* ς2*, (c)* ς3 *and (d)* $\overline{w}_*$ *for* `polblogs`.

chain are again quantitatively very close to those given in the main paper when considering all the chains.

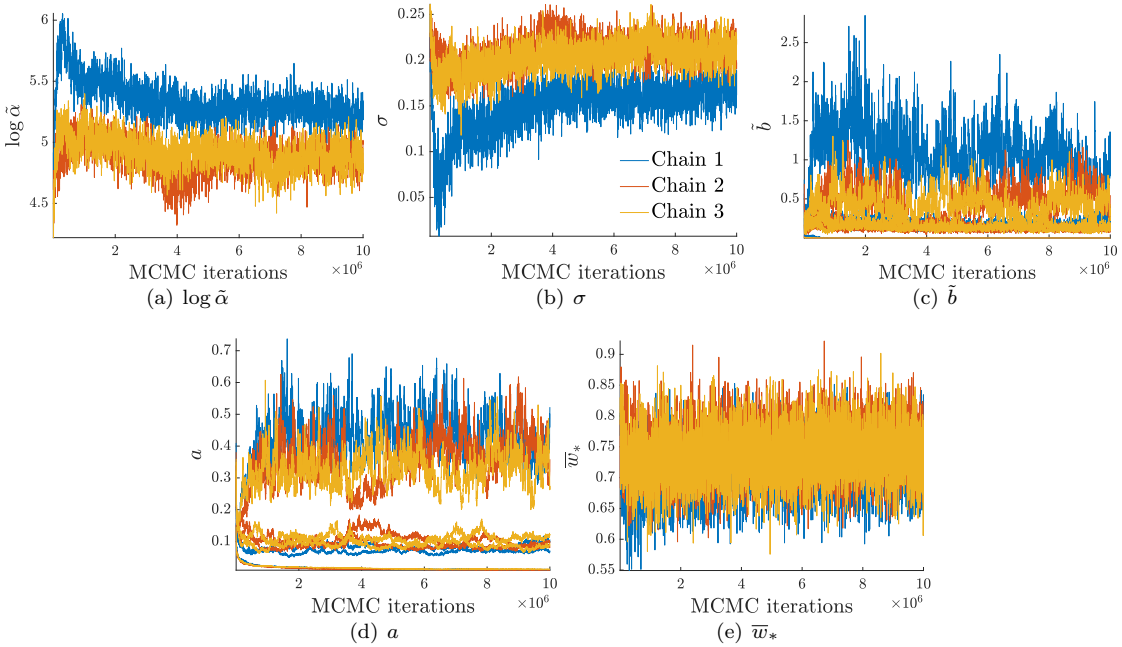


FIG S15. *MCMC trace plots of parameters (a)* $\log \tilde{\alpha}$*, (b)* σ*, (c)* $\tilde{b}$*, (d)* a *and (e)* $\overline{w}_*$ *for* `USairport`.

References

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.

O. E. Barndorff-Nielsen, J. Pedersen, and K.-I. Sato. Multivariate Subordination, Self-Decomposability and Stability. *Advances in Applied Probability*, 33:160–187, 2001.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27. Cambridge university press, 1989.

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.

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 II: General Theory and Structure*. Springer Verlag, 2nd edition edition, 2008.

I. Epifani and A. Lijoi. NONPARAMETRIC PRIORS FOR VECTORS OF SURVIVAL FUNCTIONS. *Statistica Sinica*, 20:1455–1484, 2010.

S. Favaro and Y.W. Teh. MCMC for Normalized Random Measure Mixture Models. *Statistical Science*, 28 (3):335–359, 2013.

A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv*, 4(146-171):88, 2007.

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 (Statistical Methodology)*, 75 (3):499–529, 2013.

J. Kallsen and P. Tankov. Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis*, 97(7):1551–1572, 2006.

J.F.C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

J.F.C. Kingman. *Poisson Processes*. Oxford University Press, USA, 1993.

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.

P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S.G. Walker, editors, *Bayesian nonparametrics*, volume 28, page 80. Camb. Ser. Stat. Probab. Math, 2010.

A. Lijoi, B. Nipoti, and I. Prünster. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291, 2014.

J. Møller and R. Waagepetersen. *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27 (1):23–31, 1981.

S. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer, 2013.

J. Rosiński. Tempering stable processes. *Stochastic processes and their applications*, 117(6):677–707, 2007.

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.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities
Publication Status	□Published ☑ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	To appear in Series B of the Journal of Royal Statistical Society.

Student Confirmation

Student Name:	Xenia Miscouridou
Contribution to the Paper	I have contributed in (a) theory: deriving exact and asymptotic graph properties (b) methodology: assessment of the methodology, accuracy of the Gaussian approximation part (c) experiments: extensive comparisons on real datasets between our model and alternative ones, considering several aspects and model fit diagnostics.

Signature  Date 05/01/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor François CARON

Supervisor comments

—

Signature  Date 08/01/2020

Chapter 3

Modeling Sparsity, Heterogeneity, Reciprocity and Community Structure in Temporal Interaction Data

Published at Neural Information Processing Systems 2018

Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data

Xenia Miscouridou[1], **François Caron**[1], **Yee Whye Teh**[1,2]
[1]Department of Statistics, University of Oxford
[2]DeepMind
{miscouri, caron, y.w.teh}@stats.ox.ac.uk

Abstract

We propose a novel class of network models for temporal dyadic interaction data. Our objective is to capture important features often observed in social interactions: sparsity, degree heterogeneity, community structure and reciprocity. We use mutually-exciting Hawkes processes to model the interactions between each (directed) pair of individuals. The intensity of each process allows interactions to arise as responses to opposite interactions (reciprocity), or due to shared interests between individuals (community structure). For sparsity and degree heterogeneity, we build the non time dependent part of the intensity function on compound random measures following (Todeschini et al., 2016). We conduct experiments on real-world temporal interaction data and show that the proposed model outperforms competing approaches for link prediction, and leads to interpretable parameters.

1 Introduction

There is a growing interest in modelling and understanding temporal dyadic interaction data. Temporal interaction data take the form of time-stamped triples (t, i, j) indicating that an interaction occurred between individuals i and j at time t. Interactions may be directed or undirected. Examples of such interaction data include commenting a post on an online social network, exchanging an email, or meeting in a coffee shop. An important challenge is to understand the underlying structure that underpins these interactions. To do so, it is important to develop statistical network models with interpretable parameters, that capture the properties which are observed in real social interaction data.

One important aspect to capture is the *community structure* of the interactions. Individuals are often affiliated to some latent communities (e.g. work, sport, etc.), and their affiliations determine their interactions: they are more likely to interact with individuals sharing the same interests than to individuals affiliated with different communities. An other important aspect is *reciprocity*. Many events are responses to recent events of the opposite direction. For example, if Helen sends an email to Mary, then Mary is more likely to send an email to Helen shortly afterwards. A number of papers have proposed statistical models to capture both community structure and reciprocity in temporal interaction data (Blundell et al., 2012; Dubois et al., 2013; Linderman and Adams, 2014). They use models based on Hawkes processes for capturing reciprocity and stochastic block-models or latent feature models for capturing community structure.

In addition to the above two properties, it is important to capture the global properties of the interaction data. Interaction data are often *sparse*: only a small fraction of the pairs of nodes actually interact. Additionally, they typically exhibit high degree (number of interactions per node) *heterogeneity*: some individuals have a large number of interactions, whereas most individuals have very few, therefore resulting in empirical degree distributions being heavy-tailed. As shown by Karrer and Newman (2011), Gopalan et al. (2013) and Todeschini et al. (2016), failing to account explicitly for

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

degree heterogeneity in the model can have devastating consequences on the estimation of the latent structure.

Recently, two classes of statistical models, based on random measures, have been proposed to capture sparsity and power-law degree distribution in network data. The first one is the class of models based on exchangeable random measures (Caron and Fox, 2017; Veitch and Roy, 2015; Herlau et al., 2016; Borgs et al., 2018; Todeschini et al., 2016; Palla et al., 2016; Janson, 2017a). The second one is the class of edge-exchangeable models (Crane and Dempsey, 2015; 2018; Cai et al., 2016; Williamson, 2016; Janson, 2017b; Ng and Silva, 2017). Both classes of models can handle both sparse and dense networks and, although the two constructions are different, connections have been highlighted between the two approaches (Cai et al., 2016; Janson, 2017b).

The objective of this paper is to propose a class of statistical models for temporal dyadic interaction data that can capture all the desired properties mentioned above, which are often found in real world interactions. These are *sparsity*, *degree heterogeneity*, *community structure* and *reciprocity*. Combining all the properties in a single model is non trivial and there is no such construction to our knowledge. The proposed model generalises existing reciprocating relationships models (Blundell et al., 2012) to the sparse and power-law regime. Our model can also be seen as a natural extension of the classes of models based on exchangeable random measures and edge-exchangeable models and it shares properties of both families. The approach is shown to outperform alternative models for link prediction on a variety of temporal network datasets.

The construction is based on Hawkes processes and the (static) model of Todeschini et al. (2016) for sparse and modular graphs with overlapping community structure. In Section 2, we present Hawkes processes and compound completely random measures which form the basis of our model's construction. The statistical model for temporal dyadic data is presented in Section 3 and its properties derived in Section 4. The inference algorithm is described in Section 5. Section 6 presents experiments on four real-world temporal interaction datasets.

2 Background material

2.1 Hawkes processes

Let $(t_k)_{k\geq 1}$ be a sequence of event times with $t_k \geq 0$, and let $\mathcal{H}_t = (t_k | t_k \leq t)$ the subset of event times between time 0 and time t. Let $N(t) = \sum_{k\geq 1} 1_{t_k \leq t}$ denote the number of events between time 0 and time t, where $1_A = 1$ if A is true, and 0 otherwise. Assume that $N(t)$ is a counting process with conditional intensity function $\lambda(t)$, that is for any $t \geq 0$ and any infinitesimal interval dt

$$\Pr(N(t+dt) - N(t) = 1 | \mathcal{H}_t) = \lambda(t)dt. \tag{1}$$

Consider another counting process $\tilde{N}(t)$ with the corresponding $(\tilde{t}_k)_{k\geq 1}, \tilde{\mathcal{H}}_t, \tilde{\lambda}(t)$. Then, $N(t), \tilde{N}(t)$ are mutually-exciting Hawkes processes (Hawkes, 1971) if the conditional intensity functions $\lambda(t)$ and $\tilde{\lambda}(t)$ take the form

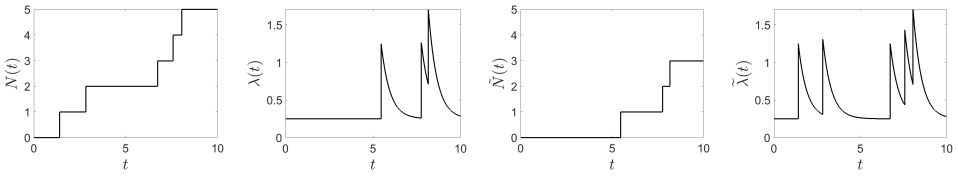
$$\lambda(t) = \mu + \int_0^t g_\phi(t-u)\, d\tilde{N}(u) \qquad \tilde{\lambda}(t) = \tilde{\mu} + \int_0^t g_{\tilde{\phi}}(t-u)\, dN(u)$$

where $\mu = \lambda(0) > 0, \tilde{\mu} = \tilde{\lambda}(0) > 0$ are the base intensities and $g_\phi, g_{\tilde{\phi}}$ non-negative kernels parameterised by ϕ and $\tilde{\phi}$. This defines a pair of processes in which the current rate of events of each process depends on the occurrence of past events of the opposite process.

Assume that $\mu = \tilde{\mu}$, $\phi = \tilde{\phi}$ and $g_\phi(t) \geq 0$ for $t > 0$, $g_\phi(t) = 0$ for $t < 0$. If g_ϕ admits a form of fast decay then this results in strong local effects. However, if it prescribes a peak away from the origin then longer term effects are likely to occur. We consider here an exponential kernel

$$g_\phi(t-u) = \eta e^{-\delta(t-u)}, t > u \tag{2}$$

where $\phi = (\eta, \delta)$. $\eta \geq 0$ determines the sizes of the self-excited jumps and $\delta > 0$ is the constant rate of exponential decay. The stationarity condition for the processes is $\eta < \delta$. Figure 1 gives an illustration of two mutually-exciting Hawkes processes with exponential kernel and their conditional intensities.



(a) (b) (c) (d)

Figure 1: Illustration of mutually-exciting Hawkes processes with exponential kernel with parameters $\mu = 0.25$, $\eta = 1$ and $\delta = 2$. (a) Counting process $N(t)$ and (b) its conditional intensity $\lambda(t)$. (c) Counting process $\tilde{N}(t)$ and its conditional intensity $\tilde{\lambda}(t)$.

2.2 Compound completely random measures

A homogeneous completely random measure (CRM) (Kingman, 1967; 1993) on $\mathbb{R}_+$ without fixed atoms nor deterministic component takes the form

$$W = \sum_{i \geq 1} w_i \delta_{\theta_i} \tag{3}$$

where $(w_i, \theta_i)_{i \geq 1}$ are the points of a Poisson process on $(0, \infty) \times \mathbb{R}_+$ with mean measure $\rho(dw)H(d\theta)$ where ρ is a Lévy measure, H is a locally bounded measure and δ_x is the dirac delta mass at x. The homogeneous CRM is completely characterized by ρ and H, and we write $W \sim \mathrm{CRM}(\rho, H)$, or simply $W \sim \mathrm{CRM}(\rho)$ when H is taken to be the Lebesgue measure. Griffin and Leisen (2017) proposed a multivariate generalisation of CRMs, called compound CRM (CCRM). A compound CRM $(W_1, \ldots, W_p)$ with independent scores is defined as

$$W_k = \sum_{i \geq 1} w_{ik} \delta_{\theta_i}, \;\; k = 1, \ldots, p \tag{4}$$

where $w_{ik} = \beta_{ik} w_{i0}$ and the scores $\beta_{ik} \geq 0$ are independently distributed from some probability distribution F_k and $W_0 = \sum_{i \geq 1} w_{i0} \delta_{\theta_i}$ is a CRM with mean measure $\rho_0(dw_0)H_0(d\theta)$. In the rest of this paper, we assume that F_k is a gamma distribution with parameters (a_k, b_k), $H_0(d\theta) = d\theta$ is the Lebesgue measure and ρ_0 is the Lévy measure of a generalized gamma process

$$\rho_0(dw) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\tau w} dw \tag{5}$$

where $\sigma \in (-\infty, 1)$ and $\tau > 0$.

3 Statistical model for temporal interaction data

Consider temporal interaction data of the form $\mathcal{D} = (t_k, i_k, j_k)_{k \geq 1}$ where $(t_k, i_k, j_k) \in \mathbb{R}_+ \times \mathbb{N}_*^2$ represents a directed interaction at time t_k from node/individual i_k to node/individual j_k. For example, the data may correspond to the exchange of messages between students on an online social network.

We use a point process $(t_k, U_k, V_k)_{k \geq 1}$ on $\mathbb{R}_+^3$, and consider that each node i is assigned some continuous label $\theta_i \geq 0$. the labels are only used for the model construction, similarly to (Caron and Fox, 2017; Todeschini et al., 2016), and are not observed nor inferred from the data. A point at location (t_k, U_k, V_k) indicates that there is a directed interaction between the nodes U_k and V_k at time t_k. See Figure 2 for an illustration.

For a pair of nodes i and j, with labels θ_i and θ_j, let $N_{ij}(t)$ be the counting process

$$N_{ij}(t) = \sum_{k | (U_k, V_k) = (\theta_i, \theta_j)} 1_{t_k \leq t} \tag{6}$$

for the number of interactions between i and j in the time interval $[0, t]$. For each pair i, j, the counting processes N_{ij}, N_{ji} are mutually-exciting Hawkes processes with conditional intensities

$$\lambda_{ij}(t) = \mu_{ij} + \int_0^t g_\phi(t-u)\, dN_{ji}(u), \qquad \lambda_{ji}(t) = \mu_{ji} + \int_0^t g_\phi(t-u)\, dN_{ij}(u) \tag{7}$$

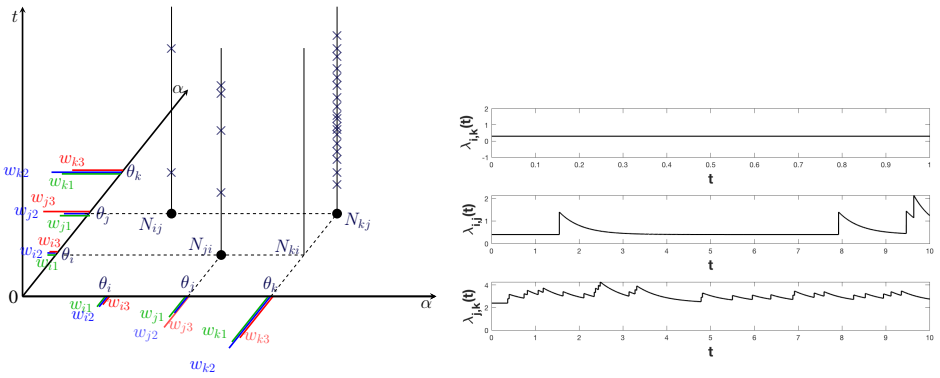


Figure 2: (Left) Representation of the temporal dyadic interaction data as a point process on $\mathbb{R}^3_+$. A point at location $(\tau, \theta_i, \theta_j)$ indicates a directed interaction from node i to node j at time τ, where $\theta_i > 0$ and $\theta_j > 0$ are labels of nodes i and j. Interactions between nodes i and j, arise from a Hawkes process N_{ij} with conditional intensity λ_{ij} given by Equation (7). Processes N_{ik} and N_{jk} are not shown for readability. (Right) Conditional intensities of processes $N_{ik}(t)$, $N_{ij}(t)$ and $N_{jk}(t)$.

where g_ϕ is the exponential kernel defined in Equation (2). Interactions from individual i to individual j may arise as a response to past interactions from j to i through the kernel g_ϕ, or via the base intensity μ_{ij}. We also model assortativity so that individuals with similar interests are more likely to interact than individuals with different interests. For this, assume that each node i has a set of positive latent parameters $(w_{i1}, \ldots, w_{ip}) \in \mathbb{R}^p_+$, where w_{ik} is the level of its affiliation to each latent community $k = 1, \ldots, p$. The number of communities p is assumed known. We model the base rate

$$\mu_{ij} = \mu_{ji} = \sum_{k=1}^{p} w_{ik} w_{jk}. \tag{8}$$

Two nodes with high levels of affiliation to the same communities will be more likely to interact than nodes with affiliation to different communities, favouring assortativity.

In order to capture sparsity and power-law properties and as in Todeschini et al. (2016), the set of affiliation parameters $(w_{i1}, \ldots, w_{ip})$ and node labels θ_i is modelled via a compound CRM with gamma scores, that is $W_0 = \sum_{i=1}^{\infty} w_{i0} \delta_{\theta_i} \sim \text{CRM}(\rho_0)$ where the Lévy measure ρ_0 is defined by Equation (5), and for each node $i \geq 1$ and community $k = 1, \ldots, p$

$$w_{ik} = w_{i0} \beta_{ik}, \text{ where } \beta_{ik} \overset{\text{ind}}{\sim} \text{Gamma}(a_k, b_k). \tag{9}$$

The parameter $w_{i0} \geq 0$ is a degree correction for node i and can be interpreted as measuring the overall popularity/sociability of a given node i irrespective of its level of affiliation to the different communities. An individual i with a high sociability parameter w_{i0} will be more likely to have interactions overall than individuals with low sociability parameters. The scores β_{ik} tune the level of affiliation of individual i to the community k. The model is defined on $\mathbb{R}^3_+$. We assume that we observe interactions over a subset $[0, T] \times [0, \alpha]^2 \subseteq \mathbb{R}^3_+$ where α and T tune both the number of nodes and number of interactions. The whole model is illustrated in Figure 2.

The model admits the following set of hyperparameters, with the following interpretation:
• The hyperparameters $\phi = (\eta, \delta)$ where $\eta \geq 0$ and $\delta \geq 0$ of the kernel g_ϕ tune the *reciprocity*.
• The hyperparameters (a_k, b_k) tune the *community structure* of the interactions. $a_k / b_k = \mathbb{E}[\beta_{ik}]$ tunes the size of community k while $a_k / b_k^2 = \text{var}(\beta_{ik})$ tunes the variability of the level of affiliation to this community; larger values imply more separated communities.
• The hyperparameter σ tunes the *sparsity* and the *degree heterogeneity*: larger values imply higher sparsity and heterogeneity. It also tunes the slope of the degree distribution. Parameter τ tunes the exponential cut-off in the degree distribution. This is illustrated in Figure 3.
• Finally, the hyperparameters α and T tune the overall number of interactions and nodes.

We follow (Rasmussen, 2013) and use vague Exponential(0.01) priors on η and δ. Following Todeschini et al. (2016) we set vague Gamma$(0.01, 0.01)$ priors on α, $1 - \sigma$, τ, a_k and b_k. The right limit for time window, T is considered known.

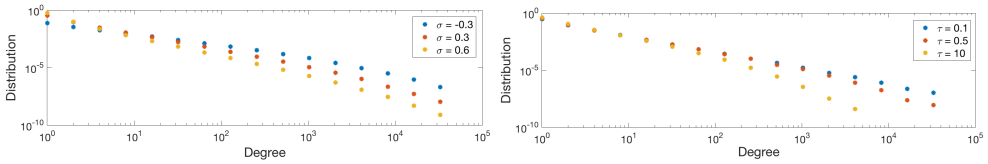


Figure 3: Degree distribution for a Hawkes graph with different values of σ (left) and τ (right). The degree of a node i is defined as the number of nodes with whom it has at least one interaction. The value σ tunes the slope of the degree distribution, larger values corresponding to a higher slope. The parameter τ tunes the exponential cut-off in the degree distribution.

4 Properties

4.1 Connection to sparse vertex-exchangeable and edge-exchangeable models

The model is a natural extension of sparse vertex-exchangeable and edge-exchangeable graph models. Let $z_{ij}(t) = 1_{N_{ij}(t)+N_{ji}(t)>0}$ be a binary variable indicating if there is at least one interaction in $[0, t]$ between nodes i and j in either direction. We assume

$$\Pr(z_{ij}(t) = 1|(w_{ik}, w_{jk})_{k=1,\dots,p}) = 1 - e^{-2t\sum_{k=1}^{p} w_{ik}w_{jk}}$$

which corresponds to the probability of a connection in the static simple graph model proposed by Todeschini et al. (2016). Additionally, for fixed $\alpha > 0$ and $\eta = 0$ (no reciprocal relationships), the model corresponds to a rank-p extension of the rank-1 Poissonized version of edge-exchangeable models considered by Cai et al. (2016) and Janson (2017a). The sparsity properties of our model follow from the sparsity properties of these two classes of models.

4.2 Sparsity

The size of the dataset is tuned by both α and T. Given these quantities, both the number of interactions and the number of nodes with at least one interaction are random variables. We now study the behaviour of these quantities, showing that the model exhibits sparsity. Let $I_{\alpha,T}, E_{\alpha,T}, V_{\alpha,T}$ be the overall number of interactions between nodes with label $\theta_i \leq \alpha$ until time T, the total number of pairs of nodes with label $\theta_i \leq \alpha$ who had at least one interaction before time T, and the number of nodes with label $\theta_i \leq \alpha$ who had at least one interaction before time T respectively.

$$V_{\alpha,T} = \sum_i 1_{\sum_{j\neq i} N_{ij}(T)1_{\theta_j\leq\alpha}>0} 1_{\theta_i\leq\alpha} \qquad I_{\alpha,T} = \sum_{i\neq j} N_{ij}(T)1_{\theta_i\leq\alpha}1_{\theta_j\leq\alpha}$$
$$E_{\alpha,T} = \sum_{i<j} 1_{N_{ij}(T)+N_{ji}(T)>0} 1_{\theta_i\leq\alpha}1_{\theta_j\leq\alpha}$$

We provide in the supplementary material a theorem for the exact expectations of $I_{\alpha,T}$, $E_{\alpha,T}$ $V_{\alpha,T}$. Now consider the asymptotic behaviour of the expectations of $V_{\alpha,T}$, $E_{\alpha,T}$ and $I_{\alpha,T}$, as α and T go to infinity.[1] Consider fixed $T > 0$ and α that tends to infinity. Then,

$$\mathbb{E}[V_{\alpha,T}] = \begin{cases} \Theta(\alpha) & \text{if } \sigma < 0 \\ \omega(\alpha) & \text{if } \sigma \geq 0 \end{cases}, \qquad \mathbb{E}[E_{\alpha,T}] = \Theta(\alpha^2), \qquad \mathbb{E}[I_{\alpha,T}] = \Theta(\alpha^2)$$

as α tends to infinity. For $\sigma < 0$, the number of edges and interactions grows quadratically with the number of nodes, and we are in the dense regime. When $\sigma \geq 0$, the number of edges and interaction grows subquadratically, and we are in the sparse regime. Higher values of σ lead to higher sparsity. For fixed α,

$$\mathbb{E}[V_{\alpha,T}] = \begin{cases} \Theta(1) & \text{if } \sigma < 0 \\ \Theta(\log T) & \text{if } \sigma = 0 \\ \Theta(T^{\sigma}) & \text{if } \sigma > 0 \end{cases}, \qquad \mathbb{E}[E_{\alpha,T}] = \begin{cases} \Theta(1) & \text{if } \sigma < 0 \\ O(\log T) & \text{if } \sigma = 0 \\ O(T^{3\sigma/2}) & \text{if } \sigma \in (0, 1/2] \\ O(T^{(1+\sigma)/2}) & \text{if } \sigma \in (1/2, 1) \end{cases}$$

[1] We use the following asymptotic notations. $X_\alpha = O(Y_\alpha)$ if $\lim X_\alpha/Y_\alpha < \infty$, $X_\alpha = \omega(Y_\alpha)$ if $\lim Y_\alpha/X_\alpha = 0$ and $X_\alpha = \Theta(Y_\alpha)$ if both $X_\alpha = O(Y_\alpha)$ and $Y_\alpha = O(X_\alpha)$.

and $\mathbb{E}[I_{\alpha,T}] = \Theta(T)$ as T tends to infinity. Sparsity in T arises when $\sigma \geq 0$ for the number of edges and when $\sigma > 1/2$ for the number of interactions. The derivation of the asymptotic behaviour of expectations of $V_{\alpha,T}$, $E_{\alpha,T}$ and $I_{\alpha,T}$ follows the lines of the proofs of Theorems 3 and 5.3 in (Todeschini et al., 2016) ($\alpha \to \infty$) and Lemma D.6 in the supplementary material of (Cai et al., 2016) ($T \to \infty$), and is omitted here.

5 Approximate Posterior Inference

Assume a set of observed interactions $\mathcal{D} = (t_k, i_k, j_k)_{k\geq 1}$ between V individuals over a period of time T. The objective is to approximate the posterior distribution $\pi(\phi, \xi|\mathcal{D})$ where ϕ are the kernel parameters and $\xi = ((w_{ik})_{i=1,\ldots,V,k=1,\ldots,p}, (a_k, b_k)_{k=1,\ldots,p}, \alpha, \sigma, \tau)$, the parameters and hyperparameters of the compound CRM. One possible approach is to follow a similar approach to that taken in (Rasmussen, 2013); derive a Gibbs sampler using a data augmentation scheme which associates a latent variable to each interaction. However, such an algorithm is unlikely to scale well with the number of interactions. Additionally, we can make use of existing code for posterior inference with Hawkes processes and graphs based on compound CRMs, and therefore propose a two-step approximate inference procedure, motivated by modular Bayesian inference (Jacob et al., 2017).

Let $\mathcal{Z} = (z_{ij}(T))_{1\leq i,j\leq V}$ be the adjacency matrix defined by $z_{ij}(T) = 1$ if there is at least one interaction between i and j in the interval $[0, T]$, and 0 otherwise. We have

$$\pi(\phi, \xi|\mathcal{D}) = \pi(\phi, \xi|\mathcal{D}, \mathcal{Z}) = \pi(\xi|\mathcal{D}, \mathcal{Z})\pi(\phi|\xi, \mathcal{D}).$$

The idea of the two-step procedure is to (i) Approximate $\pi(\xi|\mathcal{D}, \mathcal{Z})$ by $\pi(\xi|\mathcal{Z})$ and obtain a Bayesian point estimate $\widehat{\xi}$ then (ii) Approximate $\pi(\phi|\xi, \mathcal{D})$ by $\pi(\phi|\widehat{\xi}, \mathcal{D})$.

The full posterior is thus approximated by $\widetilde{\pi}(\phi, \xi) = \pi(\xi|\mathcal{Z})\pi(\phi|\widehat{\xi}, \mathcal{D})$. As mentioned in Section 4.1, the statistical model for the binary adjacency matrix $\mathcal{Z}$ is the same as in (Todeschini et al., 2016). We use the MCMC scheme of (Todeschini et al., 2016) and the accompanying software package SNetOC[2] to perform inference. The MCMC sampler is a Gibbs sampler which uses a Metropolis-Hastings (MH) step to update the hyperparameters and a Hamiltonian Monte Carlo (HMC) step for the parameters. From the posterior samples we compute a point estimate $(\widehat{w}_{i1}, \ldots, \widehat{w}_{ip})$ of the weight vector for each node. We follow the approach of Todeschini et al. (2016) and compute a minimum Bayes risk point estimate using a permutation-invariant cost function. Given these point estimates we obtain estimates of the base intensities $\widehat{\mu}_{ij}$. Posterior inference on the parameters ϕ of the Hawkes kernel is then performed using Metropolis-Hastings, as in (Rasmussen, 2013). Details of the two-stage inference procedure are given in the supplementary material.

Empirical investigation of posterior consistency. To validate the two-step approximation to the posterior distribution, we study empirically the convergence of our approximate inference scheme using synthetic data. Experiments suggest that the posterior concentrates around the true parameter value. More details are given in the supplementary material.

6 Experiments

We perform experiments on four temporal interaction datasets from the Stanford Large Network Dataset Collection[3] (Leskovec and Krevl, 2014):

• The EMAIL dataset consists of emails sent within a large European research institution over 803 days. There are 986 nodes, 24929 edges and 332334 interactions. A separate interaction is created for every recipient of an email.

• The COLLEGE dataset consists of private messages sent over a period of 193 days on an online social network (Facebook-like platform) at the University of California, Irvine. There are 1899 nodes, 20296 edges and 59835 interactions. An interaction (t, i, j) corresponds to a user i sending a private message to another user j at time t.

• The MATH overflow dataset is a temporal network of interactions on the stack exchange website Math Overflow over 2350 days. There are 24818 nodes, 239978 edges and 506550 interactions.An

[2] https://github.com/misxenia/SNetOC

[3] https://snap.stanford.edu/data/

interaction (t, i, j) means that a user i answered another user's j question at time t, or commented on another user's j question/response.
• The UBUNTU dataset is a temporal network of interactions on the stack exchange website Ask Ubuntu over 2613 days. There are 159316 nodes, 596933 edges and 964437 interactions. An interaction (t, i, j) means that a user i answered another user's j question at time t, or commented on another user's j question/response.

Comparison on link prediction. We compare our model (Hawkes-CCRM) to five other benchmark methods: (i) our model, without the Hawkes component (obtained by setting $\eta = 0$), (ii) the Hawkes-IRM model of Blundell et al. (2012) which uses an infinite relational model (IRM) to capture the community structure together with a Hawkes process to capture reciprocity, (iii) the same model, called Poisson-IRM, without the Hawkes component, (iv) a simple Hawkes model where the conditional intensity given by Equation (7) is assumed to be same for each pair of individuals, with unknown parameters δ and η, (v) a simple Poisson process model, which assumes that interactions between two individuals arise at an unknown constant rate. Each of these competing models capture a subset of the features we aim to capture in the data: sparsity/heterogeneity, community structure and reciprocity, as summarized in Table 1. The models are given in the supplementary material. The only model to account for all the features is the proposed Hawkes-CCRM model.

Table 1: (Left) Performance in link prediction. (Right) Properties captured by the different models.

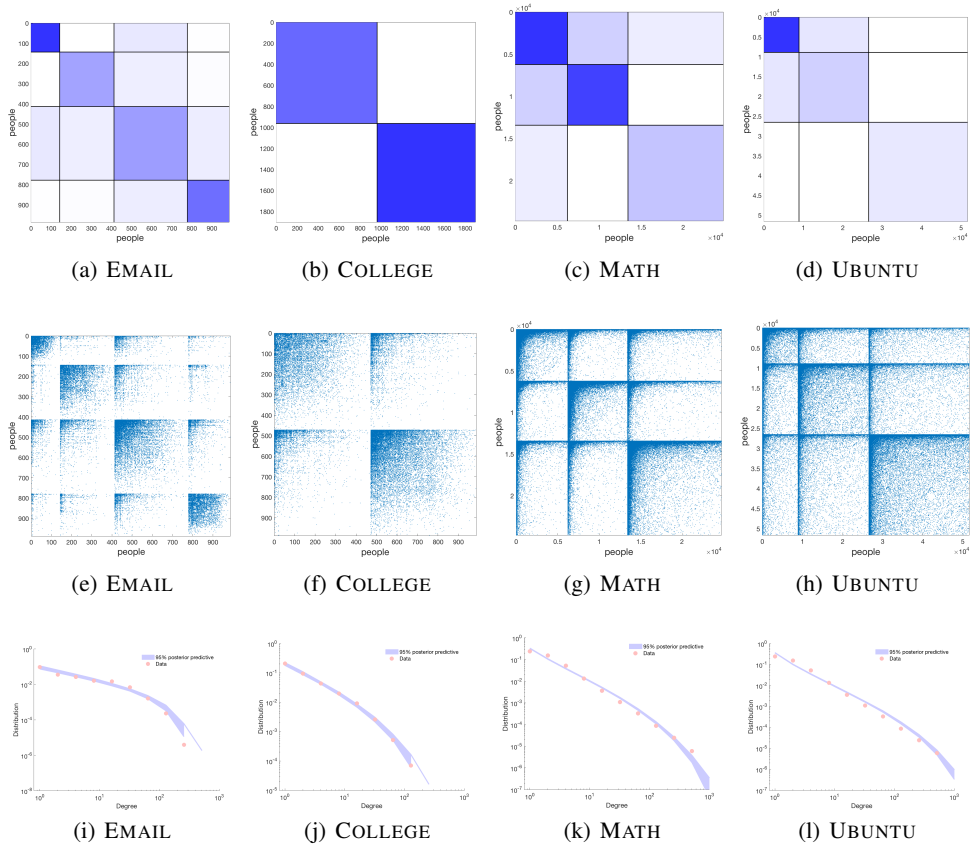
	email	college	math	ubuntu
Hawkes-CCRM	**10.95**	**1.88**	**20.07**	**29.1**
CCRM	12.08	2.90	89.0	36.5
Hawkes-IRM	14.2	3.56	96.9	59.5
Poisson-IRM	31.7	15.7	204.7	79.3
Hawkes	154.8	153.29	220.10	191.39
Poisson	$\sim 10^3$	$\sim 10^4$	$\sim 10^4$	$\sim 10^4$

sparsity/ heterogeneity	community structure	reciprocity
✓	✓	✓
✓	✓	
	✓	✓
	✓	
		✓

We perform posterior inference using a Markov chain Monte Carlo algorithm. For our Hawkes-CCRM model, we follow the two-step procedure described in Section 5. For each dataset, there is some background information in order to guide the choice of the number p of communities. The number of communities p is set to $p = 4$ for the EMAIL dataset, as there are 4 departments at the institution, $p = 2$ for the COLLEGE dataset corresponding to the two genders, and $p = 3$ for the MATH and UBUNTU datasets, corresponding to the three different types of possible interactions. We follow Todeschini et al. (2016) regarding the choice of the MCMC tuning parameters and initialise the MCMC algorithm with the estimates obtained by running a MCMC algorithm with $p = 1$ feature with fewer iterations. For all experiments we run 2 chains in parallel for each stage of the inference. We use 100000 iterations for the first stage and 10000 for the second one. For the Hawkes-IRM model, we also use a similar two-step procedure, which first obtains a point estimate of the parameters and hyperparameters of the IRM, then estimates the parameters of the Hawkes process given this point estimate. This allows us to scale this approach to the large datasets considered. We use the same number of MCMC samples as for our model for each step.

We compare the different algorithms on link prediction. For each dataset, we make a train-test split in time so that the training datasets contains 85% of the total temporal interactions. We use the training data for parameter learning and then use the estimated parameters to perform link prediction on the held out test data. We report the root mean square error between the predicted and true number of interactions for each directed pair in the test set . The results are reported in Table 1. On all the datasets, the proposed Hawkes-CCRM approach outperforms other methods. Interestingly, the addition of the Hawkes component brings improvement for both the IRM-based model and the CCRM-based model.

Community structure, degree distribution and sparsity. Our model also estimates the latent structure in the data through the weights w_{ik}, representing the level of affiliation of a node i to a community k. For each dataset, we order the nodes by their highest estimated feature weight, obtaining a clustering of the nodes. We represent the ordered matrix $(z_{ij}(T))$ of binary interactions in Figure 4 (a)-(d). This shows that the method can uncover the latent community structure in the different datasets. Within each community, nodes still exhibit degree heterogeneity as shown in Figure 4 (e)-(h). where the nodes are sorted within each block according to their estimated sociability $\widehat{w}_{i0}$. The ability of the approach to uncover latent structure was illustrated by Todeschini et al. (2016), who demonstrate that models which do not account for degree heterogeneity, cannot capture latent community estimation



(a) EMAIL (b) COLLEGE (c) MATH (d) UBUNTU

(e) EMAIL (f) COLLEGE (g) MATH (h) UBUNTU

(i) EMAIL (j) COLLEGE (k) MATH (l) UBUNTU

Figure 4: Top: Sorted adjacency matrix for each dataset. The vertices are classified to one of the communities based to their highest affiliation. Darker color correspond to more interactions. Middle: Sorted adjacency matrix. The nodes are grouped according to their highest affiliation and then sorted according to their estimated sociability parameter $\widehat{w}_{i0}$. Bottom: Empirical degree distribution (red) and posterior predictive distribution (blue).

but they rather cluster the nodes based on their degree. We also look at the posterior predictive degree distribution based on the estimated hyperparameters, and compare it to the empirical degree distribution in the test set. The results are reported in Figure 4 (i)-(l) showing a reasonable fit to the degree distribution. Finally we report the 95% posterior credible intervals (PCI) for the sparsity parameter σ for all datasets. Each PCI is $(-0.69, -0.50), (-0.35, -0.20), (0.15, 0.18), (0.51, 0.57)$ respectively. The range of σ is $(-\infty, 1)$. EMAIL and COLLEGE give negative values corresponding to denser networks whereas MATH and UBUNTU datasets are sparser.

7 Conclusion

We have presented a novel statistical model for temporal interaction data which captures multiple important features observed in such datasets, and shown that our approach outperforms competing models in link prediction. The model could be extended in several directions. One could consider asymmetry in the base intensities $\mu_{ij} \neq \mu_{ji}$ and/or a bilinear form as in (Zhou, 2015). Another important extension would be the estimation of the number of latent commmunities/features p.

Acknowledgments. The authors thank the reviewers and area chair for their constructive comments. XM, FC and YWT acknowledge funding from the ERC under the European Union's 7th Framework programme (FP7/2007-2013) ERC grant agreement no. 617071. FC acknowledges support from EPSRC under grant EP/P026753/1 and from the Alan Turing Institute under EPSRC grant EP/N510129/1. XM acknowledges support from the A. G. Leventis Foundation.

References

C. Blundell, K. Heller, and J. Beck. Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems 25*, volume 15, pages 5249–5262. Curran Associates, Inc., 2012.

C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research*, 18:1–71, 2018.

D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4249–4257. Curran Associates, Inc., 2016.

F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society B*, 79(5), 2017.

H. Crane and W. Dempsey. A framework for statistical network modeling. *arXiv preprint arXiv:1509.08185*, 2015.

H. Crane and W. Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018.

C. Dubois, C. Butts, and P. Smyth. Stochastic blockmodeling of relational event dynamics. In *Artificial Intelligence and Statistics*, pages 238–246, 2013.

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.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(7):5249–5262, 1971.

T. Herlau, M. N. Schmidt, and M. Mørup. Completely random measures for modelling block-structured sparse networks. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

P. E. Jacob, M. L. M., H. C. C., and R. C. P. Better together? statistical learning in models made of modules. *ArXiv preprint arXiv:1708.08719*, 2017.

S. Janson. On convergence for graphexes. *arXiv preprint arXiv:1702.06389*, 2017a.

S. Janson. On edge exchangeable random graphs. *To appear in Journal of Statistical Physics.*, 2017b.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

J. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

J. Kingman. *Poisson processes*, volume 3. Oxford University Press, USA, 1993.

J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

S. Linderman and R. Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

Y. C. Ng and R. Silva. A dynamic edge exchangeable model for sparse temporal networks. *arXiv preprint arXiv:1710.04008*, 2017.

K. Palla, F. Caron, and Y. Teh. Bayesian nonparametrics for sparse dynamic networks. *arXiv preprint arXiv:1607.01624*, 2016.

J. G. Rasmussen. Bayesian inference for Hawkes processes. *Methodological Computational Applied Probability*, 15:623–642, 2013.

A. Todeschini, X. Miscouridou, and F. Caron. Exchangeable random measure for sparse networks with overlapping communities. *arXiv:1602.02114*, 2016.

V. Veitch and D. M. Roy. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.

S. Williamson. Nonparametric network models for link prediction. *Journal of Machine Learning Research*, 17(202):1–21, 2016.

M. Zhou. Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 1135–1143, San Diego, California, USA, 09–12 May 2015. PMLR.

Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data

Xenia Miscouridou[1], **François Caron**[1], **Yee Whye Teh**[1,2]
[1]Department of Statistics, University of Oxford
[2]DeepMind
{miscouri, caron, y.w.teh}@stats.ox.ac.uk

1 Background on compound completely random measures

We give the necessary background on compound completely random measures (CCRM). An extensive account of this class of models is given in (Griffin and Leisen, 2017). In this article, we consider a CCRM $(W_1, \dots, W_p)$ on $\mathbb{R}_+$ characterized, for any $(t_1, \dots, t_p) \in \mathbb{R}_+^p$ and measurable set $A \subset \mathbb{R}_+$, by

$$\mathbb{E}[e^{-\sum_{k=1}^p t_k W_k(A)}] = \exp(-H_0(A)\psi(t_1, \dots, t_p))$$

where H_0 is the Lebesgue measure and ψ is the multivariate Laplace exponent defined by

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

The multivariate Lévy measure ρ takes the form

$$\rho(dw_1, \dots, dw_p) = \int_0^\infty w_0^{-p} \prod_{k=1}^p F_k\left(\frac{dw_k}{w_0}\right) \rho_0(dw_0) \tag{2}$$

where F_k is the distribution of a Gamma random variable with parameters a_k and b_k and ρ_0 is the Lévy measure on $(0, \infty)$ of a generalized gamma process

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

where $\sigma \in (-\infty, 1)$ and $\tau > 0$.

Denote $\boldsymbol{w}_i = (w_{i1}, \dots, w_{ip})^T$, $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})^T$ and $\rho(d\boldsymbol{w}) = \rho(dw_1, \dots, dw_p)$. w_0 always refers to the scalar weight corresponding to the measure ρ_0.

2 Expected number of interactions, edges and nodes

Recall that $I_{\alpha,T}$, $E_{\alpha,T}$ and $V_{\alpha,T}$ are respectively the overall number of interactions between nodes with label $\theta_i \leq \alpha$ until time T, the total number of pairs of nodes with label $\theta_i \leq \alpha$ who had at least one interaction before time T, and the number of nodes with label $\theta_i \leq \alpha$ who had at least one interaction before time T respectively, and are defined as

$$I_{\alpha,T} = \sum_{i \neq j} N_{ij}(T) 1_{\theta_i \leq \alpha} 1_{\theta_j \leq \alpha}$$

$$E_{\alpha,T} = \sum_{i<j} 1_{N_{ij}(T)+N_{ji}(T)>0} 1_{\theta_i \leq \alpha} 1_{\theta_j \leq \alpha}$$

$$V_{\alpha,T} = \sum_i 1_{\sum_{j \neq i} N_{ij}(T) 1_{\theta_j \leq \alpha} > 0} 1_{\theta_i \leq \alpha}$$

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Theorem 1 *The expected number of interactions $I_{\alpha,T}$, edges $E_{\alpha,T}$ and nodes $V_{\alpha,T}$ are given as follows:*

$$\mathbb{E}[I_{\alpha,T}] = \alpha^2 \boldsymbol{\mu}_w^T \boldsymbol{\mu}_w \left(\frac{\delta}{\delta - \eta} T - \frac{\eta}{(\eta - \delta)^2} \left(1 - e^{-T(\delta - \eta)} \right) \right)$$

$$\mathbb{E}[E_{\alpha,T}] = \frac{\alpha^2}{2} \int_{\mathbb{R}_+^p} \psi(2Tw_1, \ldots, 2Tw_p) \rho(dw_1, \ldots, dw_p)$$

$$\mathbb{E}[V_{\alpha,T}] = \alpha \int_{\mathbb{R}_+^p} \left(1 - e^{-\alpha \psi(2Tw_1, \ldots, 2Tw_p)} \right) \rho(dw_1, \ldots, dw_p)$$

where $\mu_w = \int_{\mathbb{R}_+^p} \boldsymbol{w} \rho(dw_1, \ldots, dw_p)$.

The proof of Theorem 1 is given below and follows the lines of Theorem 3 in (Todeschini et al., 2016).

Mean number of nodes $\mathbb{E}[V_{\alpha,T}]$

We have

$$\begin{aligned} \mathbb{E}\left[V_{\alpha,T}\right] &= \mathbb{E}\left[\sum_i (1 - 1_{N_{ij}(T)=0, \forall j \neq i | \theta_j \leq \alpha)} \right] 1_{\theta_i \leq \alpha} \\ &= \sum_i \left\{ 1 - \mathbb{P}(N_{ij}(T) = 0, \forall j \neq i | \theta_j \leq \alpha) \right\} 1_{\theta_i \leq \alpha} \end{aligned}$$

Using the Palm/Slivnyak-Mecke formula and Campbell's theorem, see e.g. (Møller and Waagepetersen, 2003, Theorem 3.2) and (Kingman, 1993), we obtain

$$\begin{aligned} \mathbb{E}[V_{\alpha,T}] &= \mathbb{E}\left[\mathbb{E}\left[V_{\alpha,T} | W_1, \ldots, W_p\right]\right] \\ &= \mathbb{E}\left[\sum_i \left(1 - e^{-2T \boldsymbol{w}_i^T \sum_{j \neq i} \boldsymbol{w}_j 1_{\theta_j \leq \alpha}} \right) 1_{\theta_i \leq \alpha} \right] \\ &= \alpha \int \mathbb{E}\left(1 - e^{-2T \boldsymbol{w}^T \sum_j \boldsymbol{w}_j 1_{\theta_j \leq \alpha}} \right) \rho(d\boldsymbol{w}) \\ &= \alpha \int_{\mathbb{R}_+^p} \left(1 - e^{-\alpha \psi(2Tw_1, \ldots, 2Tw_p)} \right) \rho\left(dw_1, \ldots, dw_p\right) \end{aligned}$$

Mean number of edges $\mathbb{E}[E_{\alpha,T}]$

Using the extended Slivnyak-Mecke formula, see e.g. (Møller and Waagepetersen, 2003, Theorem 3.3)

$$\begin{aligned} \mathbb{E}\left[E_{\alpha,T}\right] &= \mathbb{E}\left[\mathbb{E}\left[E_{\alpha,T} | W_1, \ldots, W_p\right]\right] \\ &= \mathbb{E}\left[\sum_i 1_{\theta_i < \alpha} \frac{1}{2} \sum_{j \neq i} 1_{\theta_j \leq \alpha} (1 - e^{-2T \boldsymbol{w}_i^T \boldsymbol{w}_j}) \right] \\ &= \frac{\alpha^2}{2} \int_{\mathbb{R}_+^p} \psi(2Tw_1, \ldots, 2Tw_p) \rho(dw_1, \ldots, dw_p) \end{aligned}$$

Mean number of interactions $\mathbb{E}[I_{\alpha,T}]$

We have

$$\begin{aligned}
\mathbb{E}[I_{\alpha,T}] &= \mathbb{E}\left[\mathbb{E}\left[I_{\alpha,T}|W_1,\ldots,W_p\right]\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}\left(\mathbb{E}\left[\int_0^T \lambda_{ij}(t)dt \mid W_1,\ldots,W_p\right]\right)\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}\left(\int_0^T \mathbb{E}\left[\mu_{ij}\frac{\delta}{\delta-\eta} - \mu_{ij}\frac{\eta}{\delta-\eta}e^{-t(\delta-\eta)} \mid W_1,\ldots,W_p\right]dt\right)\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}\boldsymbol{w}_i^T\boldsymbol{w}_j\right]\int_0^T\left[\frac{\delta}{\delta-\eta} - \frac{\eta}{\delta-\eta}e^{-t(\delta-\eta)}\right]dt \\
&= \alpha^2\boldsymbol{\mu}_w^T\boldsymbol{\mu}_w\left(\frac{\delta}{\delta-\eta}T - \frac{\eta}{(\delta-\eta)^2}(1-e^{-T(\delta-\eta)})\right)
\end{aligned}$$

where the third line follows from (Dassios and Zhao, 2013), the last line follows from another application of the extended Slivnyak-Mecke formula for Poisson point processes and $\boldsymbol{\mu}_w = \int_{\mathbb{R}_+^p}\boldsymbol{w}\rho(dw_1,\ldots,dw_p)$.

3 Details of the approximate inference algorithm

Here we provide additional details on the two-stage procedure for approximate posterior inference. The code is publicly available at `https://github.com/OxCSML-BayesNP/HawkesNetOC`.

Given a set of observed interactions $\mathcal{D} = (t_k, i_k, j_k)_{k\geq 1}$ between V individuals over a period of time T, the objective is to approximate the posterior distribution $\pi(\phi,\xi|\mathcal{D})$ where $\phi = (\eta,\delta)$ are the kernel parameters and $\xi = ((w_{ik})_{i=1,\ldots,V,k=1,\ldots,p}, (a_k,b_k)_{k=1,\ldots,p}, \alpha,\sigma,\tau)$, the parameters and hyperparameters of the compound CRM. Given data $\mathcal{D}$, let $\mathcal{Z} = (z_{ij}(T))_{1\leq i,j\leq V}$ be the adjacency matrix defined by $z_{ij}(T) = 1$ if there is at least one interaction between i and j in the interval $[0,T]$, and 0 otherwise.

For posterior inference, we employ an approximate procedure, which is formulated in two steps and is motivated by modular Bayesian inference (Jacob et al., 2017). It also gives another way to see the two natures of this type of temporal network data. Firstly we focus on the static graph i.e. the adjacency matrix of the pairs of interactions. Secondly, given the node pairs that have at least one interaction, we learn the rate for the appearance of those interactions assuming they appear in a reciprocal manner by mutual excitation.

We have

$$\pi(\phi,\xi|\mathcal{D}) = \pi(\phi,\xi|\mathcal{D},\mathcal{Z}) = \pi(\xi|\mathcal{D},\mathcal{Z})\pi(\phi|\xi,\mathcal{D}).$$

The idea of the two-step procedure is to

1. Approximate $\pi(\xi|\mathcal{D},\mathcal{Z})$ by $\pi(\xi|\mathcal{Z})$ and obtain a Bayesian point estimate $\widehat{\xi}$.
2. Approximate $\pi(\phi|\xi,\mathcal{D})$ by $\pi(\phi|\widehat{\xi},\mathcal{D})$.

3.1 Stage 1

As mentioned in Section 5 in the main article, the joint model $\pi(\mathcal{Z},\xi)$ on the binary undirected graph is equivalent to the model introduced by (Todeschini et al., 2016), and we will use their Markov chain Monte Carlo (MCMC) algorithm and the publicly available code SNetOC[1] in order to approximate the posterior $\pi(\xi|\mathcal{Z})$ and obtain a Bayesian point estimate $\widehat{\xi}$. Let $w_{*k} = W_k([0,\alpha]) - \sum_i w_{ik}$ corresponding to the overall level of affiliation to community k of all the nodes with no interaction (recall that in our model, the number of nodes with no interaction may be infinite). For each

[1] https://github.com/misxenia/SNetOC

undirected pair i, j such that $z_{ij}(T) = 1$, consider latent count variables $\widetilde{n}_{ijk}$ distributed from a truncated multivariate Poisson distribution, see (Todeschini et al., 2016, Equation (31)). The MCMC sampler to produce samples asymptotically distributed according to $\pi(\xi|\mathcal{Z})$ then alternates between the following steps:

1. Update $(w_{ik})_{i=1,\dots,V,k=1,\dots,p}$ using an Hamiltonian Monte Carlo (HMC) update,
2. Update $(w_{*k}, a_k, b_k)_{k=1,\dots,p}, \alpha, \sigma, \tau$ using a Metropolis-Hastings step
3. Update the latent count variables using a truncated multivariate Poisson distribution

We use the same parameter settings as in (Todeschini et al., 2016). We use $\epsilon = 10^{-3}$ as truncation level to simulate the w_{*k}, and set the number of leapfrog steps $L = 10$ in the HMC step. The stepsizes of both the HMC and the random walk MH are adapted during the first 50 000 iterations so as to target acceptance ratios of 0.65 and 0.23 respectively.

The minimum Bayes point estimates $(\widehat{w}_{ik})_{i=1,\dots,V,k=1,\dots,p}$ are then computed using a permutation-invariant cost function, as described in (Todeschini et al., 2016, Section 5.2). This allows to compute point estimates of the base intensity measures of the Hawkes processes, for each $i \neq j$

$$\widehat{\mu}_{ij} = \sum_{k=1}^{p} \widehat{w}_{ik}\widehat{w}_{jk}.$$

3.2 Stage 2

In stage 2, we use a MCMC algorithm to obtain samples approximately distributed according to

$$\pi(\phi|\widehat{\xi}, \mathcal{D}) = \pi(\phi|(\widehat{\mu}_{ij}), \mathcal{D})$$

where ϕ are the parameters of the Hawkes kernel. For each ordered pair (i, j) such that $n_{ij} = N_{ij}(T) > 0$, let $(t_{ij}^{(1)} < t_{ij}^{(2)} < \ldots < t_{ij}^{(n_{ij})})$ be the times of the observed directed interactions from i to j. The intensity function is

$$\lambda_{ij}(t) = \widehat{\mu}_{ij} + \sum_{\ell | t_{ji}^{(\ell)} < t} \frac{\eta}{\delta} f_\delta(t - t_{ji}^{(\ell)}))$$

where

$$\frac{\eta}{\delta} f_\delta(t - t_{ji}^{(\ell)}) = \eta \times e^{-\delta(t - t_{ji}^{(\ell)})}.$$

We write the kernel in this way to point out that it is equal to a density function f_δ, (here exponential density), scaled by the step size η/δ. We denote the distribution function of f_δ by F_δ. Then, by Proposition 7.2.III in (Daley and Vere-Jones, 2008) we obtain the likelihood

$$L(\mathcal{D} \mid \phi, (\widehat{\mu}_{ij})) = \prod_{(i,j)|N_{ij}(T)>0} \left[\exp(-\Lambda_{ij}(T)) \prod_{l=1}^{n_{ij}} \lambda_{ij}(t_{ij}^{(\ell)}) \right]$$

where

$$\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)\, du = t\widehat{\mu}_{ij} + \sum_{\ell | t_{ji}^{(\ell)} < t} \frac{\eta}{\delta} F_\delta(t - t_{ji}^{(\ell)}).$$

We derive a Gibbs sampler with Metropolis Hastings steps to estimate the parameters η, δ conditionally on the estimates $\widehat{\mu}_{ij}$. As mentioned in Section 3 of the main article we follow (Rasmussen, 2013) for the choice of vague exponential priors $p(\eta), p(\delta)$. For proposals we use truncated Normals with variances $\left(\sigma_\eta^2, \sigma_\delta^2\right) = (1.5, 2.5)$.

The posterior is given by

$$\pi(\phi \mid \mathcal{D}, (\widehat{\mu}_{ij})) \propto \exp\left(- \sum_{(i,j)|N_{ij}(T)>0} \Lambda_{ij}(T) \right) \left[\prod_{(i,j)|N_{ij}(T)>0} \prod_{\ell=1}^{n_{ij}} \lambda_{ij}(t_{ij}^{(\ell)}) \right] \times p(\eta)p(\delta)$$

We use an efficient way to compute the intensity at each time point $t_{ij}^{(\ell)}$ by writing it in the form

$$\lambda_{ij}(t_{ij}^{(\ell)}) = \widehat{\mu}_{ij} + \eta S_{ij}^{(\ell)}(\delta),$$

where

$$S_{ij}^{(\ell)}(\delta) = e^{-\delta t_{ij}^{(\ell)}} \sum_{k=1}^{n_{ji}} e^{\delta t_{ji}^{(k)}} 1_{t_{ji}^{(k)} < t_{ij}^{(\ell)}},$$

and then derive a recursive relationship of $S_{ij}^{(\ell)}(\delta)$ in terms of $S_{ij}^{(\ell-1)}(\delta)$. In this way, we can precompute several terms by ordering the event times and arrange them in bins defined by the event times of the opposite process.

4 Posterior consistency

We simulate interaction data from the Hawkes-CCRM model described in Section 3 in the main article, using parameters $p = 4, \alpha = 50, \sigma = 0.3, \tau = 1, a_k = 0.08, \phi = (0.85, 3), T = 300$. We perform the two-step inference procedure with data of increasing sample size, and check empirically that the approximate posterior $\pi(\phi|\widehat{\xi}, \mathcal{D})$ concentrates around the true value as the sample size increases. Figure 1 below shows the plots of the approximate marginal posterior distribution of δ and η. Experiments suggest that the posterior still concentrates around the true parameter value under this approximate inference scheme.

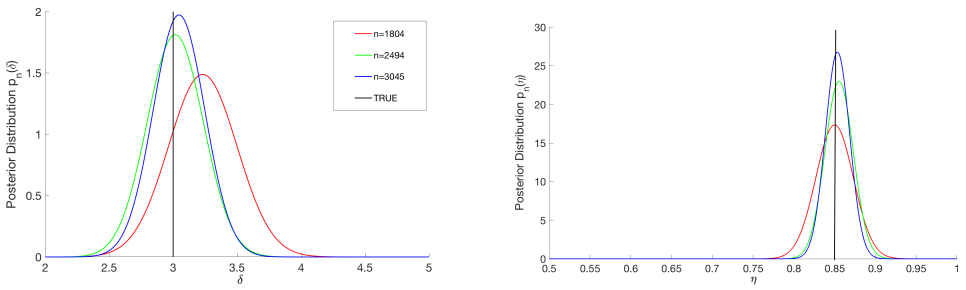


Figure 1: (Left) Approximate marginal posterior distribution of δ given n interaction data. (Right) Approximate marginal posterior distribution of η given n interaction data. Posterior concentrates around the true value, with increasing sample size n.

5 Experiments

We perform experiments in which we compare our Hawkes-CCRM model to five other competing models. The key part in all cases is the conditional intensity of the point process. which we give below. In all cases we use $\{t_{ij}^{(k)}\}_{k\geq 1}$ to refer to the set of events from i to j, i.e. the interactions for the directed pair (i, j).

Hawkes-CCRM

For each directed pair of nodes (i, j), $i \neq j$
$N_{ij}(t) \sim$ Hawkes Process$(\lambda_{ij}(t))$ where $\lambda_{ij}(t) = \sum_k w_{ik} w_{jk} + \sum_{t_{ji}^{(k)} < t} \eta e^{-\delta(t - t_{ji}^{(k)})}$.

Hawkes-IRM (Blundell et al., 2012)

For each directed pair of clusters (p, q), $p \neq q$
$N_{pq}(t) \sim$ Hawkes Process$(\lambda_{pq}(t))$ where $\lambda_{pq}(t) = n_p n_q \gamma_{pq} + \sum_{t_{qp}^{(k)} < t} \eta e^{-\delta(t - t_{qp}^{(k)})}$.

For the details of the model see (Blundell et al., 2012).

Poisson-IRM (as explained in (Blundell et al., 2012))

For each directed pair of clusters (p, q), $p \neq q$
$N_{pq}(t) \sim$ Poisson$(\lambda_{pq}(t))$ where $\lambda_{pq}(t) = n_p n_q \gamma_{pq}$.

For the details of the model see (Blundell et al., 2012).

CCRM (Todeschini et al., 2016)

$N_{ij}(t) \sim$ Poisson Process$(\lambda_{ij}(t))$ where $\lambda_{ij}(t) = \sum_k w_{ik} w_{jk}$.

Hawkes

$N_{ij}(t) \sim$ Hawkes Process $(\lambda_{ij}(t))$ where $\lambda_{ij}(t) = \mu + \sum_{t_{ji}^{(k)} < t} \eta e^{-\delta(t - t_{ji}^{(k)})}$.

Poisson

$N_{ij}(t) \sim$ Poisson Process$(\lambda_{ij}(t))$ where $\lambda_{ij}(t) = \mu$.

References

C. Blundell, K. Heller, and J. Beck. Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems 25*, volume 15, pages 5249–5262. Curran Associates, Inc., 2012.

D. 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.

A. Dassios and H. Zhao. Exact simulation of hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18(62):1–13, 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. Kingman. *Poisson processes*, volume 3. Oxford University Press, USA, 1993.

J. Møller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, 2003.

J. G. Rasmussen. Bayesian inference for Hawkes processes. *Methodological Computational Applied Probability*, 15:623–642, 2013.

A. Todeschini, X. Miscouridou, and F. Caron. Exchangeable random measure for sparse networks with overlapping communities. *arXiv:1602.02114*, 2016.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data
Publication Status	☑Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Published in the proceedings of the International Conference on Neural Information Processing Systems, 2018.

Student Confirmation

Student Name:	Xenia Miscouridou
Contribution to the Paper	I have been the primary contributor of the paper. On (a) theory: formulated the model based on mutually exciting processes, derived the exact and asymptotic properties of the graph class in two axes of growth (static and dynamic) (b) methodology: proposed and implemented the inference approach (c) applications: simulation experiments and comparisons of this model alternative models in literature

Signature  Date 05/01/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Prof François CARON
Supervisor comments	—

Signature  Date 08/01/2020

Chapter 4

A Unified Construction for Series Representations and Finite Approximations of Completely Random Measures

Under review by the Bernoulli Journal

Submitted to Bernoulli
arXiv: arXiv:1905.10733

A unified construction for series representations and finite approximations of completely random measures

JUHO LEE[1,*] XENIA MISCOURIDOU[2,**] and FRANÇOIS CARON[2,†]

[1] *AITRICS, Seoul, Republic of Korea*
E-mail: [*] juho@aitrics.com

[2] *Department of Statistics, University of Oxford*
E-mail: [**] xenia.miscouridou@spc.ox.ac.uk; [†] caron@stats.ox.ac.uk,

Infinite-activity completely random measures (CRMs) have become important building blocks of complex Bayesian nonparametric models. They have been successfully used in various applications such as clustering, density estimation, latent feature models, survival analysis or network science. Popular infinite-activity CRMs include the (generalized) gamma process and the (stable) beta process. However, except in some specific cases, exact simulation or scalable inference with these models is challenging and finite-dimensional approximations are often considered. In this work, we propose a general and unified framework to derive both series representations and finite-dimensional approximations of CRMs. Our framework can be seen as a generalization of constructions based on size-biased sampling of Poisson point process [Perman et al., 1992]. It includes as special cases several known series representations and finite approximations as well as novel ones. In particular, we show that one can get novel series representations for the generalized gamma process and the stable beta process. We show how these construction can be used to derive novel algorithms for posterior inference, including a generalization of the slice sampler for normalized CRMs mixture models introduced by Griffin and Walker [2011]. We also provide some analysis of the truncation error.

Keywords: Bayesian nonparametrics, Poisson Random Measures, Completely Random Measure, Generalized Gamma Process, Slice sampling.

1. Introduction

Probabilistic models with infinite-dimensional parameters arise as building blocks of numerous modern structured statistical models. Of particular interest are models building on infinite-activity completely random measures (CRMs), and more generally functionals of infinite-activity Poisson random measures. Examples include clustering and density estimation [Regazzini et al., 2003, James, 2005, James et al., 2009], spatial statistics [Brix, 1999, Wolpert and Ickstadt, 1998a, Moller and Waagepetersen, 2003], latent factor/trait models [Griffiths and Ghahramani, 2006, Paisley and Carin, 2009, Zhou et al., 2012, Campbell et al., 2018, Ayed and Caron, 2019], network modeling [Caron, 2012, Zhou,

2015, Cai et al., 2016, Caron and Fox, 2017], recommendation systems [Gopalan et al., 2014], prediction, risk management and option pricing of financial assets [Cont and Tankov, 2004] or survival analysis [Hjort, 1990, Nieto-Barajas et al., 2004]; see Lijoi and Prünster [2010] for a review. Popular CRMs include the (generalized) gamma random measure (also known as tempered stable) [Hougaard, 1986, Brix, 1999] or the (stable) beta random measure [Hjort, 1990, Teh and Gorur, 2009]. Some standard random probability measures are also obtained by normalization (Dirichlet process) or transformation (Pitman-Yor process) of CRMs.

The use of statistical models based on infinite-activity CRMs poses a number of practical challenges regarding posterior inference and estimation. Except in some specific cases where the infinite-dimensional parameter can be integrated out, most algorithms, either based on Gibbs sampling [Ishwaran and James, 2001, Zhou, 2015], slice sampling [Neal, 2003, Griffin and Walker, 2011, Foti and Williamson, 2012], mean-field variational inference [Blei and Jordan, 2006, Doshi et al., 2009, Lee et al., 2016] or sequential Monte Carlo [Barndorff-Nielsen and Shephard, 2001],[Andrieu et al., 2010, Section 3.2.], [Griffin, 2017], require the use of a finite-dimensional approximation of the CRM. Finite-dimensional approximations can either be obtained by (i) truncating a series representation of the CRM, with stochastically decreasing weights, or (ii) by considering a finite measure with n atoms and iid weights, converging in distribution to the CRM as n tends to infinity. Consider for example the beta process with scale parameter $\alpha > 0$ and probability distribution H. Its inverse Lévy series representation is given by [Teh et al., 2007]

$$G = \sum_{i=1}^{\infty} W_i \delta_{\theta_i} \tag{1.1a}$$

where, for $i = 1, 2, \ldots$,

$$W_i = \prod_{j=1}^{i} \beta_j, \quad \beta_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1), \quad \theta_i \overset{\text{i.i.d.}}{\sim} H. \tag{1.1b}$$

A standard finite-dimensional approximation of the beta process with iid weights is [Griffiths and Ghahramani, 2006]

$$\widetilde{G}_n = \sum_{i=1}^{n} \widetilde{W}_{n,i} \delta_{\widetilde{\theta}_{n,i}} \tag{1.2a}$$

where, for $i = 1, \ldots, n$,

$$\widetilde{W}_{n,i} \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha/n, 1), \quad \widetilde{\theta}_{n,i} \overset{\text{i.i.d.}}{\sim} H. \tag{1.2b}$$

Both representations are routinely used in Markov chain Monte Carlo and variational Bayes approximate inference algorithms [Doshi et al., 2009, Paisley et al., 2011, 2012]. Series and iid approximations are similarly used for the gamma process [Zhou, 2015, Roychowdhury and Kulis, 2015] and the generalized gamma process [Lee et al., 2016]. Since the early work of Khintchine [1937] and Ferguson and Klass [1972] on the so-called

inverse Lévy representation, various generic series representations of Poisson random measures have been proposed [Bondesson, 1982, Perman et al., 1992, Rosiński, 1990, 2001a, Campbell et al., 2019].

The objective of this article is to present a general framework to obtain both series and iid representations of CRMs. Our construction builds on the definition of a Poisson random measure on an extended space; it generalizes the size-biased approach of Perman et al. [1992], and admits as special cases both the size-biased and inverse-Lévy representations. We show that under this construction, one can draw connections between existing series and iid representations that appeared unrelated, and it allows to derive new series and iid representations. More precisely, we show how the iid representation of Lee et al. [2016] is related to the size-biased construction of Perman et al. [1992], derive novel series and iid representations of the generalized gamma and stable beta random measures. Additionally we explain how we can use our framework to develop novel slice sampling algorithms for exact posterior inference, generalizing the sampler of Griffin and Walker [2011]. Finally, we provide an asymptotic analysis of the truncation error for this class of approximations.

This article is organised as follows. In Section 2 we provide background material on completely random measures and some existing series representations for CRMs and describe the objectives. Section 3 describes the general construction for obtaining series and iid representations of CRMs. In Section 4 we describe a number of specific constructions, showing how one recovers some existing constructions as a particular case of our framework. Related approaches are discussed in Section 5. In Section 6 we show how we can use the series representation to derive a slice sampler for posterior inference. In Section 7 we provide an analysis of the asymptotic truncation error. Some proofs, additional background material and derivations are provided in the Supplementary Material.

Notations. For a measure ν on S and a positive measurable function h on S, write $\nu(h) = \int_S h(x)\nu(dx)$. Let $(\xi_1, \xi_2, \dots,)$ be the ordered points of a unit-rate Poisson point process on $(0, \infty)$, that is $\xi_1, \xi_2 - \xi_1, \xi_3 - \xi_2, \dots$ are iid unit-rate exponential random variables. For a random variable X and a probability distribution F, $X \sim F$ means X has distribution F. For two strictly positive functions $f(t)$ and $g(t)$, $f(t) \overset{t\to\infty}{\sim} g(t)$ means $\lim_{t\to\infty} f(t)/g(t) = 1$. Similarly, for sequences of strictly positive random variables $(R_n)_{n\geq 1}$ and $(S_n)_{n\geq 1}$, $R_n \overset{n\to\infty}{\sim} S_n$ a.s. means $\lim_{n\to\infty} S_n/R_n = 1$ almost surely. With a slight abuse of notation, we use the same notation for the distribution of a random variable and its pdf. For instance, the probability density function (pdf) of a gamma random variable $X \sim \text{Gamma}(a, b)$, evaluated at x, is written as $\text{Gamma}(x; a, b)$.

2. Background

2.1. Completely random measures

Let $(S, \mathcal{S})$ be a complete and separable metric space endowed with its Borel σ-field, where $S = (0, \infty) \times \Theta$. For any point $X = (W, \theta) \in S$, we refer to $W > 0$ as the **size** of X.

Let N be a Poisson random measure on S with mean measure $\nu(dw, d\theta) = \rho(dw)\mu_w(d\theta)$ where ρ is a Borel measure on $(0, \infty)$, called size measure, satisfying

$$\int_0^\infty (1 - e^{-w})\rho(dw) < \infty \text{ and } \int_0^\infty \rho(dw) = \infty \tag{2.1}$$

and $\mu_w(\cdot)$ is a Markov probability kernel from $(0, \infty)$ to Θ. The linear functional

$$G(d\theta) = \int_0^\infty wN(dw, d\theta)$$

is an infinite-activity completely random measure [Kingman, 1967] on Θ with random weights and random atoms. We write $G \sim \text{CRM}(\rho, \mu_w)$. The conditions (2.1) imply that the atomic random measures N and W have an infinite number of atoms, and $W(\Theta)$ is almost surely finite. If $\mu_w = H$ does not depend on w, the CRM is said to be homogeneous. We assume in the rest of this article that one can easily simulate from μ_w (or H) and/or it admits a tractable density with respect to some reference measure (e.g. Lebesgue).

Two popular examples of CRMs are the generalized gamma process (GGP) [Hougaard, 1986, Brix, 1999], also known as (exponentially) tilted stable process, with size measure

$$\rho(dw) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\tau w} dw \tag{2.2}$$

where $\alpha > 0$, $\sigma \in (0, 1)$ and $\tau \geq 0$, or $\sigma \leq 0$ and $\tau > 0$, and the stable beta process (SBP) [Hjort, 1990, Teh and Gorur, 2009] with

$$\rho(dw) = \frac{\alpha}{B(1-\sigma, c+\sigma)} w^{-\sigma-1}(1-w)^{c+\sigma-1} \mathbb{1}_{\{0<w<1\}} dw, \tag{2.3}$$

where $\sigma \in (-\infty, 1)$, $c > -\sigma$, $\alpha > 0$ and $B(\cdot, \cdot)$ is the beta function. When $\sigma \geq 0$, both random measures are infinite-activity.

***Remark* 2.1.** The constructions described in this paper hold more generally when the first condition in Equation (2.1) is not satisfied, but $\int_x^\infty \rho(dw) < \infty$ for all $x > 0$. Note that in this case $W(\Theta) = \infty$ almost surely. An example of this more general case is given in Section 4.4 where $\rho(dw) = w^{-2}dw$.

2.2. Objective

Our objective is to derive general series representations for the Poisson random measure N, or equivalently the CRM G, of the form

$$G = \sum_{i=1}^\infty W_i \delta_{\theta_i} \tag{2.4}$$

where the sizes $(W_1, W_2, \ldots)$, are stochastically ordered. That is, for any $w > 0$, $\Pr(W_{i+1} > w) \leq \Pr(W_i > w)$. We write $W_1 \succeq W_2 \succeq \ldots$ and $X_i = (W_i, \theta_i)$. Denote by G_n the measure obtained by truncating the above series after n points

$$G_n = \sum_{i=1}^{n} W_i \delta_{\theta_i} = \sum_{i=1}^{n} \overline{W}_{n,i} \delta_{\overline{\theta}_{n,i}} \tag{2.5}$$

where $(\overline{X}_{n,1}, \ldots, \overline{X}_{n,n})$ is a finitely exchangeable random sequence defined by $\overline{X}_{n,i} = X_{\pi_n(i)}$ where π_n is a random permutation of the set $\{1, \ldots, n\}$. We will refer to the sequence $(X_1, \ldots, X_n)$ (or $(W_1, \ldots, W_n)$) as the **sequential truncated representation**, and $(\overline{X}_{n,1}, \ldots, \overline{X}_{n,n})$ (or $(\overline{W}_{n,1}, \ldots, \overline{W}_{n,n})$) as the **exchangeable truncated representation**. In Section 3 we will show that the exchangeable truncated representation can be approximated by a **finite iid representation**, which will be denoted $(\widetilde{X}_{n,1}, \ldots, \widetilde{X}_{n,n})$.

2.3. Existing representations of CRMs

In this section we present some existing series representations which we will show are special cases of the novel general construction introduced in Section 3.

Inverse-Lévy representation. For any $x > 0$, let

$$\overline{\rho}(x) = \int_x^{\infty} \rho(dw) \tag{2.6}$$

be the tail intensity of the size measure ρ, and denote by $\overline{\rho}^{-1}(y) = \inf\{x > 0 \mid \overline{\rho}(x) \leq y\}$ its generalized inverse. The inverse Lévy representation [Khintchine, 1937, Ferguson and Klass, 1972] is given by

$$W_i = \overline{\rho}^{-1}(\xi_i).$$

In this case, the sizes are ordered $W_1 \geq W_2 \geq \ldots$ and it therefore leads to the best possible approximation in terms of the sizes. While this representation has been used in many applications [Wolpert and Ickstadt, 1998b, Nieto-Barajas et al., 2004, Griffin and Walker, 2011, Barrios et al., 2013, Argiento et al., 2016a, Arbel and Prünster, 2017] its main limitation is that $\bar{\rho}^{-1}$ is in general non-tractable. Two exceptions are the beta random measure, whose inverse Lévy representation is given by Equation (1.1), and the stable random measure (corresponding to the measure (2.2) with $\sigma \in (0, 1)$ and $\tau = 0$) where the inverse Lévy representation is given by $W_i = (\xi_i \sigma \Gamma(1 - \sigma)/\alpha)^{-1/\sigma}$.

Size-biased representation. The size-biased sequential and exchangeable representations $(W_1, \ldots, W_n)$ and $(\overline{W}_{n,1}, \ldots, \overline{W}_{n,n})$, introduced by [Perman et al., 1992, Section 4], are given as follows[1]. Let $0 < T_1 \leq T_2 \leq \ldots$ be defined as

$$T_i = \Psi^{-1}(\xi_i)$$

[1]Note that this is different from what Campbell et al. [2019] call a size-biased representation.

where Ψ^{-1} is the generalized inverse of the Laplace exponent $\Psi(t) = \int_0^\infty (1-e^{-wt})\rho(dw)$ and

$$\mathbb{P}(W_i \in dw \mid T_i = t) = \frac{we^{-wt}\rho(dw)}{\psi(t)}$$

where $\psi(t) = \Psi'(t) = \int_0^\infty we^{-wt}\rho(dw)$. Additionally, given $T_{n+1} = t_{n+1}$, we have $\overline{W}_{n,1}, \ldots, \overline{W}_{n,n}$ are iid with distribution

$$\mathbb{P}(\overline{W}_{n,i} \in dw \mid T_{n+1} = t_{n+1}) = \frac{(1-e^{-wt_{n+1}})\,\rho(dw)}{\Psi(t_{n+1})}.$$

The term size-biased comes from the fact that the atoms are ordered by successively sampling without replacement according to their size W

$$\Pr\left(0 < T_1 \leq T_2 \leq T_3 \leq \ldots \mid G\right) = \prod_{k=1}^{\infty} \frac{W_k}{\sum_{j\geq k} W_j}.$$

In the case of the gamma random measure, which corresponds to Equation (2.2) with $\sigma = 0$, Perman et al. [1992] show that the series representation corresponds to Bondesson [1982]'s representation and is given by $T_i = \Psi^{-1}(\xi_i) = \tau(e^{\xi_i/\alpha} - 1)$ and $W_i \mid T_i = t_i \sim \text{Gamma}(1, \tau + t_i)$.

3. Series representations and finite approximations of CRMs

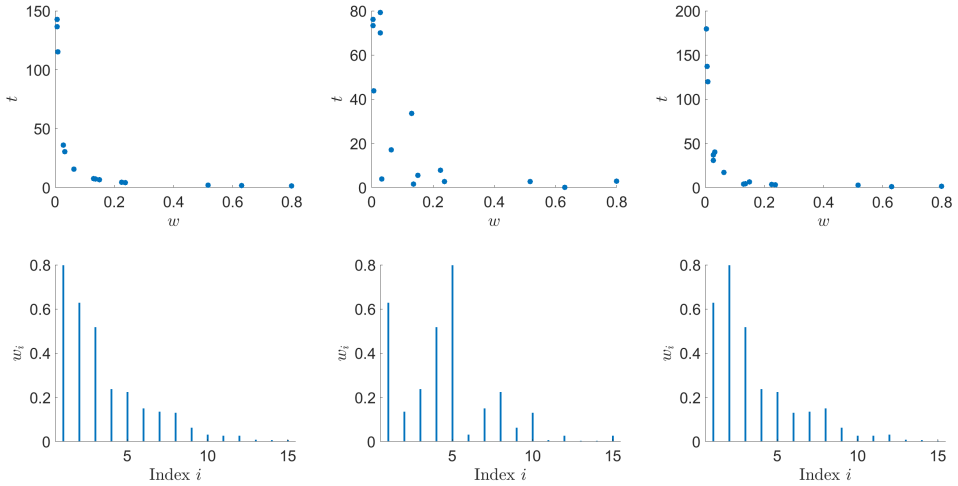
3.1. Arrival-time augmentation

Let $\lambda_w(dt)$ be some Markov probability kernel from $(0,\infty)$ to $(0,\infty)$ with cdf $\Lambda_w(t) = \int_0^t \lambda_w(du)$ satisfying, for any $t > 0$

$$\int_0^\infty \Lambda_w(t)\rho(dw) < \infty \quad \text{and} \quad \Lambda_{w_2}(t) \leq \Lambda_{w_1}(t) \text{ for all } 0 < w_2 \leq w_1. \tag{3.1}$$

That is, if $T_1 \sim \lambda_{w_1}$ and $T_2 \sim \lambda_{w_2}$, with $0 < w_2 \leq w_1$, then $T_2 \succeq T_1$.

We consider a Poisson random measure N' on the augmented space $S' = (0,\infty) \times \Theta \times (0,\infty)$ with mean measure $\nu'(dw, d\theta, dt) = \rho(dw)\mu_w(d\theta)\lambda_w(dt)$. For a point $X' = (W, \theta, T) \in S'$, we refer to T as the **arrival time** of the point X'. Indeed, the second condition in Equation (3.1) ensures that points with larger size W are more likely to have a smaller arrival time T. We may therefore consider the following analogy: atoms of the Poisson random measure are enrolled in a race, each atom having a strength W, and stronger atoms are more likely to finish faster and therefore have a smaller T. The first condition in Equation (3.1) ensures that $N'(\mathbb{R}_+, \Theta, (0,t)) < \infty$ for any $t > 0$ hence



(a) Deterministic (b) Exponential (c) Gamma ($\kappa = 10$)

Figure 1. Illustration of the Poisson process construction, for different arrival time cdfs λ_w. The top Figures represent the Poisson point process $(w_i, t_i)_{i\geq 1}$ and the bottom Figures the weights ordered according to their arrival times for (left) deterministic arrival time $\lambda_w(dt) = \delta_{1/w}(dt)$, (middle) Exponential arrival time $\lambda_w(dt) = we^{-wt}dt$ and (right) gamma arrival time $\lambda_w(dt) = \text{Gamma}(t; \kappa, \kappa w)dt$ with $\kappa = 10$.

we can order the arrival times. Let $0 < T_1 \leq T_2 \leq \ldots$ denote the sequence of ordered arrival times, and consider the augmented sequential representation

$$N' = \sum_{i=1}^{\infty} \delta_{(W_i, \theta_i, T_i)}$$

where $X_i = (W_i, \theta_i)$, $i \geq 1$ are the associated sizes and locations. By the restriction theorem [Kingman, 1993], $N = \sum_{i=1}^{\infty} \delta_{(W_i,\theta_i)}$ is a Poisson random measure with mean $\rho(dw)\mu_w(d\theta)$ and

$$G(d\theta) = \int_0^{\infty} \int_0^{\infty} wN'(dw, d\theta, dt) = \int_0^{\infty} wN(dw, d\theta) \sim \text{CRM}(\rho, \mu_w).$$

An illustration of the augmented Poisson point process construction is given in Figure 1.

We now give the general definitions of the sequential, exchangeable and iid representations of the CRM associated to the arrival time kernel λ_w. For simplicity of presentation, we assume that for any w, λ_w is absolutely continuous with respect to the Lebesgue measure, but one can also consider discontinuous cdfs Λ_w, see Section 4.1 for an example. We use the same notation $\lambda_w(dt)$ and $\lambda_w(t)$ for the distribution and its density.

3.2. Series and truncated exchangeable constructions

***Theorem* 3.1.** Let λ_w be a parametric distribution on $(0,\infty)$ with parameter $w>0$ and Λ_w be the associated parametric cumulative distribution function (cdf) satisfying condition (3.1). Consider the conditional distributions

$$\phi_t(dw) = \frac{\lambda_w(t)\rho(dw)}{\psi(t)} \quad \text{and} \quad \varphi_t(dw) = \frac{\Lambda_w(t)\rho(dw)}{\Psi(t)} \tag{3.2}$$

where

$$\Psi(t) = \int_0^\infty \Lambda_w(t)\rho(dw) \quad \text{and} \quad \psi(t) = \int_0^\infty \lambda_w(t)\rho(dw).$$

The **sequential construction** $G = \sum_{i=1}^\infty W_i\delta_{\theta_i}$ is obtained as follows, for $i \geq 1$

$$T_i = \Psi^{-1}(\xi_i) \tag{3.3a}$$
$$W_i \mid T_i = t_i \sim \phi_{t_i} \tag{3.3b}$$
$$\theta_i \mid W_i = w_i \sim \mu_{w_i}. \tag{3.3c}$$

The **truncated exchangeable construction** $G_n = \sum_{i=1}^n \overline{W}_{n,i}\delta_{\overline{\theta}_{n,i}}$ is obtained, for $i = 1,\ldots,n$ by

$$T_{n+1} = \Psi^{-1}(\xi_{n+1}) \tag{3.4a}$$
$$\overline{W}_{n,i} \mid T_{n+1} = t_{n+1} \overset{\text{i.i.d.}}{\sim} \varphi_{t_{n+1}} \tag{3.4b}$$
$$\overline{\theta}_{n,i} \mid \overline{W}_{n,i} = \overline{w}_{n,i} \sim \mu_{\overline{w}_{n,i}}. \tag{3.4c}$$

Proof. The proof is an adaption of the proof for the size-biased construction in [Perman et al., 1992, Section 4]. The mean measure ν' of the Poisson random measure N' can be expressed as

$$\begin{aligned}\nu'(dw,d\theta,dt) &= \rho(dw)\mu_w(d\theta)\lambda_w(t)dt \\ &= \psi(t)dt \times \frac{\lambda_w(t)\rho(dw)}{\psi(t)} \times \mu_w(d\theta) \\ &= \psi(t)dt \times \phi_t(dw) \times \mu_w(d\theta)\end{aligned}$$

This is the mean measure of a marked Poisson point process, where $(T_1,T_2,\ldots)$ are the points of an inhomogeneous Poisson point process with intensity $\psi(t)$, hence admit the representation of Equation (3.3), and the marks (W_i,θ_i) have conditional distribution $\phi_t(dw)\mu_w(d\theta)$ as shown in Equation (3.3). Let $(\overline{X}'_{n,1},\ldots,\overline{X}'_{n,n}) = (X'_{\pi_1},\ldots,X'_{\pi_n})$ where π_n is a random permutation of $\{1,\ldots,n\}$, and $\overline{X}'_{n,i} = (\overline{T}_{n,i},\overline{W}_{n,i},\overline{\theta}_{n,i})$. By properties of the Poisson process on the real line, the random variables $\overline{T}_{n,1},\ldots,\overline{T}_{n,n}$ are iid given $T_{n+1} = t_{n+1}$, with pdf

$$\frac{\psi(t)\mathbb{1}_{\{0<t<t_{n+1}\}}}{\Psi(t_{n+1})}.$$

Hence, given $T_{n+1} = t_{n+1}$, the marks $\overline{W}_{n,i}$ and $\overline{\theta}_{n,i}$ are also iid, with conditional distribution $\varphi_{t_{n+1}}(d\overline{w}_{n,i})\mu_{\overline{w}_{n,i}}(d\overline{\theta}_{n,i})$ where

$$\begin{aligned}\varphi_{t_{n+1}}(d\overline{w}_{n,i}) &= \mathbb{E}\left[\phi_{\overline{T}_{n,i}}(d\overline{w}_{n,i}) \mid T_{n+1} = t_{n+1}\right] \\ &= \int_0^{t_{n+1}} \frac{\lambda_w(t)\rho(d\overline{w}_{n,i})}{\psi(t)} \frac{\psi(t)}{\Psi(t_{n+1})} dt \\ &= \frac{\Lambda_w(t_{n+1})\rho(d\overline{w}_{n,i})}{\Psi(t_{n+1})}.\end{aligned}$$

□

3.3. Finite iid construction

Note that ξ_{n+1}/n tends to 1 almost surely as n tends to infinity. This therefore suggests the following **finite iid construction**, as an approximation to the truncated measure G_n

$$\widetilde{G}_n = \sum_{i=1}^{n} \widetilde{W}_{n,i}\delta_{\widetilde{\theta}_{n,i}} \tag{3.5a}$$

where for $i = 1, \ldots, n$,

$$\widetilde{W}_{n,i} \overset{\text{i.i.d.}}{\sim} \widetilde{\varphi}_n \tag{3.5b}$$

$$\widetilde{\theta}_{n,i} \mid \widetilde{W}_{n,i} = \widetilde{w}_{n,i} \sim \mu_{\widetilde{w}_{n,i}} \tag{3.5c}$$

with

$$\widetilde{\varphi}_n(dw) = \varphi_{\Psi^{-1}(n)}(dw) = \frac{\Lambda_w(\Psi^{-1}(n))\rho(dw)}{n}. \tag{3.6}$$

***Proposition* 3.1.** Let $\widetilde{G}_n$ be the finite iid approximation defined by Equation (3.5). Then $\widetilde{G}_n$ converges weakly in distribution to $G \sim \mathrm{CRM}(\rho, \mu_w)$ as $n \to \infty$.

Proof. The proof is similar to that of [Lee et al., 2016, Section 3.1]. Let $\mathcal{M}_+(\Theta)$ be the set of non-negative, totally finite measures on Θ, equipped with the topology of weak convergence. Note that, almost surely, $\widetilde{G}_n, G \in \mathcal{M}_+(\Theta)$. It suffices to show that $\mathbb{E}\left[e^{-\widetilde{G}_n(f)}\right] \to \mathbb{E}\left[e^{-\widetilde{G}(f)}\right]$ for all $f \in C_b^+(\Theta)$, the set of measurable, bounded and con-

tinuous functions $f : \Theta \to [0, \infty)$. For any $f \in C_b^+(\Theta)$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{-\widetilde{G}_n(f)}\right] &= \mathbb{E}\left[e^{-\sum_{i=1}^n \widetilde{W}_{n,i} f(\widetilde{\theta}_{n,i})}\right] \\
&= \mathbb{E}\left[e^{-\widetilde{W}_{n,1} f(\widetilde{\theta}_{n,1})}\right]^n \\
&= \left(\int_S e^{-wf(\theta)} \widetilde{\varphi}_n(dw)\mu_w(d\theta)\right)^n \\
&= \left(1 - \int_S (1 - e^{-wf(\theta)})\widetilde{\varphi}_n(dw)\mu_w(d\theta)\right)^n \\
&= \left(1 - \frac{1}{n}\int_S (1 - e^{-wf(\theta)})\Lambda_w(\Psi^{-1}(n))\rho(dw)\mu_w(d\theta)\right)^n.
\end{aligned}$$

First assume $\int_S (1 - e^{-wf(\theta)})\rho(dw)\mu_w(d\theta) < \infty$. Note that $\Lambda_w(\Psi^{-1}(n)) \leq 1, \forall n$ and $\Lambda_w(\Psi^{-1}(n)) \to 1$ as n tends to infinity. By the dominated convergence theorem, we therefore have

$$\int_S (1 - e^{-wf(\theta)})\Lambda_w(\Psi^{-1}(n))\rho(dw)\mu_w(d\theta) \to \int_S (1 - e^{-wf(\theta)})\rho(dw)\mu_w(d\theta)$$

as $n \to \infty$. Additionally, for any real sequence $(a_n)_{n\geq 1}$ converging to a we have $(1 - a_n/n)^n \to e^{-a}$ as $n \to \infty$. We therefore obtain

$$\mathbb{E}\left[e^{-\widetilde{G}_n(f)}\right] \to \exp\left(-\int_S (1 - e^{-wf(\theta)})\rho(dw)\mu_w(d\theta)\right),$$

where the right-handside is equal to the Laplace functional $\mathbb{E}[e^{-G(f)}]$ of the CRM $G \sim \mathrm{CRM}(\rho, \mu_w)$ by Campbell's theorem [Kingman, 1993].

Consider now that $\int_S (1 - e^{-wf(\theta)})\rho(dw)\mu_w(d\theta) = \infty$. Note that $\Lambda_w(\Psi^{-1}(m)) < \Lambda_w(\Psi^{-1}(n))$ for any $m < n$, hence

$$\mathbb{E}\left[e^{-\widetilde{G}_n(f)}\right] \leq \left(1 - \frac{1}{n}\int_S (1 - e^{-wf(\theta)})\Lambda_w(\Psi^{-1}(m))\rho(dw)\mu_w(d\theta)\right)^n.$$

Taking first the limit as $n \to \infty$, then $m \to \infty$ gives $\mathbb{E}\left[e^{-\widetilde{G}_n(f)}\right] \to 0$ by sandwiching. □

For the iid construction, one needs to evaluate $\Psi^{-1}(n)$ only once, and this can be done numerically if there is an analytic form for Ψ. Instead of the distribution $\widetilde{\varphi}_n = \varphi_{\Psi^{-1}(n)}$, we can alternatively use more general distributions $\widetilde{\varphi}_n = \varphi_{f(n)}$ where f is an increasing function such that $\Psi(f(n)) \overset{n\to\infty}{\sim} n$. Proposition 3.1 also holds as the proof can be straightforwardly adapted to this case. Note that if $\Psi(t) \overset{t\to\infty}{\sim} ct^\sigma$ for some constant c and $\sigma > 0$, then we can take $f(n) = (n/c)^{1/\sigma}$. Proposition B.2 in the Supplementary Material can be used to find admissible functions f under generic assumptions on ρ and Λ_w.

4. Examples

We first show how the inverse Lévy and size-biased constructions described in Section 2.3 can be recovered as special cases of the general construction introduced in Section 3. We then derive novel constructions within this framework.

4.1. Deterministic arrival times (inverse-Lévy construction)

Assume that the arrival times are deterministic given the size $W = w$, and inversely proportional to it, that is

$$\lambda_w(dt) = \delta_{1/w}(dt). \tag{4.1}$$

The distribution does not admit a density with respect to the Lebesgue measure, but one can still obtain expressions for the different quantities of interest. We obtain

$$\Lambda_w(t) = \mathbb{1}_{\{t \geq 1/w\}}, \qquad \Psi(t) = \overline{\rho}(1/t), \qquad \Psi^{-1}(\xi) = 1/\overline{\rho}^{-1}(\xi),$$
$$\phi_t(dw) = \delta_{1/t}(dw), \qquad \varphi_t(dw) = \frac{\rho(dw)\mathbb{1}_{\{w>1/t\}}}{\overline{\rho}(1/t)}, \qquad \widetilde{\varphi}_n(dw) = \frac{\rho(dw)\mathbb{1}_{\{w>\overline{\rho}^{-1}(n)\}}}{n}.$$

The sequential construction corresponds to the inverse-Lévy construction described in Section 2.3. The exchangeable representation is similar to the ϵ-truncation of normalized CRMs, used in [Argiento et al., 2016a,b], except that the truncation threshold $\epsilon = 1/T_{n+1}$ is treated as a random variable here.

4.2. Exponential arrival times (size-biased construction)

Consider an exponential arrival time distribution with $\lambda_w(dt) = we^{-wt}dt$ and $\Lambda_w(t) = 1 - e^{-wt}$. This leads to Perman et al. [1992]'s size-biased sequential and exchangeable representations described in Section 2.3. While this construction is not novel, it appears that it provides a novel series representation for the generalized gamma random measure. We also show that the iid representation associated to this arrival time distribution corresponds to the finite approximation proposed by Lee et al. [2016].

Generalized gamma process. In the case of the size measure (2.2) with $\alpha > 0, \sigma \in (0,1)$ and $\tau \geq 0$, we obtain the following sequential construction for the GGP, which appears to be novel

$$W_i \mid \xi_i \sim \text{Gamma}\left(1-\sigma, (\sigma\xi_i/\alpha + \tau^\sigma)^{\frac{1}{\sigma}}\right). \tag{4.2}$$

In Section E.1 in the Supplementary Material, we compare this representation with Rosinski's series representation for the GGP [Rosiński, 2001b, Rosiński, 2007]. The conditional distribution for the exchangeable and iid constructions is given by

$$\varphi_t(dw) = \frac{\sigma w^{-1-\sigma}e^{-\tau w}(1-e^{-tw})dw}{\Gamma(1-\sigma)((t+\tau)^\sigma - \tau^\sigma))}. \tag{4.3}$$

This is the distribution of an exponentially-tilted BFRY distribution [Bertoin et al., 2006, Devroye and James, 2014, Lee et al., 2016], written as etBFRY(σ, t, τ). One can easily simulate from this distribution via inversion, see Section D.2 in the Supplementary Material. Note that

$$\Psi(t) \overset{t\to\infty}{\sim} \alpha t^{\sigma}/\sigma$$

hence we can consider the iid distribution

$$\widetilde{\varphi}_n(dw) = \varphi_{(n\sigma/\alpha)^{1/\sigma}}(dw).$$

This corresponds precisely to the finite-dimensional approximation introduced by Lee et al. [2016] for the GGP, which can therefore be seen as a particular case of our approach.

4.3. Gamma arrival times

As a generalization of the exponential arrival times, consider now a gamma arrival distribution

$$\lambda_w(dt) = \frac{t^{\kappa-1}e^{-\kappa w t}(\kappa w)^{\kappa}}{\Gamma(\kappa)}, \quad \Lambda_w(t) = \frac{\gamma(\kappa, \kappa w t)}{\Gamma(\kappa)},$$

where $\kappa \geq 1$ is a tuning parameter and $\gamma(\kappa, t) = \int_0^t x^{\kappa-1}e^{-x}dx$ is the lower incomplete gamma function. Since $\mathbb{E}[T|w] \to 1/w$ and $\mathrm{Var}(T|w) \to 0$ as $\kappa \to \infty$, T converges in probability hence in distribution to $1/w$, and therefore $\Lambda_w(t) \to \mathbb{1}_{\{t\geq 1/w\}}$ as κ tends to infinity, which corresponds to the arrival time cdf of the inverse Lévy representation. Hence, the construction based on the gamma arrival times bridges between the size-biased ($\kappa = 1$) and inverse-Lévy ($\kappa \to \infty$) constructions.

Generalized gamma process. Consider the generalized gamma process with size measure (2.2) and parameters $\alpha > 0$, $\sigma \in (0, 1)$ and $\tau \geq 0$. We have

$$\psi(t) = \eta \frac{t^{\kappa-1}}{(t+\tau/\kappa)^{\kappa-\sigma}} \quad \text{and} \quad \Psi(t) = \begin{cases} \eta \left(\frac{\tau}{\kappa}\right)^{\sigma} B_{\frac{\kappa t}{\kappa t+\tau}}(\kappa, -\sigma) & \text{if } \tau > 0 \\ \frac{\eta}{\sigma} t^{\sigma} & \text{if } \tau = 0 \end{cases}$$

where $\eta = \frac{\alpha\kappa^{\sigma}\Gamma(\kappa-\sigma)}{\Gamma(\kappa)\Gamma(1-\sigma)}$. For the sequential and exchangeable constructions, we get

$$\phi_t(dw) = \mathrm{Gamma}(w; \kappa-\sigma, \kappa t+\tau)dw \quad \text{and} \quad \varphi_t(dw) \propto w^{-\sigma-1}e^{-\tau w}\gamma(\kappa, \kappa t w)dw.$$

For the iid construction, we can use Eq. (3.6) and estimate $\Psi^{-1}(n)$ numerically or, using Proposition B.2 and Table 1, we can alternatively use $\widetilde{\varphi}_n(dw) = \varphi_{(\sigma n/\eta)^{1/\sigma}}(dw)$. The normalizing constant of φ_t (and therefore $\widetilde{\varphi}_n$) has an analytic expression via standard functions. We call the random variable having distribution φ_t an *exponentially-tilted generalized BFRY* random variable, due to the form of the pdf obtained by exponentially tilting the pdf of generalized BFRY. This distribution has a number of remarkable properties that make it amenable for tractable simulation and posterior inference. Refer to Section D.4 in the Supplementary Material for a more detailed description.

4.4. Inverse gamma arrival times

Consider now an inverse gamma arrival distribution

$$\lambda_w(dt) = \text{iGamma}(t; \kappa, \kappa/w)dt$$

where $\kappa \geq 1$ is a tuning parameter and

$$\text{iGamma}(t; a, b) = \frac{b^a}{\Gamma(a)} t^{-a-1} e^{-b/x}$$

is the pdf of an inverse gamma random variable with shape parameter $a > 0$ and scale parameter $b > 0$. By a similar argument as for the gamma arrival times, we have $\Lambda_w(t) = \frac{\Gamma(\kappa, \kappa/(wt))}{\Gamma(\kappa)} \to \mathbb{1}_{\{t \geq 1/w\}}$ as $\kappa \to \infty$ hence it also admits the inverse Lévy construction as a limiting case. The case $\kappa = 1$ is of particular interest, as it leads to a tractable novel representation for the GGP (see Section E.3 in the Supplementary Material), and provide a novel way of interpreting the classical iid approximation of the beta process.

Beta process. Consider the one-parameter beta process with size measure (2.3) with $\sigma = 0$ and $c = 1$. The bijective transformation $u = -(\alpha \log(w))^{-1}$ gives the measure $\rho(du) = u^{-2}du$ on $(0, \infty)$. Note that $\rho(du)$ is not a Lévy measure, but we can nonetheless use our construction as the tail Lévy intensity is finite. Using the inverse gamma kernel with $\kappa = 1$, we obtain $\Psi(t) = t$ and the iid distribution $\widetilde{\varphi}_n(du) = \text{iGamma}(u; 1, 1/n)du$. Applying the inverse transformation $\widetilde{W}_i = e^{-1/(\alpha \widetilde{U}_i)}$, we obtain

$$\widetilde{W}_i \sim \text{Beta}(\alpha/n, 1),$$

which corresponds to the classical iid approximation for the beta process, described in Equation (1.2). The iid construction for the beta process can alternatively be recovered using the arrival time distribution $\Lambda_w(t) = w^{\frac{\alpha}{t}}$ directly with (2.3), without change of variable.

4.5. Generalized Pareto arrival time

Consider the arrival time distribution $\lambda_w(dt) = \frac{cw}{(tw+1)^{c+1}}dt$ where $c > 0$.

Stable beta process. Consider the stable beta process with measure (2.3) with $\sigma > 0$. We have $\Psi(t) = \frac{\alpha c}{\sigma}((t+1)^\sigma - 1)$ and

$$\phi_t(dw) = \frac{w^{-\sigma}(1-w)^{c+\sigma-1}(tw+1)^{-c-1}\mathbb{1}_{\{0<w<1\}}dw}{B(1-\sigma, c+\sigma)(t+1)^{\sigma-1}}$$

$$\varphi_t(dw) = \frac{\sigma w^{-1-\sigma}(1-w)^{c+\sigma-1}}{cB(1-\sigma, c+\sigma)((t+1)^\sigma - 1)}\left(1 - \frac{1}{(tw+1)^c}\right)\mathbb{1}_{\{0<w<1\}}dw.$$

These distributions admit the same conjugacy properties as the beta distribution, and one can sample exactly from these distributions as detailed in Section E.4 in the Supplementary Material.

5. Discussion

As mentioned in the introduction, there is a large literature on series representations for CRMs. Campbell et al. [2019] provided a recent survey of existing and novel (nested) series representations together with a truncation analysis. Our series construction can be seen as a special case of Rosinski's shot-noise series representation [Rosiński, 2001a] (which includes as special cases most series constructions, see Rosiński [2001a]), using the disintegration

$$\rho(dw) = \int_0^\infty \nu(\xi, dw)d\xi$$

where

$$\nu(\xi, dw) = \lambda_w(\Psi^{-1}(\xi))\rho(dw)/\psi(\Psi^{-1}(\xi))$$

is a Markov kernel (noting that $\int_0^\infty \lambda_w(\Psi^{-1}(\xi))/\psi(\Psi^{-1}(\xi)))d\xi = 1$). Nested series representations have also been recently proposed for some specific CRMs [Paisley et al., 2011, 2012, Roychowdhury and Kulis, 2015]. Regarding finite iid approximations, the literature is sparser. Finite iid representations can be obtained using the infinite divisibility properties of CRMs [Kingman, 1975] but as noted by Lee et al. [2016], it generally does not lead to tractable representations, except in the gamma process case. Other ways of obtaining iid constructions, which does not rely on a latent Poisson construction, are described in [Huggins et al., 2017] for some family of CRMs. Note that our construction, while very general, does not encompass all standard iid approximations. For example, the standard iid approximation for the Gamma process $\widetilde{W}_{n,i} \sim \text{Gamma}(\alpha/n, 1)$ cannot be recovered with our construction.

6. Application to Posterior Inference in normalized completely random measure mixture models

We consider here an application to posterior inference in mixture models [Lo, 1984] where the mixing distribution is a normalized CRM (NCRM) [Regazzini et al., 2003]. Such models have been developed by Lijoi et al. [2005, 2007] and James et al. [2009]; see also the overview by Lijoi and Prünster [2010].

Assume that the data $Y_1, \ldots, Y_m \in \mathcal{Y}$ are drawn from a continuous probability density function $f_G(y)$ admitting the following mixture form

$$f_G(y) = \frac{1}{G(S)} \int_\Theta k_\theta(y) dG(d\theta) = \frac{1}{\sum_{i=1}^\infty W_i} \sum_{i=1}^\infty W_i k_{\theta_i}(y) \tag{6.1}$$

where k_θ is a continuous probability density function on $\mathcal{Y}$ for any $\theta \in \Theta$ and $G \sim \text{CRM}(\rho, H)$ is an infinite-activity CRM, where ρ satisfies conditions (2.1). These conditions ensure that $0 < G(S) < \infty$ almost surely. Introducing latent variables $(\vartheta_1, \ldots, \vartheta_m)$,

the model admits the following hierarchical construction,

$$\text{for } j = 1, \dots, m, \quad \vartheta_j \mid G \overset{\text{i.i.d.}}{\sim} \frac{G}{G(S)}, \quad Y_j | \vartheta_j \overset{\text{ind.}}{\sim} k_{\vartheta_j}.$$

We also introduce allocation variables $s_j \in \mathbb{N}$ such that $\vartheta_j = \theta_{s_j}$, indicating from which component of the CRM originates the latent variable ϑ_j.

6.1. Slice Sampler

Slice sampling for infinite mixture models has been originally proposed by Walker [2007] for the mixing distribution being a Dirichlet Process, and extended by Kalli et al. [2011] to stick-breaking priors. Slice samplers have also been developed for mixture models with a Poisson-Kingman prior [Favaro and Walker, 2013, Favaro et al., 2014].

We consider here an extension of the slice sampling algorithm proposed by Griffin and Walker [2011] for NCRM mixture models (see also [Favaro and Teh, 2013]). Griffin and Walker [2011] proposed two slice samplers, Slice 1 and Slice 2. We focus here on extending the Slice 1 algorithm; a generalized version of slice 2 can be developed along the same lines.

Let $N' = \sum_{i=1}^{\infty} \delta_{(W_i, \theta_i, T_i)}$ be a Poisson random measure on $S' = (0, \infty) \times \Theta \times (0, \infty)$ with mean measure $\nu'(dw, d\theta, dt) = \rho(dw)\mu_w(d\theta)\lambda_w(dt)$, as introduced in Section 3. Then $G = \int_0^{\infty} \int_0^{\infty} wN'(dw, \cdot, dt) \sim \text{CRM}(\rho, \mu_w)$ and the mixture model of Equation (6.1) can be equivalently written as

$$f_{N'}(y) = \frac{1}{G(S)} \int_{S'} w k_\theta(y) dN'(dw, d\theta, dt) \tag{6.2}$$

For each observation Y, we introduce a slice variable U which has joint density with Y given by

$$f_{N'}(u, y) = \frac{1}{G(S)} \int_{S'} \mathbb{1}_{\{u < 1/t\}} w t k_\theta(y) dN'(dw, d\theta, dt) \tag{6.3}$$

$$= \frac{1}{\sum_{i=1}^{\infty} W_i} \sum_{i=1}^{\infty} \mathbb{1}_{\{u < 1/T_i\}} W_i T_i k_{\theta_i}(y) \tag{6.4}$$

This slice data augmentation generalizes the approach of Griffin and Walker [2011], which is obtained as a special case. Indeed, under the deterministic arrival time distribution (4.1), we have $T_i = 1/W_i$ and Equation (6.4) reduces to

$$\frac{1}{\sum_{i=1}^{\infty} W_i} \sum_{i=1}^{\infty} \mathbb{1}_{\{u < W_i\}} k_{\theta_i}(y)$$

which is the classical slice sampling augmentation for mixture models, used by Griffin and Walker [2011]. The slice sampler of Griffin and Walker [2011] can be rather straightforwardly adapted to the more general slice variables. As the derivations are quite lengthy,

and require the introduction of a number of additional latent variables, details are given in Section F of the Supplementary material. We show there that, by carefully choosing the arrival time distribution, one can obtain a Gibbs sampler with simpler and more tractable updates than the one based on the deterministic arrival times.

6.2. Inference for the approximate iid model

An alternative to performing posterior inference for the original model, is to perform inference on a finite-dimensional approximation. We approximate the infinite dimensional process G with a finite iid process $\widetilde{G}_n$ using gamma arrival times with $\kappa \geq 1$. Using Proposition B.2, we have

$$\widetilde{\varphi}_n(dw) = \text{etgBFRY}\left(w; \kappa, \sigma, \left(\frac{\sigma\Gamma(\kappa)\Gamma(1-\sigma)n}{\alpha\Gamma(\kappa-\sigma)}\right)^{\frac{1}{\sigma}}, \tau\right)dw. \tag{6.5}$$

where the pdf etgBFRY is defined in Equation (D.7) in the Supplementary Material. Note that we used the function $f(n) = \left(\frac{\sigma\Gamma(\kappa)\Gamma(1-\sigma)n}{\alpha\kappa^{\sigma}\Gamma(\kappa-\sigma)}\right)^{\frac{1}{\sigma}}$ in place of $\Psi^{-1}(n)$, thus both evaluation of the pdf and sampling can be done without any numerical approximation. Adding a latent variable $V \sim \text{Gamma}(n, \sum_{i=1}^{n} \widetilde{w}_{n,i})$ and the allocation variable $s_{1:m}$, the joint density of the mixture model is proportional to

$$v^{n-1}e^{-v\sum_i \widetilde{w}_{n,i}} \prod_{i=1}^{n} \widetilde{w}_{n,i}^{m_i} \widetilde{\varphi}_n(\widetilde{w}_{n,i}) \left[\prod_{j:s_j=i} k_{\widetilde{\theta}_{n,i}}(y_j)\right] h(\widetilde{\theta}_{n,i}), \tag{6.6}$$

The etgBFRY is a conjugate prior for a gamma likelihood, and one can sample exactly from this distribution, as discussed in Section D.4 in the Supplementary Material. It is therefore straightforward to derive a Gibbs sampler where all the conditional distributions are available. Alternatively, approximate inference algorithms can also be derived, such as variational inference or stochastic gradient MCMC as in [Lee et al., 2016].

7. Truncation error analysis

7.1. Error on functionals of the CRM

For a measurable function $f : S \to \mathbb{R}_+$ such that $N(f) < \infty$ a.s., the error term associated with the truncation is defined as

$$R_n = N(f) - N_n(f) = \sum_{i>n} f(W_i, \theta_i). \tag{7.1}$$

Taking for example $f(w, \theta) = w$ corresponds to the L_1 error between the CRM W and W_n.

***Proposition* 7.1.** For $\xi \sim \text{Gamma}(n+1,1)$, R_n has the following moment generating function

$$\mathbb{E}[e^{-\lambda R_n}] = \mathbb{E}_\xi\left[\exp\left(-\int_S (1-e^{-\lambda f(w,\theta)})(1-\Lambda_w(\Psi^{-1}(\xi)))\rho(dw)\mu_w(dw)\right)\right]. \quad (7.2)$$

Proof. By the marking theorem for Poisson point processes [Kingman, 1993, Chapter 5], given $T_{n+1} = t_{n+1}$, the random measure $\sum_{i|T_i \geq t_{n+1}} \delta_{X_i}$ is a Poisson random measure with mean measure $(1-\Lambda_w(t_{n+1}))\rho(dw)\mu_w(d\theta)$. The result follows from Campbell's theorem and the fact that $T_{n+1} = \Psi^{-1}(\xi_{n+1})$. □

We now consider results for the special case $f(w,\theta) = w$. The mean and variance of the truncation error R_n given T_{n+1} are given by

$$\mathbb{E}[R_n \mid T_{n+1}] = \int_0^\infty w(1-\Lambda_w(T_{n+1}))\rho(dw)$$
$$\mathbb{V}[R_n \mid T_{n+1}] = \int_0^\infty w^2(1-\Lambda_w(T_{n+1}))\rho(dw).$$

The next proposition provides an asymptotic expression for the error term, giving insights on how the error relates to the choice of the arrival time distribution $\lambda_w(t)$. The proposition makes some assumptions of regular variation on the mean measure ρ. Background on regular variation and Mellin transforms is given in Section A in the Supplementary Material and the proof of Proposition 7.2 is given in Section B in the Supplementary Material.

***Proposition* 7.2.** Assume that the mean measure $\rho(dw)$ is absolutely continuous with respect to the Lebesgue measure with density function $\varrho(w)$ such that

$$\varrho(w) \overset{w\to 0}{\sim} \zeta_0 w^{-1-\sigma} \quad (7.3)$$

where $\sigma \in (0,1)$ and $\zeta_0 > 0$. Assume additionally that $\Lambda_w(t) = 1-k(wt)$ where k is a positive function on $(0,\infty)$ such that its Mellin transform

$$\check{k}(z) = \int_0^\infty t^{-z-1}k(t)dt$$

is defined in some open interval containing $[\sigma-2, \sigma-1]$. Assume additionally that either (i) k is differentiable with derivative k' and that the Mellin transform $\check{k}'$ of k' is defined in some open interval containing $\sigma-1$, or (ii) that $k(x) = \mathbb{1}_{\{x\leq 1\}}$. Then we have

$$R_n \overset{n\to\infty}{\sim} C_1(\sigma)\zeta_0^{1/\sigma}\sigma^{1-1/\sigma}n^{1-1/\sigma} \quad \text{almost surely} \quad (7.4)$$

where the constant $C_1(\sigma)$ is given by $C_1(\sigma) = (1-\sigma)^{-1}$ if $k(x) = \mathbb{1}_{\{x\leq 1\}}$ and $C_1(\sigma) = \frac{\check{k}(\sigma-1)}{(-\check{k}'(\sigma-1))^{1-1/\sigma}}$ if k is differentiable, and only depends on the arrival time distribution Λ_w and σ.

Table 1. Kernels, Mellin transforms and asymptotic constants for different arrival time distributions

	Determin.	Exponential	Gamma	Inv. gamma	Gen. Pareto
$\Lambda_w(t)$	$\mathbb{1}_{\{t\geq 1/w\}}$	$1-e^{-wt}$	$1-\frac{\Gamma(\kappa,\kappa wt)}{\Gamma(\kappa)}$	$\frac{\Gamma(\kappa,\kappa/(wt))}{\Gamma(\kappa)}$	$1-(1+wt)^{-c}$
$k(x)$	$\mathbb{1}_{\{x\leq 1\}}$	e^{-x}	$\frac{\Gamma(\kappa,x\kappa)}{\Gamma(\kappa)}$	$\frac{\gamma(\kappa,\kappa/x)}{\Gamma(\kappa)}$	$(1+x)^{-c}$
$\check{k}(-z)$	z^{-1}	$\Gamma(z)$	$\frac{\Gamma(z+\kappa)}{z\Gamma(\kappa)\kappa^z}$	$\frac{\Gamma(\kappa-z)}{z\Gamma(\kappa)\kappa^{-z}}$	$B(z,c-z)$
$k'(x)$	–	$-e^{-x}$	$\frac{-\kappa^\kappa x^{\kappa-1}e^{-\kappa x}}{\Gamma(\kappa)}$	$-\frac{\kappa^\kappa e^{-\kappa/x}}{x^{\kappa+1}\Gamma(\kappa)}$	$-c(1+x)^{-c-1}$
$-\check{k}'(-z)$	–	$\Gamma(z)$	$\frac{\kappa^{1-z}\Gamma(\kappa+z-1)}{\Gamma(\kappa)}$	$\frac{\kappa^{z-1}\Gamma(-z+\kappa+1)}{\Gamma(\kappa)}$	$cB(z,c+1-z)$
$C_1(\sigma)$	$(1-\sigma)^{-1}$	$\Gamma(1-\sigma)^{1/\sigma}$	$\frac{(\kappa-\sigma)\Gamma(\kappa-\sigma)^{1/\sigma}}{(1-\sigma)\Gamma(\kappa)^{1/\sigma}}$	$\frac{(\Gamma(\kappa+\sigma)/\Gamma(\kappa))^{1/\sigma}}{(1-\sigma)(\kappa+\sigma-1)}$	$\frac{B(1-\sigma,c+\sigma-1)}{(cB(1-\sigma,c+\sigma))^{1-1/\sigma}}$

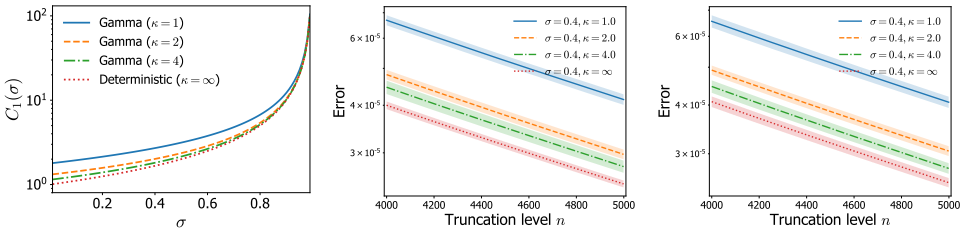


Figure 2. (Left) Constant $C_1(\sigma)$ for deterministic and gamma arrival times; (Middle-Right) Simulated error R_n with gamma arrival times for (Middle) stable process and (Right) GGP.

The deterministic, gamma, inverse-gamma (for $\kappa > 2-\sigma$) and generalized Pareto (for $c > 2-\sigma$) arrival time distributions discussed in Section 4 all verify the assumptions of Proposition 7.2. The associated kernels, Mellin transforms and constants $C_1(\sigma)$ are given in Table 1. Figure 2(a) shows the value of the constant $C_1(\sigma)$ for the deterministic and gamma arrival time, with different values of κ. As indicated in Section 4.3, the approximation gets closer to the deterministic/inverse Lévy construction as κ increases. Both the GGP and the SBP with $\sigma > 0$ verify Equation (7.3), with $\zeta_0 = \frac{\alpha}{\Gamma(1-\sigma)}$ for the GGP and $\zeta_0 = \frac{\alpha}{B(1-\sigma,c+\sigma)}$ for the SBP. We run a simulation study in order to investigate the finite-n properties of the proposed approximations. We report in Figure 2(b-c) the mean and variance of R_n for gamma arrival times, for the stable process and the GGP with $\sigma = 0.4$. For the stable process, we also compare to the inverse-Lévy approximation, as it has an analytic form. As expected, the approximation gets better as the value κ increases. Additional simulations for other arrival time distributions are given in Section G in the Supplementary Material.

7.2. L_1 error on the marginal likelihood

In this section we discuss the L_1 error on the marginal likelihood when truncated CRMs are used for hierarchical Bayesian models under the framework described in [Campbell

et al., 2018]. Let $W \sim \text{CRM}(\rho, \mu_w)$, and W_n be its approximation with n atoms. Let $H(\cdot|w)$ be a probability distribution on $\mathbb{N} \cup \{0\}$ for all w, and denote $\pi(w) := H(0|w)$. Consider a hierarchical Bayesian model for m observations $X_{1:m} := \{X_j\}_{j=1}^m$, $Z_j|w_j \sim H(w_j), \quad X_j|z_j \sim F(z_j), \quad j = 1, \dots, m$, and denote $p_{m,\infty}(X_{1:m})$ the marginal likelihood for this model. Similarly, denote by $p_{m,n}(X_{1:m})$ the marginal likelihood of the model with the same generative process, except for W_n instead of W. Following Campbell et al. [2018], we analyze the quality of approximation by comparing $p_{m,\infty}(X_{1:m})$ and $p_{m,n}(X_{1:m})$. For the inverse Lévy case, one recovers the bound derived in [Campbell et al., 2018, Theorem D.3.].

***Proposition* 7.3.** We have the bound

$$0 \leq \frac{1}{2}||p_{m,\infty}(X_{1:m}) - p_{m,n}(X_{1:m})|| \leq 1 - e^{-B_{m,n}},$$

where

$$B_{m,n} \leq m \int_0^\infty \int_0^\infty (1 - \pi(w))(1 - \Lambda_w(\Psi^{-1}(\xi))\rho(dw)\,\text{Gamma}(\xi; n+1, 1)d\xi.$$

Proof. From [Campbell et al., 2018], we have the protobound

$$0 \leq \frac{1}{2}||p_{m,\infty}(X_{1:m}) - p_{m,n}(X_{1:m})|| \leq 1 - \mathbb{P}(\text{supp}(X_{1:m}) \subseteq \text{supp}(G_n))$$

In our case,

$$\begin{aligned}
&\mathbb{P}(\text{supp}(X_{1:m}) \subseteq \text{supp}(G_n)) = \mathbb{E}\left[\prod_{i=n+1}^\infty \pi(W_i)^m\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{m\sum_{i\geq n+1}\log\pi(W_i)} \mid T_{n+1}\right]\right] \\
&= \mathbb{E}\left[\exp\left(-\int_0^\infty (1 - e^{m\log\pi(w)})(1 - \Lambda_w(T_{n+1}))\rho(dw)\right)\right] \\
&= \int_0^\infty \exp\left(-\int_0^\infty (1 - \pi(w)^m)(1 - \Lambda_w(\Psi^{-1}(\gamma))\rho(dw)\right)\text{Gamma}(\gamma; n+1, 1)d\gamma \\
&\geq \exp\left(-\int_0^\infty \int_0^\infty (1 - \pi(w)^m)(1 - \Lambda_w(\Psi^{-1}(\gamma))\rho(dw)\,\text{Gamma}(\gamma; n+1, 1)d\gamma\right) \\
&\geq \exp\left(-m\int_0^\infty \int_0^\infty (1 - \pi(w))(1 - \Lambda_w(\Psi^{-1}(\gamma))\rho(dw)\,\text{Gamma}(\gamma; n+1, 1)d\gamma\right)
\end{aligned}$$

where the last two inequalities follow from Jensen's inequality and from the fact that $1 - x^m \leq m(1 - x)$ for $m \geq 1$ respectively. □

Acknowledgements

FC acknowledges support from EPSRC under grant EP/P026753/1.

References

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Marlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

J. Arbel and I. Prünster. A moment-matching ferguson & klass algorithm. *Statistics and Computing*, 27(1):3–17, 2017.

R. Argiento, I. Bianchini, and A. Guglielmi. A blocked Gibbs sampler for NGG-mixture models via a priori truncation. *Statistics and Computing*, 26(3):641–661, 2016a.

R. Argiento, I. Bianchini, and A.s Guglielmi. Posterior sampling from ε-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547, 2016b.

F. Ayed and F. Caron. Nonnegative Bayesian nonparametric factor models with completely random measures for community detection. *arXiv:1902.10693*, 2019.

O. E. Barndorff-Nielsen and N. Shephard. Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241, 2001.

E. Barrios, A. Lijoi, L. E. Nieto-Barajas, and I. Prünster. Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334, 2013.

J. Bertoin, T. Fujita, B. Roynette, and M. Yor. On a particular class of self-decomposable random variables : the durations of Bessel excursions straddling independent exponential times. *Probability and Mathematical Statistics*, 26(2):315–366, 2006.

D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

L. Bondesson. On simulation from infinitely divisible distributions. *Advances in Applied Probability*, 14(4):855–869, 1982.

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

D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4249–4257. Curran Associates, Inc., 2016.

T. Campbell, D. Cai, and T. Broderick. Exchangeable trait allocations. *Electronic Journal of Statistics*, 12(2):2290–2322, 2018.

T. Campbell, J. H. Huggins, J. P. How, and T. Broderick. Truncated random measures. *Bernoulli*, 2019.

F. Caron. Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems 25*. 2012.

F. Caron and E. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society B*, 79:1295–1366, 2017. Part 5.

R. Cont and P. Tankov. *Financial modelling with jump processes*, volume 2. CRC press, 2004.

L. Devroye and L. F. James. On simulation and properties of the stable law. *Statistical Methods and Applications*, 23(3):307–343, 2014.

F. Doshi, K. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 137–144, 2009.

S. Favaro and Y. W. Teh. MCMC for normalized random measure mixture models. *Statist. Sci.*, 28(3):335–359, 08 2013. . URL http://dx.doi.org/10.1214/13-STS422.

S. Favaro and S. G. Walker. Slice sampling σ-stable Poisson-Kingman mixture models. *Journal of Computational and Graphical Statistics*, 22(4):830–847, 2013.

S. Favaro, M. Lomeli, N. Bernardo, and Y. W. Teh. On the stick-breaking representation of σ-stable poisson-kingman models. *Electronic Journal of Statistics*, 8(1):1063–1085, 2014. . URL http://dx.doi.org/10.1214/13-STS422.

T. S. Ferguson and M. J. Klass. A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643, 1972.

N. Foti and S. Williamson. Slice sampling normalized kernel-weighted completely random measure mixture models. In *Advances in Neural Information Processing Systems*, pages 2240–2248, 2012.

P. Gopalan, F. J. Ruiz, R. Ranganath, and D. M. Blei. Bayesian nonparametric Poisson factorization for recommendation systems. In *AISTATS*, pages 275–283, 2014.

J. E. Griffin. Sequential monte carlo methods for mixtures with normalized random measures with independent increments priors. *Statistics and Computing*, 27(1):131–145, 2017.

J. E. Griffin and S. G. Walker. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011.

Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, 2006.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294, 1990.

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

J. Huggins, L. Masoero, T. Broderick, and L. Mackey. Generic finite approximations for practical bayesian nonparametrics. In *NIPS Workshop on "Advances in Approximate Bayesian Inference"*, 2017.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

L. F. James. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics*, 33(4):1771–1799, 2005.

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.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

A. Khintchine. Zur theorie der unbeschränkt teilbaren verteilungsgesetze. *Rec. Math. Moscou, n. Ser*, 2:79–117, 1937.

J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21 (1):59–78, 1967. URL http://projecteuclid.org/euclid.pjm/1102992601.

J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):1–15, 1975.

J.F.C. Kingman. *Poisson processes*, volume 3. Oxford University Press, USA, 1993.

J. Lee, L. F. James, and S. Choi. Finite-dimensional BFRY priors and variational Bayesian inference for power-law models. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472): 1278–1291, 2005.

A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

A. Y. Lo. On a class of bayesian nonparametric estimates: I. density estimates. *Ann. Statist.*, 12(1):351–357, 03 1984. . URL http://dx.doi.org/10.1214/aos/1176346412.

J. Moller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, 2003.

R. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

L. E. Nieto-Barajas, I. Prünster, and S. G. Walker. Normalized random measures driven by increasing additive processes. *The Annals of Statistics*, 32(6):2343–2360, 2004.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784. ACM, 2009.

J. Paisley, L. Carin, and D. Blei. Variational inference for stick-breaking Beta process priors. In *ICML*, pages 889–896. Citeseer, 2011.

J. Paisley, D. Blei, and M.l Jordan. Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics*, pages 850–858, 2012.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *Annals of Statistics*, 31(2):560–585, 04 2003. . URL http://dx.doi.org/10.1214/aos/1051027881.

J. Rosiński. On series representations of infinitely divisible random vectors. *The Annals of Probability*, pages 405–430, 1990.

J. Rosiński. Series representations of Lévy processes from the perspective of point processes. In Ole E. Barndoff-Nielson, Thomas Mikosch, and Sidney I. Resnick, editors, *Lévy Processes - Theory and Applications*, pages 401–415. Birkhäuser, Boston, 2001a.

J. Rosiński. Contribution to the discussion of the paper "Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics" by O. Barndorff-Nielsen and N. Shephard. *Journal of the Royal Statistical Society, Series B*, 63:230–231, 2001b.

J. Rosiński. Tempering stable processes. *Stochastic processes and their applications*, 117 (6):677–707, 2007.

A. Roychowdhury and B. Kulis. Gamma processes, stick-breaking, and variational inference. In *Artificial Intelligence and Statistics*, pages 800–808, 2015.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In *Advances in neural information processing systems*, pages 1838–1846, 2009.

Y. W. Teh, D. Gorür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 556–563, 2007.

S. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.

R. L. Wolpert and K. Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267, 1998a.

R. L. Wolpert and K. Ickstadt. Simulation of Lévy random fields. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 227–242. Springer, 1998b.

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Artificial Intelligence and Statistics*, pages 1135–1143, 2015.

M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1462–1471, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Submitted to Bernoulli
arXiv: arXiv:1905.10733

Supplementary Material for A unified construction for series representations and finite approximations of completely random measures

JUHO LEE[1,*] XENIA MISCOURIDOU[2,**] and FRANÇOIS CARON[2,†]

[1] *AITRICS, Seoul, Republic of Korea*
E-mail: [*] juho@aitrics.com

[2] *Department of Statistics, University of Oxford*
E-mail: [**] xenia.miscouridou@spc.ox.ac.uk; [†] caron@stats.ox.ac.uk,

Appendix A: Background on regular variation and Mellin transforms

This background material comes from the book of Bingham et al. [1989].

A.1. Definitions

***Definition* A.1** (Slowly varying function)**.** A function $\ell : (0,\infty) \to (0,\infty)$ is *slowly varying* at infinity if for all $c > 0$,

$$\frac{\ell(cx)}{\ell(x)} \to 1 \text{ as } x \to \infty. \tag{A.1}$$

***Definition* A.2** (Regularly varying function)**.** A function $f : (0,\infty) \to (0,\infty)$ is *regularly varying* at infinity with exponent $\rho \in \mathbb{R}$ if $f(x) = x^{\rho}\ell(x)$ for some slowly varying ℓ. A function f is regularly varying at 0 if $f(1/x)$ is regularly varying at infinity, i.e., $f(x) = x^{-\rho}\ell(1/x)$ for some $\rho \in \mathbb{R}$ and slowly varying ℓ.

A.2. Basic theorems for regularly varying functions

Let U be a regularly function with exponent ρ and slowly varying function ℓ locally bounded on $(0,\infty)$.

***Theorem* A.1** (Karamata's theorem)**.** [Bingham et al., 1989, Propositions 1.5.8 and 1.5.10]. Suppose that $U(t) \overset{t\to\infty}{\sim} t^\rho \ell(t)$.

- When $\rho > -1$,
$$\int_0^x U(t)dt \overset{x\to\infty}{\sim} \frac{1}{\rho+1} x^{\rho+1}\ell(x).$$
- When $\rho < -1$,
$$\int_x^\infty U(t)dt \overset{x\to\infty}{\sim} -\frac{1}{\rho+1} x^{\rho+1}\ell(x).$$

***Corollary* A.1.** This also holds when U is regularly varying at 0. When $\rho < -1$ and $U(s) \overset{s\to 0}{\sim} s^\rho \ell(1/s)$,
$$\int_x^\infty U(s)ds \overset{x\to 0}{\sim} \frac{-1}{\rho+1} x^{\rho+1}\ell(1/x).$$

A.3. Generalized Abelian theorem

***Definition* A.3.** Given a measurable kernel $k : (0,\infty) \to \infty$ let
$$\check{k}(z) = \int_0^\infty t^{-z-1}k(t)dt = \int_0^\infty u^{z-1}k(1/u)du \tag{A.2}$$
be its Mellin transform, for $z \in \mathbb{C}$ such that the integral converges.

***Remark* A.1.** If $k(x)$ has a Mellin transform $\check{k}(z)$ which converges in (z_1, z_2), then $h(x) = k(1/x)$ has a Mellin transform $\check{h}(z) = \check{k}(-z)$ which converges in $(-z_2, -z_1)$.

***Theorem* A.2.** [Bingham et al., 1989, Theorem 4.1.6 page 201] Let the Mellin transform $\check{k}$ of k converge at least in the strip $\sigma \le \mathrm{Re}(z) \le \tau$, where $-\infty < \sigma < \tau < \infty$. Let $\rho \in (\sigma, \tau)$, ℓ a slowly varying function, $c \in \mathbb{R}$. If f is measurable, $f(x)/x^\sigma$ is bounded on every interval $(0, a]$ and
$$f(x) \overset{x\to\infty}{\sim} cx^\rho \ell(x)$$
then
$$\int_0^\infty k(x/t)f(t)t^{-1}dt \overset{x\to\infty}{\sim} c\check{k}(\rho)x^\rho\ell(x).$$

The next result is a trivial corollary of Theorem A.2, considering limits as x tends to 0.

***Corollary* A.2.** Let the Mellin transform $\check{k}$ of k converge at least in the strip $\tau_1 \leq \mathrm{Re}(z) \leq \tau_2$, where $-\infty < \tau_1 < \tau_2 < \infty$. Let $\rho \in (\tau_1, \tau_2)$, ℓ a slowly varying function, $c \in \mathbb{R}$. If f is measurable, $f(x)x^{-\tau_2}$ is bounded on every interval $[a, \infty)$ and

$$f(x) \overset{x\to 0}{\sim} cx^{\rho}\ell(1/x)$$

then

$$\int_0^{\infty} k(x/t)f(t)t^{-1}dt \overset{x\to 0}{\sim} c\check{k}(\rho)x^{\rho}\ell(1/x).$$

Proof.

$$\begin{aligned}\int_0^{\infty} k(x/t)f(t)t^{-1}dt &= \int_0^{\infty} k(xu)f(1/u)u^{-1}du \\ &= \int_0^{\infty} \widetilde{k}\left(1/\left(xu\right)\right)\widetilde{f}(u)u^{-1}du\end{aligned}$$

where $\widetilde{f}(x) = f(1/x)$, $\widetilde{f}(x)/x^{-\tau_2}$ bounded on every interval $(0, 1/a]$ with

$$\widetilde{f}(x) = f(1/x) \overset{x\to\infty}{\sim} cx^{-\rho}\ell(x)$$

and $\widetilde{k}(x) = k(1/x)$ is such that its Mellin transform converges in the strip $-\tau_2 \leq \mathrm{Re}(z) \leq -\tau_1$. Theorem A.2 above therefore gives the result. □

Appendix B: Proof of Proposition 7.2

We state a slightly more general version of Proposition 7.2, where the constant ζ_0 in Equation (7.3) can more generally be any slowly varying function $\ell_0(1/x)$. We then prove this generalized proposition.

***Proposition* B.1.** [Slight generalization of Proposition 7.2] Assume that the mean measure $\rho(dw)$ is absolutely continuous with respect to the Lebesgue measure with density function $\varrho(w)$ such that

$$\varrho(w) \overset{w\to 0}{\sim} w^{-1-\sigma}\ell_0(1/w) \tag{B.1}$$

where $\sigma \in (0, 1)$ and ℓ_0 is a slowly varying function. Assume additionally that $\Lambda_w(t) = 1 - k(wt)$ where k is a positive function on $(0, \infty)$ such that its Mellin transform (see Definition A.3) $\check{k}$ converges in some open interval containing $[\sigma - 2, \sigma - 1]$. Assume additionally that either (i) k is differentiable with derivative k' and that the Mellin transform $\check{k}'$ of k' is defined in some open interval containing $\sigma - 1$, or (ii) that $k(x) = \mathbb{1}_{\{x\leq 1\}}$. Then we have

$$R_n \overset{n\to\infty}{\sim} C_1(\sigma)n^{1-1/\sigma}\ell_{**}(n) \quad \text{almost surely} \tag{B.2}$$

where ℓ_{**} is some slowly varying function that depends on ℓ_0 and σ but not λ_w and the constant $C_1(\sigma)$ is given by $C_1(\sigma) = (1-\sigma)^{-1}$ if $k(x) = \mathbb{1}_{\{x\leq 1\}}$ and $C_1(\sigma) = \frac{\check{k}(\sigma-1)}{\left(-\check{k}'(\sigma-1)\right)^{1-1/\sigma}}$ if k is differentiable, and only depends on the arrival time distribution Λ_w and σ.

In order to prove Proposition B.1, we first state the following proposition.

***Proposition* B.2.** Assume that

$$\overline{\rho}(x) \overset{x\to 0}{\sim} x^{-\sigma}\ell(1/x) \tag{B.3}$$

where $\sigma \in (0,1)$ and ℓ is a slowly varying function. Assume additionally that

$$\Lambda_w(t) = 1 - k(wt) \tag{B.4}$$

where k is a positive and differentiable function on $(0,\infty)$, with derivative k'. Assume that the Mellin transform $\check{k}'$ of k' is defined in some open interval containing $\sigma - 1$. Then

$$\Psi(t) \overset{t\to\infty}{\sim} -\check{k}'(\sigma-1)t^{\sigma}\ell(t) \quad \text{and} \quad \Psi^{-1}(t) \overset{t\to\infty}{\sim} \left(-\check{k}'(\sigma-1)\right)^{-1/\sigma} t^{1/\sigma}\ell_*(t) \tag{B.5}$$

where ℓ_* is another slowly varying function depending on ℓ and σ, and defined in Equation (B.8). If $\ell(x) = c$ is constant, then we simply have $\ell_*(x) = c^{-1/\sigma}$. In particular, this is the case for both the generalized gamma process (GGP) and the stable beta process (SBP), which verify condition (B.3) when $\sigma > 0$ with $\ell(x) = \frac{\alpha}{\sigma\Gamma(1-\sigma)}$ for the GGP and $\ell(x) = \frac{\alpha}{\sigma B(1-\sigma,c+\sigma)}$ for the SBP.

Proof of Proposition B.2. The assumptions (B.1) and the first condition in Equation (3.1) both imply that

$$\lim_{w\to 0} \Lambda_w(t)\overline{\rho}(w) = 0.$$

Using integration by parts

$$\begin{aligned}
\Psi(t) &= \int_0^\infty \Lambda_w(t)\rho(dw) \\
&= \left[-\Lambda_w(t)\overline{\rho}(w)\right]_{w=0}^{\infty} + \int_0^\infty \frac{\partial\Lambda_w(t)}{dw}\overline{\rho}(w)dw \\
&= \int_0^\infty \frac{\partial\Lambda_w(t)}{dw}\overline{\rho}(w)dw.
\end{aligned}$$

Now assume $\Lambda_w(t) = 1 - k(wt)$ where k is differentiable on $(0,\infty)$. Then

$$\frac{\partial\Lambda_w(t)}{dw} = -tk'(wt).$$

If the Mellin transform $\check{k}'$ of k' is defined in some open interval containing $\sigma - 1$, then Corollary A.2 implies

$$\Psi(t) \overset{t\to\infty}{\sim} -\check{k}'(\sigma - 1)t^{\sigma}\ell(t). \tag{B.6}$$

In the case $k(x) = \mathbb{1}_{\{x\leq 1\}}$, k is not differentiable, but we have directly

$$\Psi(t) = \overline{\rho}(1/t) \overset{t\to\infty}{\sim} t^{\sigma}\ell(t).$$

Now we use inversion formulas for regularly varying function to get the asymptotic regime for $\Psi^{-1}(t)$. Assume $\sigma > 0$, then [Gnedin et al., 2007, Lemma 22] implies

$$\Psi^{-1}(t) \overset{t\to\infty}{\sim} \left(-\check{k}'(\sigma - 1)\right)^{-1/\sigma} t^{1/\sigma}\ell_*(t) \tag{B.7}$$

where ℓ_* is a slowly varying function defined by

$$\ell_*(t) = \left(\ell^{1/\sigma}\left(t^{1/\sigma}\right)\right)^{\#} \tag{B.8}$$

where $\ell^{\#}$ denotes the de Bruijn conjugate of the slowly varying function ℓ [Bingham et al., 1989, Theorem 1.5.13]. Note that ℓ_* only depends on ℓ and σ, but not the arrival time distribution $\Lambda_w(t)$.

□

We now prove Proposition B.1. Assume that the mean measure $\rho(dw)$ is absolutely continuous with respect to the Lebesgue measure with density function $\varrho(w)$ verifying Equation (B.1) for some $\sigma \in (0, 1)$ and some slowly varying function ℓ_0. [Bingham et al., 1989, Proposition 1.5.8] imply that

$$\overline{\rho}(x) \overset{x\to 0}{\sim} x^{-\sigma}\ell(1/x) \tag{B.9}$$

where ℓ is a slowly varying function defined by

$$\ell(1/x) = \ell_0(1/x)/\sigma.$$

Assume additionally that $\Lambda_w(t) = 1 - k(wt)$ where k is a positive function on $(0, \infty)$ such that its Mellin transform (see Definition A.3) $\check{k}$ converges in some open interval containing $[\sigma - 2, \sigma - 1]$. Note that

$$\mathbb{E}[R_n \mid T_{n+1}] = \int_0^{\infty} w(1 - \Lambda_w(T_{n+1}))\rho(dw)$$
$$\mathbb{V}[R_n \mid T_{n+1}] = \int_0^{\infty} w^2(1 - \Lambda_w(T_{n+1}))\rho(dw).$$

T_{n+1} tends to infinity almost surely as n tends to infinity and then Corollary A.2 implies

$$\mathbb{E}[R_n \mid T_{n+1}] \overset{n\to\infty}{\sim} \check{k}\,(\sigma - 1)\,T_{n+1}^{\sigma-1}\ell_0\,(T_{n+1}) \tag{B.10}$$
$$\mathbb{V}[R_n \mid T_{n+1}] \overset{n\to\infty}{\sim} \check{k}\,(\sigma - 2)\,T_{n+1}^{\sigma-2}\ell_0\,(T_{n+1}) \tag{B.11}$$

almost surely.

As $T_{n+1} = \Psi^{-1}(\xi_{n+1})$ where $\xi_{n+1} \overset{n\to\infty}{\sim} n$ almost surely, using Proposition B.2 we obtain

$$T_{n+1} \overset{n\to\infty}{\sim} \left(-\check{k}'(\sigma-1)\right)^{-1/\sigma} n^{1/\sigma}\ell_*(n) \tag{B.12}$$

almost surely. Combining Equation (B.12) with Equations (B.10) and (B.11), we obtain

$$\mathbb{E}[R_n \mid T_{n+1}] \overset{n\to\infty}{\sim} \frac{\check{k}(\sigma-1)}{\left(-\check{k}'(\sigma-1)\right)^{1-1/\sigma}} n^{1-1/\sigma}\ell_*^{\sigma-1}(n)\ell_0\left(n^{1/\sigma}\ell_*(n)\right) \tag{B.13}$$

$$\mathbb{V}[R_n \mid T_{n+1}] \overset{n\to\infty}{\sim} \frac{\check{k}(\sigma-2)}{\left(-\check{k}'(\sigma-1)\right)^{1-2/\sigma}} n^{1-2/\sigma}\ell_*^{\sigma-2}(n)\ell_0\left(n^{1/\sigma}\ell_*(n)\right) \tag{B.14}$$

almost surely. Note that if $\ell_0(t) = \zeta_0$ is constant, then all the other slowly varying functions are also constant with

$$\ell(x) = \frac{\zeta_0}{\sigma}, \quad \ell_*(x) = \ell(x)^{-1/\sigma} = \left(\frac{\zeta_0}{\sigma}\right)^{-1/\sigma}. \tag{B.15}$$

Finally, Equation (B.2) follows similarly to the proof of [Gnedin et al., 2007, Proposition 2]. Using Chebyshev's inequality

$$\Pr\left(\left|\frac{R_n}{\mathbb{E}[R_n \mid T_{n+1}]} - 1\right| > \epsilon \mid T_{n+1}\right) \leq \frac{\mathbb{V}[R_n \mid T_{n+1}]}{\epsilon^2 \mathbb{E}[R_n \mid T_{n+1}]^2}$$

Take $a_n = n^2$. Using Equations (B.13) and (B.14)

$$\frac{\mathbb{V}[R_n \mid T_{n+1}]}{\mathbb{E}[R_n \mid T_{n+1}]^2} \overset{n\to\infty}{\sim} n^{-1}L(n),$$

for some slowly varying function L. By the Borel-Cantelli lemma, given T_n,

$$R_{a_n} \overset{n\to\infty}{\sim} \mathbb{E}[R_{a_n} | T_{a_n+1}]$$

almost surely. As R_n is decreasing, we have, for any $m^2 \leq n \leq (m+1)^2$

$$\frac{R_{(m+1)^2}}{\mathbb{E}[R_{m^2} \mid T_{m^2+1}]} \leq \frac{R_n}{\mathbb{E}[R_n \mid T_{n+1}]} \leq \frac{R_{m^2}}{\mathbb{E}[R_{(m+1)^2} \mid T_{(m+1)^2+1}]}$$

and it follows by sandwiching that

$$R_n \overset{n\to\infty}{\sim} \mathbb{E}[R_n | T_{n+1}]$$

almost surely. Combining this with Equation (B.13) gives the final result, with ℓ_{**} the slowly varying function defined by

$$\ell_{**}(t) = \ell_*^{\sigma-1}(t)\ell_0\left(t^{1/\sigma}\ell_*(t)\right).$$

Note that in the case of the GGP, the different slowly varying functions are all constant functions

$$\ell_0(x) = \frac{\alpha}{\Gamma(1-\sigma)} \tag{B.16}$$

$$\ell(x) = \frac{\alpha}{\sigma\Gamma(1-\sigma)} \tag{B.17}$$

$$\ell_*(x) = \ell(x)^{-1/\sigma} = \left(\frac{\alpha}{\sigma\Gamma(1-\sigma)}\right)^{-1/\sigma} \tag{B.18}$$

$$\ell_{**}(x) = \left(\frac{\alpha}{\sigma\Gamma(1-\sigma)}\right)^{-1+1/\sigma} \frac{\alpha}{\Gamma(1-\sigma)} = \left(\frac{\alpha}{\Gamma(1-\sigma)}\right)^{1/\sigma} \sigma^{1-1/\sigma} \tag{B.19}$$

For the SBP, we have

$$\ell_0(x) = \frac{\alpha}{B(1-\sigma, c+\sigma)} \tag{B.20}$$

$$\ell(x) = \frac{\alpha}{\sigma B(1-\sigma, c+\sigma)} \tag{B.21}$$

$$\ell_*(x) = \ell(x)^{-1/\sigma} = \left(\frac{\alpha}{\sigma B(1-\sigma, c+\sigma)}\right)^{-1/\sigma} \tag{B.22}$$

$$\ell_{**}(x) = \left(\frac{\alpha}{B(1-\sigma, c+\sigma)}\right)^{1/\sigma} \sigma^{1-1/\sigma} \tag{B.23}$$

Appendix C: Mellin transforms

C.1. Deterministic kernel

Take $k(t) = \mathbb{1}_{\{t\leq 1\}}$. Then

$$\check{k}(z) = \int_0^\infty t^{-z-1} k(t) dt = \int_0^1 t^{-z-1} = -z^{-1}$$

if $z < 0$.

C.2. Gamma kernel

Take $k(t) = \frac{\Gamma(\kappa,\kappa t)}{\Gamma(\kappa)}$ for $\kappa \geq 1$.

$$\begin{aligned}
\check{k}(z) &= \int_0^\infty u^{-z-1} k(u) du \\
&= \frac{1}{\Gamma(\kappa)} \int_0^\infty u^{-z-1} \int_{\kappa u}^\infty v^{\kappa-1} e^{-v} dv du \\
&= \frac{1}{\Gamma(\kappa)} \int_0^\infty \int_0^{v/\kappa} u^{-z-1} v^{\kappa-1} e^{-v} du dv \\
&= \frac{-\kappa^z}{z\Gamma(\kappa)} \int_0^\infty v^{\kappa-z-1} e^{-v} dv \\
&= \frac{-\Gamma(\kappa-z)\kappa^z}{z\Gamma(\kappa)}
\end{aligned}$$

which converges for $z < 0$. Additionally, $k'(t) = -\frac{\kappa^\kappa t^{\kappa-1} e^{-\kappa t}}{\Gamma(\kappa)}$ hence

$$\begin{aligned}
\check{k}'(z) &= -\int_0^\infty t^{-z-1} k'(t) dt \\
&= -\frac{\kappa^\kappa}{\Gamma(\kappa)} \int_0^\infty t^{\kappa-z-2} e^{-\kappa t} dt \\
&= -\frac{\Gamma(\kappa-z-1)\kappa^{z+1}}{\Gamma(\kappa)}
\end{aligned}$$

which converges for $z < \kappa - 1$.

C.3. Inverse gamma kernel

Take $k(t) = \frac{\gamma(\kappa,\kappa/t)}{\Gamma(\kappa)}$. Note that if $\kappa = 1$, $k(t) = (1 - e^{-1/t})$. Then

$$\begin{aligned}
\check{k}(z) &= \int_0^\infty t^{-z-1} k(t) dt = \int_0^\infty u^{z-1} k(1/u) du \\
&= \int_0^\infty u^{z-1} \frac{\gamma(\kappa,\kappa u)}{\Gamma(\kappa)} du = \frac{1}{\Gamma(\kappa)} \int_0^\infty u^{z-1} \int_0^{\kappa u} v^{\kappa-1} e^{-v} dv du \\
&= \frac{1}{\Gamma(\kappa)} \int_0^\infty v^{\kappa-1} e^{-v} \int_{v/\kappa}^\infty u^{z-1} du dv \\
&= -\frac{1}{z\Gamma(\kappa)\kappa^z} \int_0^\infty v^{\kappa+z-1} e^{-v} dv \text{ if } z < 0 \\
&= -\frac{\Gamma(\kappa+z)}{z\Gamma(\kappa)\kappa^z} \text{ if } \kappa + z > 0
\end{aligned}$$

therefore defined for $z \in (-\kappa, 0)$. Note that $\check{k}(z) \to -1/z$ as κ tends to infinity, which corresponds to the inverse-Lévy case.

We have

$$k'(t) = -\frac{\kappa^\kappa}{t^{\kappa+1}\Gamma(\kappa)} e^{-\kappa/t}$$

and

$$\begin{aligned}\check{k}'(z) &= \int_0^\infty t^{-z-1} k'(t) dt = \int_0^\infty u^{z-1} k'(1/u) du \\ &= -\frac{\kappa^\kappa}{\Gamma(\kappa)} \int_0^\infty u^{z+\kappa} e^{-\kappa u} du \\ &= -\frac{\kappa^{-z-1}\Gamma(z+\kappa+1)}{\Gamma(\kappa)}\end{aligned}$$

defined for $z \in (-1-\kappa, \infty)$. Note again that $\check{k}'(z) \to -1$ as $\kappa \to \infty$ (inverse Lévy case).

C.4. Generalized Pareto kernel

Take $k(t) = \frac{1}{(t+1)^c}$. Then

$$\check{k}(z) = \int_0^\infty \frac{t^{-z-1}}{(t+1)^c} dt = B(-z, c+z)$$

for $z \in (-c, 0)$. We have $k'(t) = -\frac{c}{(t+1)^{c+1}}$ hence

$$\check{k}'(z) = -c \int_0^\infty \frac{t^{-z-1}}{(t+1)^{c+1}} dt = -cB(-z, c+1+z)$$

defined for $z \in (-c-1, 0)$.

Appendix D: BFRY and related distributions

D.1. BFRY distribution

The BFRY distribution, first named in Devroye and James [2014] after the work of Bertoin, Fujita, Roynette, and Yor [Bertoin et al., 2006], arises much earlier in various contexts [Pitman and Yor, 1997, Winkel, 2005]. Recently it was highlighted in Lee et al. [2016] as a finite-dimensional approximate distribution for stable process, generalized gamma process, and a special case of the stable beta process. The density of a BFRY distribution with parameter $\sigma \in (0, 1)$ is written as

$$\mathrm{BFRY}(w; \sigma) = \frac{\sigma w^{-1-\sigma}(1 - e^{-w}).}{\Gamma(1-\sigma)} \tag{D.1}$$

One can easily verify that the distribution can be simulated as a ratio of a independent gamma and beta random variables.

$$G \sim \text{Gamma}(1-\sigma, 1), \quad B \sim \text{Beta}(\sigma, 1), \quad \frac{G}{B} \stackrel{d}{=} \text{BFRY}(\sigma). \tag{D.2}$$

D.2. Exponentially-tilted BFRY distribution

In Lee et al. [2016], James et al. [2015], the exponentially-tilted version of BFRY distribution was discussed. The density of exponentially-tiled random variable with parameters $\sigma \in (0,1)$, $c > 0$ and $\tau > 0$ is

$$\text{etBFRY}(w; \sigma, t, \tau) = \frac{\sigma w^{-1-\sigma} e^{-\tau w}(1 - e^{-tw})}{\Gamma(1-\sigma)\left((t+\tau)^{\sigma} - \tau^{\sigma}\right)}. \tag{D.3}$$

Then it is easy to show that for $G \sim \text{Gamma}(1-\sigma, 1)$ and $U \sim \text{Unif}(0,1)$,

$$G \times \left((t+\tau)^{\sigma}(1-U) + \tau^{\sigma} U\right)^{-\frac{1}{\sigma}} \stackrel{d}{=} \text{etBFRY}(\sigma, t, \tau). \tag{D.4}$$

D.3. Generalized BFRY distribution

The generalized BFRY distribution, introduced by Ayed et al. [2019], is a generalization of the BFRY distribution. A generalized BFRY random variable with parameter $\sigma \in (0,1)$ and $\kappa > \sigma$ is obtained as

$$G \sim \text{Gamma}(\kappa - \sigma, 1), \quad B \sim \text{Beta}(\sigma, 1), \quad \frac{G}{B} \stackrel{d}{=} \text{gBFRY}(\kappa, \sigma). \tag{D.5}$$

By a simple algebra, we obtain the density as

$$\text{gBFRY}(w; \kappa, \sigma) = \frac{\sigma w^{-1-\sigma} \gamma(\kappa, w)}{\Gamma(\kappa - \sigma)}, \tag{D.6}$$

where $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function.

D.4. Exponentially-tilted generalized BFRY distribution

The density of the exponentially-tilted generalized BFRY distribution with parameter $\sigma \in (0,1), \kappa > \sigma, t > 0$ and $\tau > 0$ is

$$\text{etgBFRY}(w; \kappa, \sigma, t, \tau) = \frac{w^{-\sigma-1} e^{-\tau w} \gamma(\kappa, \kappa t w)}{\tau^{\sigma} \Gamma(\kappa - \sigma) B_{\frac{\kappa t}{\kappa t + \tau}}(\kappa, -\sigma)}, \tag{D.7}$$

where $B_x(\cdot,\cdot)$ is the incomplete beta function. A random variable having this distribution can be simulated by rejection sampling. Alternatively, using the power series expansion for the incomplete gamma function, we have

$$\begin{aligned}
\text{etgBFRY}(w;\kappa,\sigma,t,\tau) &\propto w^{-\sigma-1}e^{-\tau w}\gamma(\kappa,tw) \\
&= \sum_{j=0}^{\infty} \frac{\Gamma(\kappa)(\kappa t w)^{j+\kappa}e^{-\kappa t w}}{\Gamma(\kappa+j+1)} w^{-\sigma-1}e^{-\tau w} \\
&= \sum_{j=0}^{\infty} \frac{\Gamma(\kappa)}{\Gamma(\kappa+j+1)} (\kappa t)^{\kappa+j} w^{\kappa+j-\sigma-1}e^{-(\kappa t+\tau)w} \\
&= \sum_{j=0}^{\infty} \frac{(\kappa t)^{\kappa+j}\Gamma(\kappa)\Gamma(\kappa+j-\sigma)}{(\kappa t+\tau)^{\kappa+j-\sigma}\Gamma(\kappa+j+1)} \,\text{Gamma}(w;\kappa+j-\sigma,\kappa t+\tau).
\end{aligned}$$

which means that the distribution is an infinite mixture of gamma distributions

$$\text{etgBFRY}(w;\kappa,\sigma,t,\tau) = \sum_{j=0}^{\infty} \pi_j \,\text{Gamma}(w;\kappa+j-\sigma,\kappa t+\tau) \tag{D.8}$$

with mixing proportions

$$\pi_j = \frac{\Gamma(\kappa)(\kappa t+\tau)^{\sigma}}{\tau^{\sigma}\Gamma(\kappa-\sigma)B_{\frac{\kappa t}{\kappa t+\tau}}(\kappa,-\sigma)} \left(\frac{\kappa t}{\kappa t+\tau}\right)^{\kappa+j} \frac{\Gamma(\kappa+j-\sigma)}{\Gamma(\kappa+j+1)}. \tag{D.9}$$

Hence, one can sample from the exponentially-tilted generalized BFRY distribution by first sampling the component j from the above infinite discrete distribution, then sampling from a gamma distribution with the corresponding parameters.

The exponentially-tilted generalized BFRY distribution has a nice property of being a conjugate prior for Poisson, gamma, normal with fixed mean, and Pareto distributions. Let $W \sim \text{etgBFRY}(\kappa,\sigma,t,\tau)$. Then, for Poisson,

$$X|W \sim \text{Poisson}(\lambda W) \Rightarrow W|X \sim \text{etgBFRY}(\kappa,\sigma-X,t,\tau+\lambda). \tag{D.10}$$

For gamma,

$$X|W \sim \text{Gamma}(a,W) \Rightarrow W|X \sim \text{etgBFRY}(\kappa,\sigma-a,t,\tau+X). \tag{D.11}$$

For normal,

$$X|W \sim \mathcal{N}(\mu,1/W) \Rightarrow W|X \sim \text{etgBFRY}\left(\kappa,\sigma-\frac{1}{2},t,\tau+\frac{(X-\mu)^2}{2}\right). \tag{D.12}$$

For Pareto,

$$X|W \sim \text{Pareto}(x_0,W) \Rightarrow W|X \sim \text{etgBFRY}(\kappa,\sigma-1,t,\tau+\log(X/x_0)). \tag{D.13}$$

D.5. Inverse generalized BFRY

One can also consider the counterpart of generalized BFRY distribution where gamma is replaced with inverse gamma. We define the inverse generalized BFRY distribution, whose pdf is written as

$$\begin{aligned}\text{igBFRY}(w;\kappa,\sigma) &= \frac{\sigma}{\Gamma(\kappa+\sigma)} w^{-1-\sigma}\Gamma(\kappa, w^{-1}) \\ &= \frac{\sigma}{\Gamma(\kappa+\sigma)} \int_1^\infty w^{-\kappa-\sigma-1} v^{\kappa-1} e^{-v/w} dv \\ &= \int_1^\infty v^{-1} \times \frac{(w/v)^{-\kappa-\sigma-1} e^{-v/w}}{\Gamma(\kappa+\sigma)} \times \sigma v^{-\sigma-1} dv \\ &= \int_0^1 u \times \text{iGamma}(wu;\kappa+\sigma,1) \times \text{Beta}(u;\sigma,1) du.\end{aligned}$$

Hence, one can realize that

$$G \sim \text{iGamma}(\kappa+\sigma,1), \quad B \sim \text{Beta}(\sigma,1), \quad \frac{G}{B} \stackrel{d}{=} \text{igBFRY}(\kappa,\sigma). \tag{D.14}$$

D.6. Exponentially-tilted inverse generalized BFRY

Finally, we consider an exponentially tilted version of inverse generalized BFRY distribution, whose pdf is written as

$$\text{etigBFRY}(w;\kappa,\sigma,t,\tau) \propto w^{-1-\sigma} e^{-\tau w} \Gamma(\kappa, (tw)^{-1}).$$

Unfortunately, we do not have an analytic expression for the normalizing constant. We can still sample from this distribution via rejection sampling. This distribution arises as the truncated exchangeable density $\varphi_t(w)$ of the generalized gamma process with inverse gamma arrival times.

Appendix E: Detailed derivations of the results in Section 4 and additional examples

E.1. Exponential arrival times

Generalized gamma. In the case of the size measure (2.2) with $\alpha > 0, \sigma \in (0,1)$ and $\tau \geq 0$, we have

$$\Psi(t) = \frac{\alpha}{\sigma}\{(t+\tau)^\sigma - \tau^\sigma\}, \quad \psi(t) = \frac{\alpha}{(\tau+t)^{1-\sigma}}. \tag{E.1}$$

The arrival times are thus generated as

$$T_i = \Psi^{-1}(\xi_i) = \left(\frac{\sigma}{\alpha}\xi_i + \tau^\sigma\right)^{\frac{1}{\sigma}} - \tau. \tag{E.2}$$

The conditional distribution for the sequential construction is

$$\phi_t(dw) = \text{Gamma}(w; 1-\sigma, t+\tau)dw. \tag{E.3}$$

In summary, the sequential construction for the GGP, is given by

$$W_i \mid \xi_i \sim \text{Gamma}\left(1-\sigma, \left(\frac{\sigma}{\alpha}\xi_i + \tau^\sigma\right)^{\frac{1}{\sigma}}\right). \tag{E.4}$$

Comparison to Rosinski's series representation for the GGP Rosinski Rosiński [2001], Rosiński [2007] proposed the following series representation for the GGP/tempered stable process

$$W_i = \min\left(\left(\frac{\xi_i \sigma \Gamma(1-\sigma)}{\alpha}\right)^{-1/\sigma}, e_i u_i^{1/\sigma}\right)$$

where $e_i \overset{iid}{\sim} \text{Exp}(\tau)$, $u_i \overset{iid}{\sim} \text{Unif}(0,1)$. For i large, $W_i = \left(\frac{\xi_i \sigma \Gamma(1-\sigma)}{\alpha}\right)^{-1/\sigma}$ with high probability, which corresponds to the inverse-Lévy construction for the stable process, and this construction has the same asymptotic error rate as the inverse-Lévy construction for the GGP. The asymptotic error of Rosinski's representation is therefore lower than the asymptotic error for the series defined by Equation (E.4), by a factor $\Gamma(1-\sigma)^{1/\sigma}(1-\sigma) \in (1,2)$, according to Table 1 in the main paper.

E.2. Gamma arrival times

Generalized gamma. Consider the generalized gamma process with (2.2), $\alpha > 0$, $\sigma \in (0,1)$ and $\tau \geq 0$. We have

$$\begin{aligned}
\psi(t) &= \int_0^\infty \frac{\kappa^\kappa w^\kappa t^{\kappa-1} e^{-\kappa w t}}{\Gamma(\kappa)} \times \frac{\alpha}{\Gamma(1-\sigma)} w^{-\sigma-1} e^{-\tau w} dw \\
&= \frac{\alpha \kappa^\sigma \Gamma(\kappa-\sigma)}{\Gamma(\kappa)\Gamma(1-\sigma)} \frac{t^{\kappa-1}}{(t+\tau/\kappa)^{\kappa-\sigma}} \\
\Psi(t) &= \int_0^t \psi(s) ds \\
&= \frac{\alpha \kappa^\sigma \Gamma(\kappa-\sigma)}{\Gamma(\kappa)\Gamma(1-\sigma)} \int_0^t \frac{s^{\kappa-1}}{(s+\tau/\kappa)^{\kappa-\sigma}} ds \\
&= \begin{cases} \frac{\alpha \tau^\sigma \Gamma(\kappa-\sigma)}{\Gamma(\kappa)\Gamma(1-\sigma)} B_{\frac{\kappa t}{\kappa t+\tau}}(\kappa, -\sigma) & \text{if } \tau > 0 \\ \frac{\alpha \kappa^\sigma \Gamma(\kappa-\sigma)}{\sigma \Gamma(\kappa)\Gamma(1-\sigma)} t^\sigma & \text{if } \tau = 0 \end{cases}
\end{aligned}$$

where $B_x(a,b)$ is the incomplete beta function. For $\tau = 0$, Ψ^{-1} has the analytic expression

$$\Psi^{-1}(\xi) = \left(\frac{\sigma\Gamma(\kappa)\Gamma(1-\sigma)\xi}{\alpha\kappa^\sigma\Gamma(\kappa-\sigma)}\right)^{\frac{1}{\sigma}}.$$

For $\tau > 0$, there is no analytic expression for Ψ^{-1}. For the sequential construction, we get

$$\phi_t(dw) = \text{Gamma}(w; \kappa - \sigma, \kappa t + \tau)dw. \tag{E.5}$$

For the exchangeable and iid constructions, we obtain

$$\varphi_t(dw) = \begin{cases} \frac{w^{-\sigma-1}e^{-\tau w}\gamma(\kappa,\kappa t w)}{\tau^\sigma\Gamma(\kappa-\sigma)B_{\frac{\kappa t}{\kappa t+\tau}}(\kappa,-\sigma)}dw & \text{if } \tau > 0 \\ \frac{\sigma w^{-1-\sigma}\gamma(\kappa,\kappa t w)}{(\kappa t)^\sigma\Gamma(\kappa-\sigma)} & \text{if } \tau = 0 \end{cases} \tag{E.6}$$

When $\tau = 0$, $\varphi_t(dw)$ is the distribution of a generalized BFRY distribution (Section D.3). When $\tau > 0$, $\varphi_t(dw)$ corresponds to the distribution of exponentially-tilted generalized BFRY (Section D.4).

E.3. Inverse gamma arrival times

Take

$$\lambda_w(dt) = \text{iGamma}(t; \kappa, \kappa/w)dt$$
$$\Lambda_w(t) = \frac{\Gamma(\kappa, \kappa/(wt))}{\Gamma(\kappa)}.$$

Stable process. Consider the stable process with size measure (2.2), $\alpha > 0$, $\sigma \in (0,1)$ and $\tau = 0$. We have

$$\psi(t) = \frac{\alpha\Gamma(\kappa+\sigma)}{\kappa^\sigma\Gamma(1-\sigma)\Gamma(\kappa)t^{1-\sigma}}, \quad \Psi(t) = \frac{\alpha\Gamma(\kappa+\sigma)t^\sigma}{\sigma\kappa^\sigma\Gamma(1-\sigma)\Gamma(\kappa)}, \tag{E.7}$$

$$T_i = \left(\frac{\sigma\kappa^\sigma\Gamma(\kappa)\Gamma(1-\sigma)\gamma_i}{\alpha\Gamma(\kappa+\sigma)}\right)^{\frac{1}{\sigma}}. \tag{E.8}$$

For the sequential construction, we have

$$\phi_t(dw) = \text{iGamma}(w; \kappa + \sigma, \kappa/t)dw. \tag{E.9}$$

For the exchangeable construction, we get

$$\varphi_t(dw) = \frac{\sigma\kappa^\sigma}{t^\sigma\Gamma(\kappa+\sigma)}w^{-1-\sigma}\Gamma\left(\kappa, \frac{\kappa}{tw}\right). \tag{E.10}$$

which correspond to an inverse generalized BFRY distribution. We can sample from this as $X \sim \text{iGamma}(\kappa+\sigma,1), Y \sim \text{Beta}(\sigma,1), W = \frac{\kappa X}{tY}$. See Section D.5 for more details.

The case $\kappa = 1$ is of particular interest, as it leads to a tractable novel representation for the GGP, and provide a novel way of interpreting the classical iid approximation of the beta process.

Generalized gamma process. Consider GGP with size measure (2.2) with $\alpha > 0, \sigma \in [0,1)$ and $\tau > 0$. Take inverse gamma arrival time with $\kappa = 1$. We have

$$\psi(t) = \frac{2\alpha(\tau t)^{\frac{\sigma+1}{2}} K_{-\sigma-1}(2\sqrt{\tau/t})}{t^2\Gamma(1-\sigma)}, \quad \Psi(t) = \frac{2\alpha(\tau t)^{\frac{\sigma}{2}} K_{-\sigma}(2\sqrt{\tau/t})}{\Gamma(1-\sigma)}, \tag{E.11}$$

where $K_\nu(\cdot)$ is a modified Bessel function of the second kind. Unfortunately, the arrival time is not given analytically, so we may resort to a numerical root finding algorithm to compute $T_i = \Psi^{-1}(\xi_i)$. The sequential construction is then given by

$$\phi_t(dw) = \frac{(\tau t)^{\frac{-\sigma-1}{2}} w^{-\sigma-2} e^{-\tau w - \frac{1}{wt}} dw}{2K_{-\sigma-1}(2\sqrt{\tau/t})} = \text{GIG}(w; -\sigma-1, 2\tau, 2/t)dw, \tag{E.12}$$

where $\text{GIG}(w; p, a, b)$ is the pdf of a generalized inverse Gaussian distribution with parameters $p \in \mathbb{R}, a > 0, b > 0$. The exchangeable construction is given by

$$\varphi_t(dw) = \frac{(\tau t)^{-\frac{\sigma}{2}} w^{-\sigma-1} e^{-\tau w - \frac{1}{wt}} dw}{2K_{-\sigma}(2\sqrt{\tau/t})} = \text{GIG}(w; -\sigma, 2\tau, 2/t)dw. \tag{E.13}$$

This particular case, which seems to be novel to the best of our knowledge, is useful because it covers the gamma process ($\sigma = 0$). It also includes the stable process ($\tau = 0$) as its limiting case i.e. as $\tau \to 0$,

$$\phi_t(dw) \to \text{iGamma}(w; \sigma+1, 1/t)dw, \quad \varphi_t(dw) \to \text{iGamma}(w; \sigma, 1/t)dw. \tag{E.14}$$

The sequential construction is impractical since we have to invert Ψ for each t_i, but the exchangeable construction requires only one inversion for t_{n+1}.

Beta process. Consider the beta process:

$$\rho(dw) = \alpha w^{-1} \mathbb{1}_{\{0<w<1\}} dw. \tag{E.15}$$

Take the bijective transformation $u = -(\alpha \log(w))^{-1}$ which gives the measure on $(0,\infty)$

$$\rho(du) = u^{-2} du \tag{E.16}$$

Note that $\rho(du)$ is not a Lévy measure. Using the inverse gamma kernel with $\kappa = 1$ to obtain a series approximation for $\rho(du)$, we obtain

$$\Psi(t) = \int_0^\infty w^{-2} e^{-1/(wt)} dw = t \tag{E.17}$$

and for the iid model we have $\widetilde{U}_i \sim \widetilde{\phi}_n(du)$ where

$$\widetilde{\phi}_n(du) = \frac{1}{n} u^{-2} e^{-1/(nu)} du$$

which is the distribution of an inverse gamma random variable with parameter $(1, 1/n)$. Setting the inverse transformation $\widetilde{W}_i = e^{-1/(\alpha \widetilde{U}_i)}$, we obtain

$$\widetilde{W}_i \sim \text{Beta}(\alpha/n, 1)$$

which corresponds to the classical iid approximation for the beta process, described in the introduction.

This construction can also be obtained directly with different arrival time distribution. Consider

$$\lambda_w(dt) = \frac{-\alpha w^{\frac{\alpha}{t}} \log w}{t^2} dt, \quad \Lambda_w(t) = w^{\frac{\alpha}{t}}. \tag{E.18}$$

A sample from this distribution can be obtained as

$$t' \sim \text{Exp}(\log w^{-\alpha}), \quad t = 1/t'. \tag{E.19}$$

Then we have

$$\Psi(t) = t, \quad \psi(t) = 1, \tag{E.20}$$

and as a result

$$\phi_t(dw) = -\frac{\alpha}{t^2} w^{\frac{\alpha}{t}-1} \log w \mathbb{1}_{\{0<w<1\}} dw, \quad \varphi_t(dw) = \frac{\alpha}{t} w^{\frac{\alpha}{t}-1} \mathbb{1}_{\{0<w<1\}} dw \tag{E.21}$$

A sample from $\phi_t(w)$ can be obtained by

$$v \sim \text{Gamma}(2, 1/t), \quad w = e^{-v/\alpha}, \tag{E.22}$$

and the exchangeable construction correspond to the iid beta approximation.

E.4. Generalized Pareto arrival time distribution

Consider the following arrival time distribution,

$$\lambda_w(dt) = \frac{cw}{(tw+1)^{c+1}} dt, \quad \Lambda_w(t) = 1 - \frac{1}{(tw+1)^c} \tag{E.23}$$

where $c > 0$.

Stable beta process. Consider the stable beta process with Lévy measure

$$\rho(dw) = \frac{\alpha}{B(1-\sigma, c+\sigma)} w^{-1-\sigma}(1-w)^{c+\sigma-1} \mathbb{1}_{\{0<w<1\}} dw. \tag{E.24}$$

With change of variable $v = \frac{w}{1-w}$, we see that

$$\begin{aligned}
\Psi(t) &= \int_0^1 \frac{\alpha w^{-1-\sigma}(1-w)^{c+\sigma-1}}{B(1-\sigma, c+\sigma)} \left(1 - \frac{1}{(tw+1)^c}\right) dw \\
&= \frac{\alpha}{B(1-\sigma, c+\sigma)} \int_0^\infty v^{-1-\sigma} \Big((1+v)^{-c} - ((t+1)v+1)^{-c} \Big) dv \\
&= \frac{\alpha}{\Gamma(c) B(1-\sigma, c+\sigma)} \int_0^\infty \int_0^\infty v^{-1-\sigma} e^{-vy} (1 - e^{-tyv}) y^{c-1} e^{-y} dy dv \\
&= \frac{\alpha \Gamma(1-\sigma)}{\sigma \Gamma(c) B(1-\sigma, c+\sigma)} \int_0^\infty \left((t+1)^\sigma - 1\right) y^{c+\sigma-1} e^{-y} dy \\
&= \frac{\alpha c}{\sigma} \left((t+1)^\sigma - 1\right).
\end{aligned}$$

It follows that $\psi(t) = \alpha c (t+1)^{\sigma-1}$. For the sequential model we obtain

$$\phi_t(dw) = \frac{w^{-\sigma}(1-w)^{c+\sigma-1}(tw+1)^{-c-1} \mathbb{1}_{\{0<w<1\}} dw}{B(1-\sigma, c+\sigma)(t+1)^{\sigma-1}}. \tag{E.25}$$

The random variable $V = \frac{W}{W-1}$ has distribution

$$\begin{aligned}
\overline{\phi}_t(dv) &= \frac{1}{B(1-\sigma, c+\sigma)(t+1)^{\sigma-1}} v^{-\sigma}(tv+v+1)^{-c-1} dv \\
&= \frac{dv}{B(1-\sigma, c+\sigma)(t+1)^{\sigma-1}\Gamma(c+1)} \int_0^\infty v^{-\sigma} y^c e^{-tvy - vy - y} dy \\
&= dv \int_0^\infty y \cdot \frac{(t+1)^{1-\sigma}(vy)^{-\sigma} e^{-(t+1)vy}}{\Gamma(1-\sigma)} \cdot \frac{y^{c+\sigma-1} e^{-y}}{\Gamma(c+\sigma)} dy,
\end{aligned}$$

and thus a sample from ϕ_t can be obtained as

$$W = \frac{V}{V+1} \tag{E.26}$$

where

$$V = \frac{Z}{Y}, \quad Z \sim \text{Gamma}(1-\sigma, 1+t), \quad Y \sim \text{Gamma}(c+\sigma, 1). \tag{E.27}$$

For the exchangeable model, we have

$$\varphi_t(dw) = \frac{\sigma w^{-1-\sigma}(1-w)^{c+\sigma-1}}{cB(1-\sigma, c+\sigma)((t+1)^\sigma - 1)} \left(1 - \frac{1}{(tw+1)^c}\right) \mathbb{1}_{\{0<w<1\}} dw, \tag{E.28}$$

and by a similar calculation we see that a sample from φ_t can be obtained as

$$Z \sim \text{etBFRY}(\sigma, t, 1), \quad Y \sim \text{Gamma}(c+\sigma, 1), \quad W = \frac{Z/Y}{Z/Y+1}. \tag{E.29}$$

Appendix F: Derivation of the slice sampler for normalized CRM mixture models

For ease of presentation, we assume that ρ and H are absolutely continuous with respect to the Lebesgue measure, and admit densities ϱ and h respectively.

F.1. Data augmentation

Adding the allocation variables s_j, $j = 1, \ldots, m$ and latent variable $V \sim \text{Gamma}(n, W(S))$, we obtain the joint pdf

$$f_{N'}(y_{1:m}, u_{1:m}, s_{1:m}, v) = v^{n-1} e^{-vW(S)} \prod_{j=1}^{m} \mathbb{1}_{\{u_j < 1/t_{s_j}\}} w_{s_j} t_{s_j} k_{\theta_{s_j}}(y_j)$$

where $y_{1:m} = (y_1, \ldots, y_m)$ and similarly for $u_{1:m}$ and $s_{1:m}$.

Let $n = \max(s_j)$, $L = T_n$, $m_i = \sum_{j=1}^{m} \mathbb{1}_{\{s_j = i\}}$ and $K = \sum_{i=1}^{n} \mathbb{1}_{\{m_i > 0\}}$. Denote $w_* = \sum_{i=n+1}^{\infty} w_i$. The joint distribution of $(y_{1:m}, s_{1:m}, v, w_{1:n}, t_{1:n}, \theta_{1:n}, w_*, n)$ ($u_{1:m}$ being integrated out), is proportional to [Pitman, 2003]

$$\begin{aligned}
&p(y_{1:m}, s_{1:m}, v, w_{1:n}, t_{1:n}, \theta_{1:n}, w_*, n) \\
&\propto v^{n-1} e^{-vw_*} g_*(w_*) \prod_{i:m_i>0} w_i^{m_i} e^{-vw_i} \varrho(w_i) h(\theta_i) \lambda_{w_i}(t_i) \prod_{j|s_j=i} k_{\theta_i}(y_j) \\
&\times \left[\prod_{i \le n | m_i = 0} e^{-vw_i} \varrho(w_i) h(\theta_i) \lambda_{w_i}(t_i) \mathbb{1}_{\{t_i < L\}} \right] e^{-\int_0^\infty \varrho(w) \Lambda_w(L) dw}
\end{aligned}$$

where g_* is the pdf of a random variable with moment generating function $\mathbb{E}[e^{-\xi w_*}] = e^{-\psi_*(\xi)}$, where

$$\psi_*(\xi) = \int_0^\infty (1 - e^{-\xi w}) \varrho(w) (1 - \Lambda_w(L)) dw. \tag{F.1}$$

Integrating over w_* gives

$$\begin{aligned}
&p(y_{1:m}, s_{1:m}, v, w_{1:n}, t_{1:n}, \theta_{1:n}, n) \\
&\propto v^{n-1} e^{-\psi_*(v)} \prod_{i:m_i>0} w_i^{m_i} e^{-vw_i} \varrho(w_i) h(\theta_i) \lambda_{w_i}(t_i) \prod_{j|s_j=i} k_{\theta_i}(y_j) \\
&\times \left[\prod_{i \le n | m_i = 0} e^{-vw_i} \varrho(w_i) h(\theta_i) \lambda_{w_i}(t_i) \mathbb{1}_{\{t_i < L\}} \right] e^{-\int_0^\infty \varrho(w) \Lambda_w(L) dw}
\end{aligned}$$

F.2. Gibbs sampler

We consider two versions of the Gibbs sampler. One where w_* is updated within the Gibbs sampler, and a second version where w_* is integrated out, as in Griffin and Walker [2011]. As we will see in the next section, for the GGP with an exponential arrival time distribution, one does not need to integrate out w_* and can obtain tractable updates. We assume here additionally that the Lévy measure ρ is parameterized by some parameter ζ that we wish to estimate, and write $\rho(w;\zeta)$ for the measure and $p(\zeta)$ for the prior on ζ.

Step 1. Update the allocation variables $s_{1:m}$. The allocation variables are updated according to

$$p(s_j = i \mid \text{rest}) \propto \mathbb{1}_{\{u_j < 1/t_i\}} w_i t_i k_{\theta_i}(y_j)$$

where $\{i \mid u_j < 1/t_i\}$ is a finite set for any j.

Step 2. Update v. When conditioning on w_*, the full conditional for v is Gamma $\left(n, w_* + \sum_{i=1}^n w_i\right)$. Otherwise, when w_* is integrated out, the conditional is proportional to

$$v^{n-1} e^{-v \sum_{i=1}^n w_i} e^{-\psi_*(v)}$$

which is typically of non-standard form, and can be sampled via rejection as suggested by Griffin and Walker [2011], or Metropolis-Hastings.

Step 3. Update $\theta_{1:n}$ The elements θ_i are conditionally independent. For i such that $m_i > 0$, we have

$$h(\theta_i) \prod_{j \mid s_j = i} k_{\theta_i}(y_j).$$

Step 4. Update jointly $(w_{1:n}, t_{1:n})$, n and the slice variables $u_{1:m}$. Let $K = \sum_{i=1}^n \mathbb{1}_{\{m_i > 0\}}$. Reorder the points $(w_i, t_i\, \theta_i)$ so that $m_i > 0$ for all $i = 1, \ldots, K$.

(a) For each $i = 1, \ldots, K$, sample w_i from the conditional proportional to

$$w_i^{m_i} e^{-v w_i} \varrho(w_i;\zeta)$$

then sample t_i given w_i from λ_{w_i}. Set $L = \max_{i=1,\ldots,K} t_i$.

(b) Sample K' points (w_{K+i}, t_{K+i}), $i = 1, \ldots, K'$ from a Poisson point process with mean measure $e^{-vw} \varrho(w;\zeta) \lambda_w(t) \mathbb{1}_{\{t<L\}}$. This can be done by first sampling K' from a (finite) Poisson distribution with mean

$$\int_0^\infty e^{-vw} \varrho(w;\zeta) \Lambda_w(L) dw, \tag{F.2}$$

then sample the points iid from the density proportional to

$$e^{-vw} \varrho(w;\zeta) \lambda_w(t) \mathbb{1}_{\{t<L\}}. \tag{F.3}$$

Finally, set $n = K + K'$ and reorder the n points according to their arrival time so that $t_1 < t_2 < \ldots < t_n$.

(c) For $j = 1, \ldots, m$, sample $u_j \sim \text{Unif}(0, 1/t_{s_j})$

Step 5. Update jointly ζ and w_*. The conditional for the parameter ζ, (with w_* marginalised out) is proportional to

$$p(\zeta)\prod_{i=1}^{n}\rho(w_i;\zeta)\times e^{-\int_0^\infty \varrho(w;\zeta)\Lambda_w(L)dw-\psi_*(v;\zeta)}$$

If needed, w_* can be sampled according to its full distribution, proportional to $e^{-vw_*}g_*(w_*)$. This is the distribution of a random variable $\widetilde{w}_*$ with moment generating function $\mathbb{E}[e^{-t\widetilde{w}_*}]=e^{-\widetilde{\psi}(t)}$ where $\widetilde{\psi}(t)=\int_0^\infty(1-e^{-wt})e^{-vt}\varrho(w)(1-\Lambda_w(L))dw$.

F.2.1. Comparison to Griffin and Walker's slice sampler for the GGP

The slice sampler 1 of Griffin and Walker [2011] is a special case of the sampler described above. Indeed, under the deterministic arrival time distribution (4.1), we have $T_i=1/W_i$ and Equation (6.4) reduces to

$$\frac{1}{\sum_{i=1}^\infty W_i}\sum_{i=1}^{\infty}\mathbb{1}_{\{u<W_i\}}k_{\theta_i}(y)$$

which is the classical slice sampling augmentation for mixture models, used by Griffin and Walker [2011].

It may appear unclear why one would want to use a more general slice sampler, based on a different series representation, as the inverse Lévy provides the best possible approximation. Using a different arrival time distribution will lead to a larger number n of points sampled at each iteration (see Section 7 in the main paper for an asymptotic analysis of the truncation error). We nonetheless show that, in the case of the GGP, the use of an alternative arrival time distribution leads to more tractable Gibbs sampler updates. Steps 1, 3 and 4(a) are essentially similar, and we focus on Steps 2, 4(b) and 5.

The sampler of Griffin and Walker requires to resort to numerical integration of the one-dimensional integrals (F.1) and (F.2) and to rejection sampling in steps 2, 4(b) and 5. Consider that we use instead an exponential arrival time distribution with $\lambda_w(dt)=we^{-wt}dt$ and $\Lambda_w(t)=1-e^{-wt}$. All the integrals involved are analytically tractable. (F.1) and (F.2) reduce to

$$\psi_*(\xi)=\frac{\alpha}{\sigma}\left((\tau+L+\xi)^\sigma-(\tau+L)^\sigma\right)$$
$$\int_0^\infty e^{-vw}\varrho(w)\Lambda_w(L)dw=\frac{\alpha}{\sigma}\left((\tau+L+v)^\sigma-(\tau+v)^\sigma\right)$$

In step 5, we have $e^{-vt}\rho(w)(1-\Lambda_w(L))=\frac{1}{\Gamma(1-\sigma)}w^{-1-\sigma}e^{-(v+L+\tau)w}$ is of the form of the mean measure of a GGP with parameters $\alpha,\sigma,\tau+v+L$. $\widetilde{w}_*$ has therefore the distribution of the total mass of a GGP, which is an exponentially tilted stable random variable. One can sample exactly from this distribution [Devroye, 2009]. Conditioning on w_*, the full conditional in Step 2 is therefore a simple gamma distribution.

Finally, in Step 4(b), one needs to sample a potentially large number K' of points from a pdf proportional to (F.3). In the case of the deterministic arrival time, (F.3) is not of standard form, and Griffin and Walker [2011] suggest to use rejection sampling. For exponential arrival times, one can sample exactly from (F.3), without resorting to rejection. One can first sample the weights from the pdf proportional to

$$w^{-1-\sigma} e^{-(\tau+v)w}(1 - e^{-Lw})$$

which is the pdf of an exponentially tilted BFRY distribution, from which one can sample exactly via inversion, see Section D. Conditionally on w_i, t_i has a right-truncated exponential distribution, and one can also sample exactly using inversion.

Appendix G: Details on simulations and additional results

We are interested in measuring the error $R_n|T_{n+1}$, apparently not tractable. Hence, we consider the approximate error $R_{n,\hat{n}}$ defined for $\hat{n} > n$ as

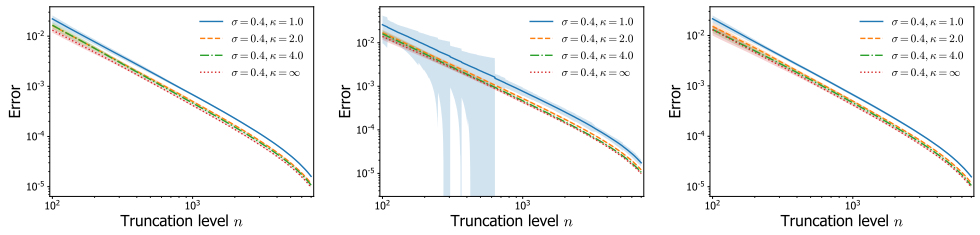
$$R_{n,\hat{n}} = \sum_{i=n}^{\hat{n}} W_i, \tag{G.1}$$

where we simulate $(W_i)_{i=1}^{\hat{n}}$ via sequential constructions. When no analytic expression is available (computing $\Psi^{-1}(t)$ for Gamma arrival times for GGP when $\kappa > 1$ or computing $\bar{\rho}^{-1}(\xi)$ for the inverse-Lévy for GGP), we resort to a numerical inversion algorithm. For each configuration, we first sample a series of arrival time sequences, and conditioning on that series, we sample 100 series of jumps $(W_i)_{i=1}^{\hat{n}}$ to compute $R_{n,\hat{n}}$. We repeat this procedure 10 times for each arrival time sequence (thus 1,000 jump simulations in total for each configuration) and report the mean and standard deviations. Unless specified otherwise, we use the hyperameters $\alpha = 2.0, \tau = 1.0$ for Stable Process (SP) and GGP, and set $\hat{n} = 10^4$. Figure 2 (Middle-Right) in the paper reports $R_{n,\hat{n}}$ for the SP and the GGP.

Figure 1 here shows the approximate errors for gamma and inverse gamma arrival times for the SP, and gamma arrival times for the GGP. We can observe that the variances quickly approach zero, except for the inverse gamma arrival time with $\kappa = 1$ case for which our theory predicts to have infinite variance.

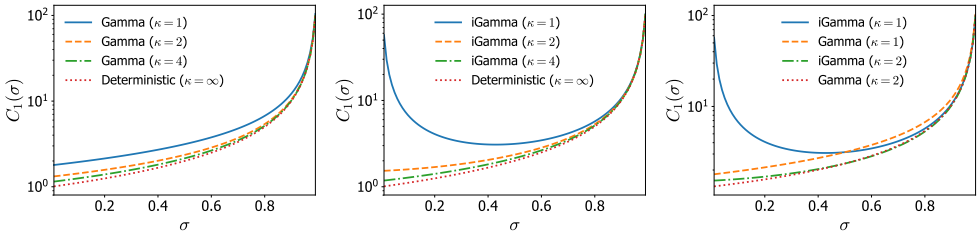
Figure 2 shows the value of constants $C_1(\sigma)$ for gamma and inverse gamma arrival times with varying σ and κ values. Note that lower $C_1(\sigma)$ implies lower expected error $\mathbb{E}[R_n|T_{n+1}]$ by Proposition 7.2. We found that gamma arrival times exhibit lower $C_1(\sigma)$ when $0 < \sigma < 0.5$, and inverse gamma has lower $C_1(\sigma)$ when $0.5 < \sigma < 1$. This observation is empirically confirmed in Figure 3.

Finally, we compared the approximate error to asymptotic value of R_n. In case of



(a) SP-Gamma (b) SP-iGamma (c) GGP-Gamma

Figure 1. Gamma arrival times, inverse gamma arrival times for stable processes (a-b) and gamma arrival times for GGP (c). Plotted for whole range. See how variances diminishes. Note also that in case of inverse gamma arrival times with $\kappa = 1$, the variances diverge as our theory predicts.



(a) Gamma (b) iGamma (c) Gamma vs. iGamma

Figure 2. $C_1(\sigma)$ values for Gamma (a), inverse gamma (b) arrival times, and comparison between them (c).

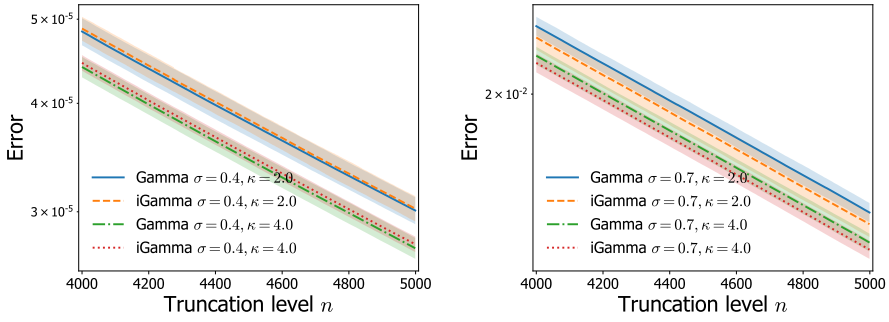
gamma arrival times for GGP, according to Proposition 7.2, we have

$$R_n \overset{n\to\infty}{\sim} \frac{(\kappa-\sigma)\Gamma(\kappa-\sigma)^{\frac{1}{\sigma}}}{(1-\sigma)\Gamma(\kappa)^{\frac{1}{\sigma}}}\left(\frac{\alpha}{\Gamma(1-\sigma)}\right)\sigma^{1-\frac{1}{\sigma}}n^{1-\frac{1}{\sigma}}. \tag{G.2}$$

Figure 4 compares the empirical approximate errors with $\hat{n} = 10^6$ to the asymptotic error as shown in Equation (G.2) for two different values of σ. We fix $\kappa = 1$. One can see that the approximate error quickly approaches the asymptotic error in both cases.

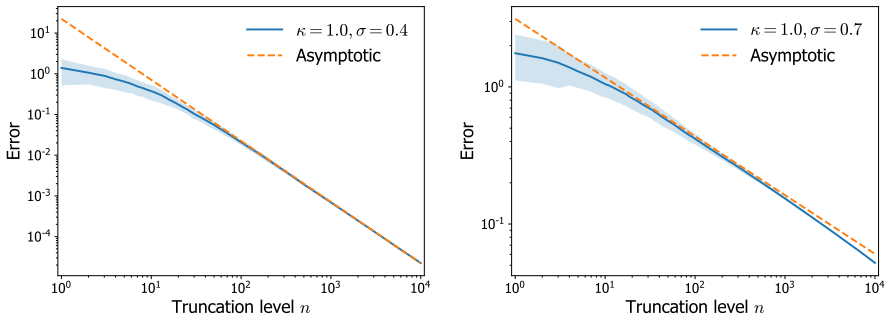
References

F. Ayed, J. Lee, and F. Caron. Beyond the Chinese restaurant and Pitman-Yor processes: statistical models with double power-law behavior. In *International Conference on Machine Learning (ICML2019)*, Long Beach, USA, 2019.



(a) $\sigma = 0.4$ (b) $\sigma = 0.7$

Figure 3. Gamma arrival times vs. inverse gamma arrival times for stable process, with $\sigma = 0.4$ (a) and $\sigma = 0.7$ (b). Gamma is better when $\sigma < 0.5$, and inverse gamma is better when $\sigma > 0.5$ as predicted in Fig. 2 (c).



(a) $\sigma = 0.4$ (b) $\sigma = 0.7$

Figure 4. Empirical approximation error $R_{n,\hat{n}}$ compared to asymptotic error R_n, for (a) $\sigma = 0.4$ and (b) $\sigma = 0.7$.

J. Bertoin, T. Fujita, B. Roynette, and M. Yor. On a particular class of self-decomposable random variables : the durations of Bessel excursions straddling independent exponential times. *Probability and Mathematical Statistics*, 26(2):315–366, 2006.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27. Cambridge university press, 1989.

L. Devroye. Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comput. Simul.*, 19(4):18–1, 2009.

L. Devroye and L. F. James. On simulation and properties of the stable law. *Statistical Methods and Applications*, 23(3):307–343, 2014.

A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys*, 4:146–171, 2007.

J. E. Griffin and S. G. Walker. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011.

L. F. James, P. Orbanz, and Y. W. Teh. Scaled subordinators and generalizations of the Indian buffet process. *arXiv:1510.07309*, 2015.

J. Lee, L. F. James, and S. Choi. Finite-dimensional BFRY priors and variational Bayesian inference for power-law models. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

J. Pitman. Poisson-kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

J. Rosiński. Contribution to the discussion of the paper "Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics" by O. Barndorff-Nielsen and N. Shephard. *Journal of the Royal Statistical Society, Series B*, 63:230–231, 2001.

J. Rosiński. Tempering stable processes. *Stochastic processes and their applications*, 117 (6):677–707, 2007.

M. Winkel. Electronic foreign-exchange markets and passage events of independent subordinators. *Journal of Applied Probability*, 42(1):138–152, 2005.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	A unified construction for series representations and finite approximations of completely random measures
Publication Status	□Published □ Accepted for Publication ☑Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Submitted in the Bernoulli Journal and it is under review.

Student Confirmation

Student Name:	Xenia Miscouridou
Contribution to the Paper	I have contributed in: (a) the conception of the idea, formulation of the motivation, objectives of this work (b) the formulation of the framework and the construction (c) deriving specific examples (d) understanding and working on the framework's properties and quantifying the error

Signature  Date 05/01/2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Prof. François CARON
Supervisor comments	/

Signature  Date 08/01/2020

Chapter 5

Discussion

In conclusion, we first present a summary of the contributions presented in the thesis. Subsequently, we discuss additional extensions and ideas for future work and outline relevant papers of other authors that build on our contributions.

5.1 Summary

This thesis presents contributions in the area of BNP Statistics. The first contribution consists of a network class model for sparse modular graphs constructed on vectors of exchangeable random measures. The driving force for this work is the tremendous interest over recent years in the study of networks. We build upon the framework of Caron and Fox [2017] which is the first type of model to allow for a whole range of dense to sparse graphs while keeping probabilistic symmetry through a different notion of exchangeability. This is the first approach to our knowledge allowing for these two network properties together and in this way defines a new research direction for the BNP community.

The first main part of the thesis presented in Chapter 2 consists of the paper Todeschini et al. [2019] which generalises the construction of Caron and Fox [2017] from a univariate construction (scalar CRMs) to a multivariate construction (vectors of CRMs). To the best of our knowledge, this is the first paper that generalises probabilistic models of overlapping communities to the sparse regime.

In Chapter 3 we present the second major contribution, which is a model for temporal interaction data that exhibit a series of properties observed in real interactions. Time-stamped interactions are represented in the form of triples (i, j, t) indicating that an event was initiated from a sender i towards a receiver j at time t. Our model

explains the underlying structure and properties that underpin these interactions, namely, community structure, reciprocity, sparsity and degree heterogeneity. An important aspect of the model is that it captures events of the type action-response in which an action from a member i to a member j encourages the same action being returned in the near future. The elegance of the proposed model is that it unifies and completes existing literature by providing a framework to generalise existing models of reciprocating relationships to the sparse and power law regime.

The third paper of the thesis presented in Chapter 4 is a methodological contribution to the area of BNP. We provide a unified way to design principled approximations of infinite-activity CRMs which are used in a variety of applications. This work proposes finite-dimensional approximations to infinite-activity CRMs by generalising the size-biased approach of Perman et al. [1992]. The practical impact of this work is the ability to develop efficient posterior inference algorithms that tackle the computationally cumbersome inference associated with infinite-activity CRMs.

5.2 Extensions

Our work opens many interesting directions of further research. Below, we discuss some ideas of future research as well as current work that builds on our contributions.

5.2.1 Sparse Networks with Core-Periphery Structure

A recent piece of work that expands on our modular graph constructions [Todeschini et al., 2019] is by Naik et al. [2019] who propose a model for sparse graphs with core-periphery structure. To motivate this idea, recall the network of flight connections, USairport. Although a community structure does exist, this dataset is a typical example of a core-periphery network as explained in Naik et al. [2019]. The core nodes correspond to central hubs that flights are routed through, while other airports are more sparsely connected between them. The precise notion of core-periphery structure is defined in Naik et al. [2019] in the scope of sparsity properties of the subgraphs that correspond to the core and the periphery nodes respectively. The main idea is that graphs consist of two classes of nodes; the core nodes that are densely connected to each other and the periphery nodes that have sparse connections between them. Therefore, the core-periphery model can be seen as an instance of our modular p community graphs to the case of $p = 2$.

Consider a Poisson point process on $[0,+\infty)^2\backslash\{0,0\} \times [0,\infty)$ with mean measure $\rho(dw_1, dw_2)d\theta$ and points denoted by $(w_{1i}, w_{2i}, \theta_i)_{i\geq 1}$. As in Chapter 2, the parameter α which indexes the graph sequences, is a size parameter and the conditions specified by Naik et al. [2019] imply that, for any $\alpha > 0$

$$card\{i|w_{1i} > 0, \theta_i \leq \alpha\} < \infty$$
$$card\{i|w_{1i} = 0 \text{ and } w_{2i} > 0, \theta_i \leq \alpha\} = \infty.$$

There exist two sets of nodes, with the first set of nodes that satisfies $w_{1i} > 0$ corresponding to the core and the other set of nodes corresponding to the periphery. Under regular variation assumptions the number of edges between nodes in the core grows quadratically with the number of nodes in the core, but the number of edges between nodes of the periphery (and between nodes overall) grows subquadratically with the number of nodes, giving the desired sparsity patterns.

5.2.2 Topological Covariates in the Link Probability

In the original graph construction of Caron and Fox [2017] as well as in our models [Todeschini et al., 2019; Miscouridou et al., 2018a] the locations of the nodes denoted by θ_i do not take part in the link probability. The introduction of topology in the graph family is another direction that could be explored. For instance we could modify the link probability of connections to depend not only on the sociability vectors w_i but also on the locations θ_i. In this way we can allow the distance between two nodes to affect how likely they are to connect. In the case of the USairport dataset of flight connections the locations of the airports could be included in the link probability. Furthermore, it may be of interest to consider other types of node covariates in the link probability and include them in various functional forms.

5.2.3 Sociability Parameters' Evolution in Time

One way to consider dynamic interactions in networks is our extension of the static model in Todeschini et al. [2019] to dynamic reciprocating relationships [Miscouridou et al., 2018a]. The latter assumes that one of the factors underpinning the interactions between members is that members have similar community affiliations. However, the model assumes that the latent network communities are constant in time. In order to have communities changing in time or new communities appearing later on, the sociability parameters must become functions of time. This idea has been addressed

by Palla et al. [2016] for graph models with no communities. In particular, they consider that each node has a set of sociability parameters indexed by discrete times

$$(w_{ti})_{t=1,\dots,T,i=1,2,\dots} \tag{5.1}$$

which then gives rise to a time dependent graph. What would be exciting but highly non-trivial is the generalisation of their model to the overlapping community case. In such a case the sociability vectors of the community affiliations of the nodes evolve in time, capturing how their memberships and interests change. In this way we can explore which groups survive in time, which groups gradually vanish and if new ones are formed. Then, it may be of interest to match the results with the ground truth if available. For example, when studying social networks we can compare our inference results to ground truth information on the socio-economic reasons that may have led to the growth or shrinkage of specific groups. The evolution of communities and interests also relates to peer influence and habit formation in social networks. This can be achieved by monitoring the evolution of the interests of the individuals in order to study and eventually predict how the interests propagate and spread in the network.

5.2.4 Statistical Consistency

Of great interest is the study on statistical consistency of the inference methods for this class of graph models. The question of what type of frequentist properties we can study on our Bayesian estimators is very interesting. For instance we would like to know whether we use the credible intervals obtained from our posterior distributions for the parameters and hyperparameters as confidence intervals. Caron and Rousseau [2017] present the asymptotic expressions for the number of edges, number of nodes and degree distributions and use their results to derive a consistent estimator for the sparsity parameter σ. We encourage similar results to be derived for our work [Todeschini et al., 2019].

Consistency results are also relevant for our work in Miscouridou et al. [2018a]. We use a two-step inference procedure motivated by modular Bayesian inference in which we first estimate the structural (static) parameters of the global graph and then use a point estimate of those to estimate the temporal parameters of reciprocating interactions. To validate the two-step approximation to the posterior distribution, we

provide empirical results on the convergence of our approximate inference scheme using simulated data. Our experiments suggest that the posterior distribution concentrates closer to the sampled parameter value with increasing data size. A theoretical investigation of this would be a great contribution to our work.

5.2.5 Assortativity and Disassortativity in Link Probability

In social networks we often observe the property of assortativity (or homophily) which means that the existence of an edge (i, j) is highly likely because the nodes i and j are similar between them. This assumption is taken in both our models [Todeschini et al., 2019; Miscouridou et al., 2018a]. We assume that two nodes with similar characteristics are more likely to interact rather than two nodes with dissimilar characteristics. This assumption could be easily generalised to allow for a whole range of assortative to disassortative mixing. A choice is the link function proposed in Zhou [2015]

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

consisting of a probability weighted by the parameters $\eta_{k\ell} \geq 0$ which tune the assortative and disassortative mixing in the network. For instance, capturing disassortative mixing is achieved by setting larger values off-diagonal than on the diagonal of the matrix $(\eta_{k\ell})_{1\leq k,\ell\leq p}$. Note that our graph properties and algorithms for simulation and posterior inference [Todeschini et al., 2019] can be trivially extended to this case.

5.2.6 Number of Communities

Our model is nonparametric in the number of nodes and edges but is parametric in the number of communities meaning that the number of communities p is fixed. Being nonparametric in the number of communities, nodes and edges is particularly complicated and highly non-trivial. An example of a sparse model in which the number of nodes is parametric but the number of communities is nonparametric, is the one of Ayed and Caron [2019] which assumes a random number of communities but a fixed number of nodes. It is noteworthy that several community detection methods that assume a fixed number of communities have been particularly successful in applications.

Placing fully nonparametric assumptions on the number of communities is not the objective of the work presented in the thesis but it remains an interesting extension of our work and a proper investigation on potential approaches is highly encouraged.

An approach has been suggested by Zhou [2015] considering that i and j connect with probability

$$\Pr(z_{ij} = 1) = 1 - \exp\left(-w_{0i}w_{0j}\sum_{k=1}^{\infty} r_k\beta_{ik}\beta_{jk}\right),$$

in which the sequence of $(r_k)_k$ can be for instance the jumps of a gamma process. MCMC inference would be particularly problematic here, as the model is doubly nonparametric in both w_{0i} and r_k. In fact, even in Zhou [2015], they do not perform inference on their (dense) model. Alternatively, they consider a truncated model at K_{max}, and perform inference on this parametric model instead, treating it as an approximation to the nonparametric model. In a similar way, we could also consider a truncated model (e.g. $K_{max} = 100$ or 200) but such an approach inherits the limitation of having to fit a parametric model with a fixed, potentially large upper bound on the number of features. Additionally, in our inference algorithms we interpret the communities based on the minimum Bayes risk point estimate in which the cost function is a permutation-invariant absolute loss on the weights. This approach would be infeasible with a large (or even moderately large) number of communities and may force us to other estimation procedures like a maximum a posteriori (MAP) estimate as done in Zhou [2015].

A non-Bayesian nonparametric approach on the number of communities is cross-validation. One could subsample both training and test graphs from the observed graph data and study the estimated prediction on summary statistics. The subsampling procedure associated to the class of graphs considered in this paper is called p-sampling and has been introduced by Veitch and Roy [2016]. This direction is statistically sound and seems promising, however it is not Bayesian like the rest of our work.

5.2.7 Generic Reciprocation Patterns

In Chapter 3 we have used Hawkes processes to model temporal interaction data, and focus on the case of an exponential kernel. This implies that the probability of an immediate response to a recent action is high and then decays exponentially. In theory, the exponential kernel can be replaced by any other type of kernel or by a general functional form. However, this is likely to have computationally complicated inference. The exponential kernel has been a popular choice in literature and cannot be considered as very restrictive. In Miscouridou et al. [2018a] we assume that

the behaviour of reciprocation is globally uniform in the datasets which means that everybody has a similar response behaviour and action effect. This assumption was motivated by what was observed empirically in the datasets we were studying but it can be trivially extended. For instance, we can introduce personalisation in the reciprocation pattern with the extreme case being individual response behaviours. Our current assumption lies on the other extreme with every member of the network following the same behaviour. The intermediate cases are of interest too, such as having groups of reciprocation patterns with the groups depending on the type of affiliations, the event type, the event duration or other node covariates.

5.2.8 Other Applications

Using our proposed model classes we have performed experiments primarily in social datasets. However, we can use our models [Todeschini et al., 2019; Miscouridou et al., 2018a] and apply our novel techniques [Lee et al., 2019] in many different contexts. For instance we could apply our models to uncover the latent structure in biological networks. A relevant example may be protein-protein interaction networks or the human microbiome.

Regarding the property of reciprocation we have studied datasets of emails, messages and online blogs but as discussed in Blundell et al. [2012] reciprocation is observed in many scenarios. An area in which mutual excitation patterns may exist is neuroscience with neurones exciting each other through their activity in the brain. A research study in this direction may be fruitful and of interdisciplinary impact.

Bibliography

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

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

D. J. Aldous. Exchangeability and related topics. In *Ecole d'été de Probabilités de Saint-Flour XIII - 1983*, pages 1–198. Springer, 1985.

F. Ayed and F. Caron. Nonnegative Bayesian nonparametric factor models with completely random measures for community detection. *arXiv:1902.10693*, 2019.

A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

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.

N. Berger, C. Borgs, J. T. Chayes, and A. Saberi. Asymptotic behavior and distributional limits of preferential attachment graphs. *Annals of Probability*, 42:140, 2014.

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

C. Blundell, K. Heller, and J. Beck. Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems 25*, volume 15, pages 5249–5262, 2012.

B. Bollobás. *Random graphs*, volume 73. Cambridge University Press, 2001.

C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research*, 18:1–71, 2018.

D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity, 2016.

F. Caron and E. B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5), 2017.

F. Caron and J. Rousseau. On sparsity and power-law properties of graphs based on exchangeable point processes. *https://arxiv.org/pdf/1708.03120.pdf*, 2017.

F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

H. Crane and W. Dempsey. A framework for statistical network modeling. *arXiv preprint arXiv:1509.08185*, 2015.

H. Crane and W. Dempsey. Edge exchangeable models for interaction networks. *arXiv preprint arXiv:1603.04571*, 2016.

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.

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

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 and S. G. Walker. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011.

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.

A. G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971a.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(7):5249–5262, 1971b.

T. Herlau, M. N. Schmidt, and M. Mørup. Completely random measures for modelling block-structured sparse networks. In *Advances in Neural Information Processing Systems*, 2016.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

D. N. Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 1979.

S. Janson. On convergence for graphexes. *arXiv preprint arXiv:1702.06389*, 2017a.

S. Janson. On edge exchangeable random graphs. *To appear in Journal of Statistical Physics.*, 2017b.

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.

M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160, 2012.

J.F.C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21 (1):59–78, 1967.

J. Lee, X. Miscouridou, and F. Caron. A unified construction for series representations and finite approximations of completely random measures. *arXiv preprint arXiv:1905.10733*, 2019.

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.

J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

A. Lijoi, B. Nipoti, and I. Prünster. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291, 2014.

J. Lloyd, P. Orbanz, Z. Ghahramani, and D. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems*, volume 25, pages 1007–1015, 2012.

K. Miller, T. Griffiths, and M. Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, 2009.

X. Miscouridou, F. Caron, and Y. W. Teh. Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. In *Neural Information Processing Systems*, 2018a.

X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Proceedings of the Third Conference on Machine Learning for Healthcare*, volume 85 of *Proceedings of Machine Learning Research*, pages 244–256, 2018b.

C. Naik, F. Caron, and J. Rousseau. Sparse networks with core-peripherie structure. *arXiv:1910.09679*, 2019.

M. Newman. *Networks: An Introduction*. OUP Oxford, 2010.

Y. C. Ng and R. Silva. A dynamic edge exchangeable model for sparse temporal networks. *arXiv preprint arXiv:1710.04008*, 2017.

K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

P. Orbanz. Nonparametric priors for exchangeable graphs and arrays. 2015b.

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, F. Caron, and Y.W. Teh. Bayesian nonparametrics for sparse dynamic networks. *arXiv preprint arXiv:1607.01624*, 2016.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

A. Todeschini, X. Miscouridou, and F. Caron. Exchangeable random measures for sparse and modular graphs with overlapping communities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019. (to appear).

U.S. Bureau of transportation statistics, 2010. URL `http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=292`.

V. Veitch and D. M. Roy. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.

V. Veitch and D. M. Roy. Sampling and estimation for (sparse) exchangeable graphs. *arXiv preprint arXiv:1611.00843*, 2016.

S. Williamson. Nonparametric network models for link prediction. *Journal of Machine Learning Research*, 17(202):1–21, 2016.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *ArXiv preprint arXiv:1309.5936*, 2013.

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 1135–1143, 2015.