Nonparametric Bayesian modeling and inference for Hawkes processes

Hyotae Kim and Athanasios Kottas *

December 14, 2024

Abstract

We propose a nonparametric Bayesian modeling and inference framework for Hawkes processes. The objective is to increase the inferential scope for this practically important class of point processes by exploring flexible models for its conditional intensity function. In particular, we develop different types of nonparametric prior models for the immigrant intensity and for the offspring density, the two functions that define the Hawkes process conditional intensity. The prior models are carefully constructed such that, along with the Hawkes process branching structure, they enable efficient handling of the complex likelihood normalizing terms in implementation of inference. We discuss prior specification for model hyperparameters, and design posterior simulation algorithms to obtain inference for different point process functionals. We also present computational methods for model assessment and for prediction, using the posterior samples for model parameters. The modeling approach is studied empirically using several synthetic data examples, and is illustrated through reexamination of a data set involving earthquake occurrences.

Keywords: Dirichlet process; Erlang mixtures; Gamma process; Markov chain Monte Carlo; Non-homogeneous Poisson process

*Hyotae Kim (hyotae.kim@duke.edu) is Postdoctoral Researcher, Department of Biostatistics & Bioinformatics, Duke University, and Athanasios Kottas (thanos@soe.ucsc.edu) is Professor, Department of Statistics, University of California, Santa Cruz. This work is part of the Ph.D. dissertation of H. Kim, completed at University of California, Santa Cruz. The research was supported in part by the National Science Foundation under award SES 1950902.

1 Introduction

We develop nonparametric Bayesian (NPB) modeling and inference methods for temporal Hawkes processes. The Hawkes process (HP), originally developed in Hawkes (1971), is a versatile stochastic model for point processes, built from structured conditional intensity functions that model *self-excitation*, i.e., the property that the occurrence of an event increases the rate of occurrence for some period of time in the future. Such a structure yields point patterns with events that are naturally clustered in time. For example, earthquake occurrences are grouped into clusters consisting of a main shock and subsequent shocks (aftershocks). Indeed, including extensions to incorporate marks and information on spatial location, HPs have found applications in seismology (e.g., Ogata 1988; Ogata 1998; Zhuang et al. 2002; Veen and Schoenberg 2008), as well as in crime data modeling (e.g., Mohler et al. 2011), finance (e.g., Da Fonseca and Zaatour 2014; Hardiman et al. 2013; Rambaldi et al. 2015), and biology (e.g., Balderama et al. 2012).

The standard form of the HP conditional intensity is expressed as

$$\lambda^*(t) \equiv \lambda(t \mid \mathcal{H}(t)) = \mu + \sum_{t_j<t} h(t-t_j), \quad t \in \mathbb{R}^+ \tag{1}$$

where $\mathcal{H}(t)$ denotes the point process history up to time t, $\mu > 0$ is the background (immigrant) intensity, and $h : \mathbb{R}^+ \to \mathbb{R}^+$ is the excitation function (offspring intensity), which accounts for the effect of previous events on the current intensity. The excitation function must be integrable over $\mathbb{R}^+$, and it can thus be equivalently represented in terms of the branching ratio, $\gamma = \int_0^\infty h(u)\mathrm{d}u$, and the offspring density $f(t) = h(t)/\gamma$. The original definition of the HP, and several of its applications, focus on constant immigrant rate μ, in which case the HP is stationary provided $\gamma \in (0, 1)$. Our methodology handles also the more general case of a non-constant immigrant intensity function, $\mu(t) : \mathbb{R}^+ \to \mathbb{R}^+$.

The *immigrant* and *offspring* terminology originates from the HP cluster representation (Hawkes and Oakes (1974)), which is key for our modeling and inference methods. Consider a HP realization, $\{0 < t_1 < ... < t_n < T\}$, observed in time window $(0, T)$. The branching

structure for the point pattern can be described by latent variables $\boldsymbol{y} = \{y_i : i = 1, ..., n\}$, such that $y_i = 0$ if t_i is an immigrant point, and $y_i = j$ if point t_i is the offspring of t_j. Hence, given $\boldsymbol{y}$, the HP point pattern is partitioned into the set of immigrants, $I = \{t_j : y_j = 0\}$, and sets of offspring, $O_j = \{t_i : y_i = j\}$, where O_j collects all offspring of t_j. Moreover, conditioning on the branching structure $\boldsymbol{y}$, the HP can be constructed as the superposition of independent Poisson processes corresponding to I and the O_j, with intensity μ (or $\mu(t)$) and $h(t - t_j) = \gamma f(t - t_j)$ for I and O_j, respectively.

The HP likelihood based on observed point pattern $\{0 < t_1 < ... < t_n < T\}$ is given by $\exp\left(-\int_0^T \lambda^*(u)\mathrm{d}u\right) \prod_{i=1}^n \lambda^*(t_i)$. Owing to the additive form of the conditional intensity, it is challenging to work with the likelihood, even under simple parametric offspring intensities. The HP cluster representation provides a practically useful alternative, using the branching structure latent variables. Consider the general case with a time-varying immigrant intensity. Then, given $\boldsymbol{y}$, the augmented likelihood can be expressed as

$$\exp\left(-\int_0^T \mu(u)\mathrm{d}u\right) \left\{\prod_{\{i:t_i \in I\}} \mu(t_i)\right\} \exp\left(-\sum_{j=1}^n \int_0^T h(u - t_j)\mathrm{d}u\right) \left\{\prod_{\{i:t_i \in O\}} h(t_i - t_{y_i})\right\} \tag{2}$$

where $O = \cup_{j=1}^n O_j$ is the set of all offspring points. The second exponential term incorporates the probability that point t_j has no offspring in $(0, T)$, for all j with $O_j = \emptyset$. The distribution for the latent variables, which yields the HP likelihood upon marginalization over $\boldsymbol{y}$, is given by $\delta_0(y_1) \prod_{i=2}^n \mathrm{Unif}(y_i \mid 0, 1, \ldots, i-1)$, where $\mathrm{Unif}(\cdot \mid 0, 1, \ldots, i-1)$ is the discrete uniform distribution on $\{0, 1, \ldots, i-1\}$, and $\delta_s(\cdot)$ denotes the point mass at s.

Regarding classical inference for HPs, which is based mainly on maximum likelihood estimation, we refer to the reviews by Laub et al. (2021) and Reinhart (2018), the latter focusing on space-time HPs. Bayesian modeling and inference for HPs has received relatively less attention in the literature. Working with constant immigrant intensity and parametric forms for the excitation function, Rasmussen (2013) compared posterior simulation methods based on either the likelihood defined directly through the conditional intensity or the

augmented likelihood that utilizes the branching structure. The more recent literature includes also a small collection of NPB methods. Donnet et al. (2020) studied NPB modeling for multivariate HPs with an emphasis on theoretical properties, such as posterior concentration rates. Inference methods based on Gaussian process priors have been investigated in the machine learning literature (e.g., Zhang et al., 2019; Zhou et al., 2021, 2020).

In this article, we seek to develop a general NPB modeling and inference framework for temporal Hawkes processes. Our key motivation is to expand the scope of modeling methods for inference and prediction under this practically important class of self-exciting point processes. We develop different NPB models, all based on mixture structures, for the components of the HP conditional intensity, including a prior model for each of the immigrant intensity, excitation function, and offspring density. Posterior simulation is facilitated by the HP branching structure cluster representation, and by the model structure of the proposed mixture formulations. Indeed, the models are carefully constructed to achieve a balance between computational efficiency in model estimation and a range of general shapes for the corresponding functions. As part of the modeling framework, we study a nonparametric prior that supports general decreasing offspring densities. This option may prove practically useful for HP applications to earthquake modeling. It incorporates the meaningful constraint of a decreasing offspring density (present in standard parametric models) while allowing for general tail behavior, including either exponential or polynomial offspring density tails. Another component of the proposed framework involves fully NPB models that combine the prior for the immigrant intensity with either of the priors for the excitation function or offspring density. As demonstrated with several synthetic data examples, the fully NPB models can uncover non-standard shapes for the different components of the HP conditional intensity. In addition to MCMC posterior simulation algorithms, we present various inference methods, including inference for the HP first-order and second-order intensities, model assessment, and prediction.

The rest of the paper is organized as follows. In Section 2, we develop the different NPB models for the HP conditional intensity function. Technical details on prior specification and posterior simulation are deferred to the Supplementary Material. Section 3 presents methods for inference beyond estimation of the models, including inference for first-order and second-order intensities, model assessment techniques, and predictive inference. In Section 6, we summarize the findings from a detailed simulation study, with the full results given in the Supplementary Material. An application involving earthquake data analysis follows in Section 5. Finally, Section 6 concludes with a summary.

2 Modeling methods for the HP intensity function

This section develops the prior models for the two components of the HP conditional intensity, the immigrant intensity function (Section 2.1) and the excitation function (Section 2.2), with fully NPB models emerging from their combination (Section 2.3).

The model for the immigrant intensity and one of the models for the excitation function build from weighted combinations of Erlang densities. Relevant to those modeling approaches is the Erlang mixture model for density estimation, which we review here. Denote by $\text{Ga}(\cdot \mid a, b)$ the gamma density or distribution (depending on the context) with mean a/b, and consider the following representation for density f on $\mathbb{R}^+$:

$$f(t) = \sum_{l=1}^{L} \psi_l \, \text{Ga}(t \mid l, \theta^{-1}), \quad t \in \mathbb{R}^+ \tag{3}$$

where $\psi_l = F(l\theta) - F((l-1)\theta)$, for $l = 1, \ldots, L-1$, and $\psi_L = 1 - F((L-1)\theta)$, for a distribution function F on $\mathbb{R}^+$. This is a structured mixture for which the Erlang densities play the role of basis densities with increasing means and a single parameter θ controlling their dispersion. The relative magnitude of the weights assigned to the basis densities is driven by the shape of distribution F. Under this formulation for the weights, as $\theta \to 0$ and $L \to \infty$, the Erlang mixture density converges pointwise to the density of F (e.g.,

Butzer 1954; Lee and Lin 2010). The convergence result for the Erlang mixture structure offers the foundation for NPB modeling, treating F as a random distribution. There has been some recent relevant work on NPB modeling for renewal process inter-arrival densities (Xiao et al., 2021) and for response densities in survival analysis (Li et al., 2023), using in both cases a Dirichlet process (DP) prior (Ferguson, 1973) for F.

2.1 Modeling the immigrant intensity function

As discussed in the Introduction, temporal HPs are typically applied with constant immigrant rate μ. To achieve general inference and prediction, we seek a flexible NPB prior model for immigrant intensity functions $\mu(t)$. In particular, in earthquake modeling, the immigrant intensity explains the dynamics of main shock occurrences, and it is thus natural to expect improved prediction if we can effectively estimate time-dependent immigrant intensities. Indeed, this is illustrated with the earthquake data analysis of Section 5.

Given the branching structure, the immigrant intensity plays the role of the non-homogeneous Poisson process (NHPP) intensity for the immigrants point pattern. In terms of applications for temporal NHPPs, existing methods typically build from Gaussian process (GP) priors for (transformations of) the NHPP intensity (e.g., Adams et al., 2009), or DP mixture priors for the NHPP density over the observation window $(0, T)$ (e.g., Taddy and Kottas, 2012). As the prior for the immigrant intensity $\mu(t)$, we employ the NPB model developed in Kim and Kottas (2022), which extends the Erlang mixture model in (3) for intensity estimation. Relative to GP-based methods, this model is significantly more computationally efficient. Compared with DP mixture models, it offers a more parsimonious representation for the intensity over $\mathbb{R}^+$, rather than only over $(0, T)$, which is important for predictive inference based on the estimated HP conditional intensity.

With a generic excitation function h, the Erlang mixture model for the immigrant

intensity yields the following formulation for the HP conditional intensity:

$$\lambda^*(t) = \sum_{l=1}^{L} \nu_l \, \mathrm{Ga}(t \mid l, \theta^{-1}) + \sum_{t_j<t} h(t - t_j), \quad t \in \mathbb{R}^+$$
$$\nu_l = G(l\theta) - G((l-1)\theta), \;\; l = 1, \ldots, L, \quad G \mid c_0, G_0 \sim \mathcal{G}(G_0, c_0). \tag{4}$$

Here, $\mathcal{G}(G_0, c_0)$ denotes the gamma process with precision parameter $c_0 > 0$, and centering cumulative intensity function G_0, such that $\mathrm{E}(G(t)) = G_0(t)$ and $\mathrm{Var}(G(t)) = G_0(t)/c_0$ (Kalbfleisch, 1978). We set $G_0(t) = t/b_{G_0}$, i.e., the cumulative hazard of the exponential distribution with scale parameter b_{G_0}. Hyperpriors are placed on c_0, b_{G_0}, and θ. Owing to the gamma process independent increments, the weights ν_l are independently gamma distributed. In particular, $\nu_l \mid \theta, c_0, b_{G_0} \overset{ind.}{\sim} \mathrm{Ga}(c_0\, \theta\, b_{G_0}^{-1}, c_0)$, under the exponential G_0.

The model for the immigrant intensity builds from a weighted combination of Erlang densities with common scale parameter θ and (specified) integer shape parameters. The weights are constructed through increments of random cumulative intensity function G. In the context of NHPPs, Kim and Kottas (2022) present a convergence result that extends the one for Erlang mixture densities: as $\theta \to 0$ and $L \to \infty$, $\mu(t) = \sum_{l=1}^{L} \nu_l \, \mathrm{Ga}(t \mid l, \theta^{-1})$ converges pointwise to the NHPP intensity that corresponds to cumulative intensity G. This result provides theoretical support for the model structure in (4) and motivates the nonparametric prior for G. The gamma process prior allows flexible cumulative intensity realizations G with their effective support concentrated on different time intervals, resulting in weights ν_l that concentrate on different subsets of the Erlang basis densities. The precision parameter c_0 controls the variability of realizations G around G_0, and thus the effective mixture weights. In general, a smaller value for c_0 implies a smaller number of practically non-zero weights. Moreover, smaller values of θ and larger values of L yield more variable, typically multimodal shapes for the intensity.

Note that we are using a parsimonious mixture formulation with the Erlang densities defined in terms of a single parameter, and playing the role of basis functions in the model representation for the immigrant intensity. Working with a specified, sufficiently large

number of basis densities, the nonparametric prior for G is key in the selection of the subset of them to be used in the representation of the intensity. An approach to specify L and the priors for c_0, b_{G_0}, and θ is discussed in the Supplementary Material.

Let $\{0 < t_1 < ... < t_n < T\}$ be the point pattern observed in time window $(0, T)$. The model is implemented using the HP cluster representation, based on the branching structure variables $\boldsymbol{y}$, along with an additional set of latent variables $\boldsymbol{\xi} = \{\xi_i : i \in I\}$, where ξ_i indicates the Erlang basis density to which immigrant time point $t_i \in I$ is assigned. Building from the HP augmented likelihood in (2), the first stage of the hierarchical model for the observed point pattern, given $\boldsymbol{y}$ and $\boldsymbol{\xi}$, can be written as

$$\begin{aligned} \exp\left(-\sum_{l=1}^{L}\nu_l\int_0^T \mathrm{Ga}(u \mid l,\theta^{-1})\mathrm{d}u\right)\left\{\prod_{\{i:t_i\in I\}}\left(\textstyle\sum_{k=1}^{L}\nu_k\right)\mathrm{Ga}(t_i \mid \xi_i,\theta^{-1})\right\} \\ \times \exp\left(-\sum_{j=1}^{n}\int_0^T h(u-t_j)\mathrm{d}u\right)\left\{\prod_{\{i:t_i\in O\}} h(t_i-t_{y_i})\right\} \end{aligned} \tag{5}$$

where the ξ_i are i.i.d., given $\boldsymbol{\nu} = \{\nu_l : l = 1, ..., L\}$, with ξ_i taking value l with probability $\nu_l/(\sum_{k=1}^{L}\nu_k)$, for $l = 1, ..., L$. Note that, marginalizing over the ξ_i, the second term in (5) becomes $\prod_{\{i:t_i\in I\}}\left\{\sum_{l=1}^{L}\nu_l\,\mathrm{Ga}(t_i \mid l,\theta^{-1})\right\} = \prod_{\{i:t_i\in I\}}\mu(t_i)$.

The Supplementary Material includes MCMC posterior simulation details for the immigrant intensity model parameters, $(\boldsymbol{\nu}, \theta, c_0, b_{G_0})$, and for latent variables $\boldsymbol{y}$ and $\boldsymbol{\xi}$. A key feature of the Erlang mixture model structure is that the weights ν_l can be sampled independently from gamma posterior full conditional distributions.

2.2 Mixture models for the excitation function

In this section, we turn our attention to modeling the excitation function. Regarding estimation, this is a challenging task, even under the augmented likelihood in (2). The challenge arises from likelihood component $\exp\left(-\sum_{j=1}^{n}\int_0^T h(u-t_j)\mathrm{d}u\right)$, which will generally include in a complicated fashion all model parameters for the excitation function.

We develop two mixture models inspired by the direct connection between the excitation function h and the offspring density f, i.e., $f(t) = h(t)/\gamma$, where $\gamma = \int_0^\infty h(u)\mathrm{d}u$ is the branching ratio. The first model (Section 2.2.1) uses a weighted combination of Erlang densities to represent the excitation function. Because the excitation function must be integrable over $\mathbb{R}^+$, the model formulation is different from the one of Section 2.1 for the immigrant intensity, and it implies the Erlang mixture model in (3) for the offspring density. Therefore, the model allows for general excitation function shapes. Moreover, it results in efficient posterior simulation, since, similarly to the model of Section 2.1, the mixture weights can be updated through gamma distributions.

Under the second modeling approach (Section 2.2.2), we work directly with the offspring density and employ a nonparametric uniform mixture structure that represents all decreasing densities on $\mathbb{R}^+$. The constraint of a decreasing offspring density is motivated by the two parametric forms that are most commonly used, i.e., the exponential or Lomax (power-law) offspring density. The former is typically chosen for convenience as it provides a simple parametric form for estimation, as well as analytical results for HP functionals dating back to the study by Hawkes (1971). The Lomax density is, to our knowledge, the only parametric form that can be potentially justified in the context of a specific scientific area, in particular, by *Omori's law* for the decay over time of the frequency of earthquake aftershocks (e.g., Ogata, 1988). From a modeling perspective, the assumption of a decreasing offspring density may be viewed as natural in other applications, given the role of the excitation function in the generation of HP point patterns. In this respect, the approach developed in Section 2.2.2 offers modeling flexibility with respect to the shape and tail behavior of the offspring density; note that commonly used parametric models support either exponential or polynomial tails.

2.2.1 Erlang mixture for the excitation function

The HP cluster representation considers the excitation function as the intensity of offspring Poisson processes, as in the immigrant intensity model for the immigrant intensity function. But, unlike the immigrant intensity, the excitation function must be integrable (over time) so that it can be factorized into the offspring density and the branching ratio. In the cluster representation, the branching ratio, γ, determines cluster sizes in that offspring points of a cluster are produced in successive generations following a Galton–Watson branching process with Poisson offspring distribution whose mean is γ. Provided $\gamma \in (0, 1)$, each cluster size is almost surely finite, with its expected size of $1 + \gamma + \gamma^2 + \ldots = 1/(1 - \gamma)$ (Hawkes and Oakes, 1974; Daley et al., 2003). Therefore, the *stability condition*, $\gamma \in (0, 1)$ is needed to prevent the process explosion caused by divergent cluster sizes. Model development should also take this into account, along with the essential requirement of excitation integrability.

As an approach to the excitation function, one might consider the Erlang mixture model of Section 2.1. But, the model describes the branching ratio as $\gamma = \sum_{l=1}^{L} \nu_l = G(L\theta)$; the cumulative hazard function, G, is infinite, as $L\theta \to \infty$, i.e., the model does not guarantee the integrability of the excitation function. Alternatively, we have developed a mixture of Erlang densities with weights (directly) assigned independent gamma priors rather than defined by increments of a cumulative hazard function as in the Erlang mixture model. The model fulfills the integrability modeling condition, while retaining all the benefits of the Erlang mixture, such as conjugate priors for the mixture weights and the effective handling of the likelihood normalizing constant term. Further, the proposed method achieves the HP stability condition by simply adjusting hyperparameters of the gamma priors, as detailed below. Assuming constant immigrant intensity μ, the proposed model is defined as

$$\begin{aligned} \lambda^*(t) = \mu + \sum_{t_i<t} \Big[\sum_{l=1}^{L} \nu_l \, \mathrm{Ga}(t - t_i \mid l, \theta^{-1}) \Big], \quad t \in \mathbb{R}^+ \\ \nu_l \overset{ind.}{\sim} \mathrm{Ga}(A_l, \eta), \end{aligned} \tag{6}$$

where $A_l \equiv \alpha_0\{F_0(l\theta) - F_0((l-1)\theta)\}$, for $l = 1, \ldots, L-1$, and $A_L \equiv \alpha_0[1 - F_0((L-1)\theta)]$.

We define F_0 as a cumulative distribution function and assign an exponential cumulative distribution for simplicity. As the mixture weights have gamma priors with a common rate η, the branching ratio, $\gamma = \sum_{l=1}^{L} \nu_l$, is also gamma distributed, which ensures the integrability of the excitation function. Under the modeling framework, the shape parameter for γ is given by $\alpha_0 = \sum_{l=1}^{L} A_l$. So, one can even manage the HP stability condition by simply adjusting the two hyperparameters α_0 and η to place large probability mass on the unit interval.

The excitation function in (6) can be expressed as $\gamma\Big[\sum_{l=1}^{L} \psi_l \,\mathrm{Ga}(t - t_i \mid l, \theta^{-1})\Big]$, where $\psi_l = \nu_l / \sum_{k=1}^{L} \nu_k = \nu_l/\gamma$. Since a Dirichlet random vector can be represented by independent gamma random variables, we can regard the weights $(\psi_1, \psi_2, \ldots, \psi_L)$ as a Dirichlet random vector with concentration parameters A_l, $l = 1, \ldots, L$. Such an offspring density function forms an Erlang mixture for the target being a density function, with a DP prior on a cumulative distribution function F such that $\psi_l = F(l\theta) - F((l-1)\theta)$ for $l = 1, \ldots, L-1$ and $\psi_L = 1 - F((L-1)\theta)$.

To complete the probability model, we place the following set of priors on the model parameters. The constant immigrant μ is assigned the $\mathrm{Exp}(a_\mu)$ prior, under which the posterior full conditional for μ is available in closed-form. Hyperparameter a_μ is chosen by linking T/a_μ, interpreted as the expected cumulative immigrant intensity over $(0, T)$ under the constant assumption, to the size of observed immigrant points (or simply $n/2$, half of the observed points). Again, the branching ratio is gamma distributed with shape α_0 and rate η. The hyperparameters are specified considering the stability condition, such that $\Pr(0 < \gamma < 1) \approx 0.9$.

Under a constant immigrant intensity in the HP augmented likelihood in (2), the hierarchical representation of the semiparametric model with the Erlang mixture for the

excitation function is as follows:

$$p(\boldsymbol{t}|\boldsymbol{\nu},\theta,\boldsymbol{\xi},\boldsymbol{y}) = \mu^{n_I}\Big[\prod_{i:t_i\in O}(\sum_{k=1}^{L}\nu_k)\mathrm{Ga}(t_i - t_{y_i}|\xi_i,\theta^{-1})\Big] \times \exp\{-\mu T\}\exp\Big\{-\sum_{t_i\in\boldsymbol{t}}\sum_{l=1}^{L}\nu_l\int_0^{T-t_i}\mathrm{Ga}(u|l,\theta^{-1})du\Big\} \tag{7}$$

$$p(\xi_i|\boldsymbol{\nu}) \propto \sum_{l=1}^{L}\frac{\nu_l}{\sum_k \nu_k}\delta_l(\xi_i),\ \ i: t_i \in O,$$

where, given the branching latent variables, n_I denotes the number of points that belong to the immigrant process, $|\{t_i : y_i = 0,\ \ i = 1,\ldots,n\}|$. The priors for model parameters ν_l, θ, and b_{F_0} are as in Section 2.2.1.

We can take posterior samples for model parameters using the general Gibbs sampling method. Although the Erlang mixture has been adjusted for the excitation function, the weights ν_l retain conjugate (gamma) priors. So, we can draw posterior samples for ν_l using ready prior-to-posterior updating. The constant immigrant intensity μ also has a well-known distribution for posterior sampling under its exponential prior. Details of the MCMC posterior simulation are available in Supplementary Material B.2.

2.2.2 Uniform mixture models for decreasing offspring densities

In HP parametric modeling, both exponential and power-law densities are commonly used as the offspring density function. But, these two densities exhibit different tail behavior, owing to their exponential tail or polynomial tail. In the data analysis of Section 5, we applied each density function to HP parametric modeling for earthquake occurrences. The choice of the offspring density function brought about remarkably different estimation results. The unfavorable implications of the model choice have motivated our new modeling approach. We propose a semiparametric modeling framework with offspring density that is capable of capturing any non-increasing densities.

Our method is based on the fact that for any non-increasing density f on the positive real line there exists a distribution function F on $[0,\infty)$ such that $f(t) \equiv f(t \mid F) =$

$\int \theta^{-1}\mathbf{1}_{[0,\theta)}(t)dF(\theta)$. By placing a nonparametric prior on the mixing distribution, F, we can construct a uniform mixture model which can represent any non-increasing density function on $\mathbb{R}^+$.

For F, we consider two stochastic process priors: the Dirichlet process (DP) and geometric weights (GW). Both priors can be defined via stick-breaking constructions, $F(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{Z_l}(\cdot)$. From the infinite-sum representation of the priors, the target function (i.e., the offspring density function such that $f = g$) can be derived as $g(t|F) = \int \theta^{-1}\mathbf{1}_{[0,\theta)}(t)dF(\theta) = \sum_{l=1}^{\infty} \omega_l Z_l^{-1}\mathbf{1}_{[0,Z_l)}(t)$, where the weights ω_l and locations Z_l are random, with $\sum_{l=1}^{\infty}\omega_l = 1$ almost surely. The weight is independent of $Z_l \mid \beta \overset{i.i.d.}{\sim} F_0$, where F_0 is the centering distribution function, taken to be the inverse gamma distribution with shape 3 and mean $\beta/2$.

By substituting the uniform mixture for the offspring density function (the target function), we can define a semiparametric model as follows:

$$\lambda^*(t) \;=\; \mu + \sum_{t_j<t} \gamma g(t-t_j) \;=\; \mu + \sum_{t_j<t} \gamma \Big[\sum_{l=1}^{\infty} \omega_l Z_l^{-1}\mathbf{1}_{[0,Z_l)}(t)\Big], \quad t \in \mathbb{R}^+ \tag{8}$$

By specifying the mixture weights, we can clarify the differences between the two process models. Denote by $\mathrm{Be}(a,b)$ the beta distribution with mean $a/(a+b)$. The weight of the Dirichlet process prior is defined through the following stick-breaking mechanism,

$$\omega_1 = \varphi_1 \qquad \omega_l = \varphi_l \prod_{r=1}^{l-1}(1-\varphi_r), \quad l \geq 2, \tag{9}$$

with $\varphi_l \mid \alpha \overset{i.i.d.}{\sim} \mathrm{Be}(1,\alpha)$, and $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$. On the other hand, the weight of the GW prior is given in the form

$$\omega_l = (1-\zeta)\,\zeta^{l-1}, \quad l = 1,2,\ldots, \tag{10}$$

with $\zeta \sim \mathrm{Be}(a_\zeta, b_\zeta)$. The method of generating ω_l in (10) can be envisioned as a simplified stick-breaking technique in which the stick is always broken with the same size, $\nu_l = 1-\zeta$ for all l. Unlike the DP prior, the GW has ordered mixture weights and, thus, identifiable

location parameters; for instance, the first location parameter is always given the most weight.

Regarding the hierarchical model representation, we have

$$p(\boldsymbol{t}|\boldsymbol{y},\mu,\gamma,\boldsymbol{\xi},\boldsymbol{Z}) \propto \mu^{n_I}\gamma^{n_O}\Big[\prod_{i:t_i\in O}\frac{1}{Z_{\xi_i}}1_{(0,Z_{\xi_i})}(t_i-t_{y_i})\Big]\exp\{-\mu T\}\exp\Big\{-\sum_{l=1}^{L}\gamma\omega_l K(Z_l)\Big\}$$
$$p(\xi_i|\boldsymbol{\omega}) = \sum_{l=1}^{L}\omega_l\delta_l(\xi_i),\ \ i:t_i\in O, \tag{11}$$

with the mixture weights given by either (9) for the DP prior or (10) for the GW prior, and

$$K(Z_l) = \begin{cases} \dfrac{1}{Z_l}\Big(\sum_{i=1}^{n}(T-t_i)\Big) \ \text{ for } \ Z_l > T-t_1; \\ r+\dfrac{1}{Z_l}\Big(\sum_{i=r+1}^{n}(T-t_i)\Big) \ \text{ for } \ T-t_{r+1} < Z_l \le T-t_r,\ \ r=1,\ldots,n-1; \\ n \ \text{ for } \ 0 < Z_l \le T-t_n. \end{cases}$$

The hierarchical representation excludes the priors for the branching structure, but they are the same as in (5), and for model parameters Z_l, μ, γ, and β, given in Section 2.2.2.

The posterior inference method for the models is based on the blocked Gibbs sampler. Specifically, we use the algorithm for estimating the offspring density function, modeled by the uniform mixture with the DP or GW prior. With an exponential prior for the constant immigrant intensity μ, we can simply draw the posterior sample for the parameter from a well-known distribution. The key parameter Z_l also has a conjugate (inverse gamma) prior. An exponential term in (11) involves the function $K(Z_l)$, which is defined differently based on the postion of Z_l in a partition of $\mathbb{R}^+$. The parameter Z_l, therefore, has a piecewise truncated (inverse gamma) distribution for its posterior sampling. A complete description of the MCMC inference method are given in Supplementary Material B.3.

2.3 Fully NPB models for the HP intensity function

To achieve fully nonparametric model extensions, we can combine the Erlang mixture model for the immigrant intensity with either of the models for the excitation function. In particular, the NPB model based on Erlang mixtures is defined as

$$\begin{aligned}
\lambda^*(t) &= \sum_{l=1}^{L_I} \nu_{l,I} \mathrm{Ga}(t|l, \theta_I^{-1}) + \sum_{t_i<t} \Big[\sum_{l=1}^{L_O} \nu_{l,O} \mathrm{Ga}(t-t_i|l, \theta_O^{-1}) \Big], \quad t \in \mathbb{R}^+ \\
\nu_{l,I} &\overset{ind.}{\sim} \mathrm{Ga}(A_{l,I}, c_0), \;\; A_{l,I} \equiv c_0[G_0(l\theta_I) - G_0((l-1)\theta_I)], \\
\nu_{l,O} &\overset{ind.}{\sim} \mathrm{Ga}(A_{l,O}, \eta), \;\; A_{l,O} \equiv \alpha_0[F_0(l\theta_O) - F_0((l-1)\theta_O)],
\end{aligned} \tag{12}$$

where $G_0(t) = t/b_{G_0}$ and $F_0(t) = 1 - \exp\{-t/b_{F_0}\}$. The prior is specified following the strategy given for each of the Erlang-mixture models in Sections 2.1 and 2.2.1.

Similarly, to accommodate decreasing offspring densities in a fully NPB model, the uniform mixtures developed in Section 2.2.2 can be substituted for the offspring Erlang mixture of (12):

$$\begin{aligned}
\lambda^*(t) &= \sum_{l=1}^{L_I} \nu_{l,I} \mathrm{Ga}(t|l, \theta_I^{-1}) + \sum_{t_i<t} \Big[\int \theta^{-1} \mathbf{1}_{[0,\theta)}(t-t_i) dF(\theta) \Big], \quad t \in \mathbb{R}^+ \\
\nu_{l,I} &\overset{ind.}{\sim} \mathrm{Ga}(A_{l,I}, c_0), \;\; A_{l,I} \equiv c_0[G_0(l\theta_I) - G_0((l-1)\theta_I)] \\
F &\sim \mathrm{DP}(F_0, \alpha_0) \;\; / \;\; \mathrm{GW}(F_0, \zeta).
\end{aligned} \tag{13}$$

Again, the prior for the model parameters is specified in the same manner as in Sections 2.1 and 2.2.2.

The hierarchical model representation for the fully NPB models is obtained by combining the structure for the semiparametric models. Given the branching structure, we can separately estimate the immigrant intensity and the excitation function, using the algorithms detailed in the Supplementary Material.

3 Inference, prediction, and model assessment

3.1 First-order and second-order intensities

Presented here are the first-order and second-order intensity functions, which can be used to study HPs in conjunction with the conditional intensity function. We begin by defining the HP conditional intensity function using the counting process, which helps clarify the notion of the two functions. The conditional intensity function can be interpreted as the conditional expected rate of arrivals at t, given the history $\mathcal{H}(t)$, times observed in $[0, t)$. So, we can write the function as $\lambda^*(t) = \mathrm{E}(dN(t)/dt|\mathcal{H}(t))$, where $N(t)$ is a counting process at t. The first-order intensity represents an averaged intensity function over the history. In other words, the first-order intensity $\lambda(t)$ is the expectation of the HP conditional intensity function, defined as $\lambda(t) = \mathrm{E}(\lambda^*(t))$. The second-order intensity $r_t(\tau)$ is the HP covariance structure, given by $r_t(\tau) \equiv r(t, \tau) = \mathrm{Cov}(dN(t)/dt, dN(t+\tau)/d\tau)$ for $\tau > 0$ and $\lambda(t)$ for $\tau = 0$. Hawkes (1971) presented derivations of the two intensities under a simple parametric model consisting of a constant immigrant intensity and an exponential offspring density function.

To derive the first-order intensity under our models, we start with a general expression of the first-order intensity, $\lambda(t) = \mu + \int_0^t h(u)\lambda(t-u)du$ (Laub et al., 2021). Taking the Laplace transform of the expression yields

$$\begin{aligned} \mathcal{L}[\lambda(t)](s) &= \mathcal{L}[\mu](s) + \mathcal{L}\Big[\int_0^t h(u)\lambda(t-u)du\Big](s), \quad \text{where} \\ \mathcal{L}[\mu](s) &= \int_0^\infty \exp\{-st\}\mu dt = \frac{\mu}{s} \\ \mathcal{L}\Big[\int_0^t h(u)\lambda(t-u)du\Big](s) &= \mathcal{L}[h(t)](s)\mathcal{L}[\lambda(t)](s). \end{aligned} \tag{14}$$

Under the offspring Erlang mixture model, for which the offspring density is defined as $h(t) = \sum_{l=1}^{L} \nu_l(\theta^{-l}/\Gamma(l))t^{l-1}\exp\{-t\theta^{-1}\}$, $\nu_l = \gamma\omega_l$, the Laplace transform of the excitation

function is derived as

$$\begin{aligned}\mathcal{L}[h(t)](s) &= \int_0^\infty \exp\{-st\}\Big(\sum_{l=1}^{L}\nu_l\frac{\theta^{-l}}{\Gamma(l)}t^{l-1}\exp\{-t\theta^{-1}\}\Big)dt \\ &= \sum_{l=1}^{L}\nu_l\int_0^\infty \frac{\theta^{-l}}{\Gamma(l)}t^{l-1}\exp\{-t(s+\theta^{-1})\}dt \\ &= \sum_{l=1}^{L}\nu_l\Big(\frac{1}{s\theta+1}\Big)^l.\end{aligned} \tag{15}$$

Substituting (15) into (14) provides an analytical form of the Laplace transform of the first-order intensity, $\mathcal{L}[\lambda(t)](s) = \mu/\big[s\big(1-\sum_{l=1}^{L}\nu_l(s\theta+1)^{-l}\big)\big]$.

Similarly, under the uniform-mixture-based models with the excitation function $h(t) = \sum_{l=1}^{L}(\nu_l/\theta_l)1_{(0,\theta_l)}(t)$, $\nu_l = \gamma\omega_l$, we can derive the Laplace transform of the first-order intensity as follows

$$\mathcal{L}[h(t)](s) = \int_0^\infty \exp\{-st\}\Big(\sum_{l=1}^{L}\frac{\nu_l}{\theta_l}1_{(0,\theta_l)}(t)\Big)dt = \frac{1}{s}\sum_{l=1}^{L}\frac{\nu_l}{\theta_l}(1-\exp\{-s\theta_l\}). \tag{16}$$

Substituting (16) into (14) yields an analytical form of the Laplace transformed first-order intensity, $\mathcal{L}[\lambda(t)](s) = \mu/\big[s\big(1-(\sum_{l=1}^{L}(\nu_l/\theta_l)(1-\exp\{-s\theta_l\}))/s\big)\big]$.

By numerical inverse transforms, we can readily obtain the first-order intensities under the two semiparametric models. Further development of the intensity can be accomplished by replacing μ with $\mu(t)$ for our nonparametric models. Since the immigrant intensity of the nonparametric models is the Erlang mixture, we can obtain an analogous result to (15) for the Laplace transform of $\mu(t)$. Therefore, Laplace transforms of the intensity of the models remain analytically available.

Unlike the first-order intensity, the proposed models do not offer such an analytical solution for the second-order intensity. Instead, we can approximate the intensity of our models using Monte Carlo integration. As the second-order intensity is a covariance function, it is denoted by

$$r_t(\tau) = \mathrm{E}\big[(dN(t)/dt)(dN(t+\tau)/d\tau)\big] - \mathrm{E}\big[dN(t)/dt\big]\mathrm{E}[dN(t+\tau)/d\tau\big].$$

The first expectation can be expressed via the double expectation theorem as

$$\mathrm{E}\big[(dN(t)/dt)(dN(t+\tau)/d\tau)\big] = \mathrm{E}\Big[\mathrm{E}\big[(dN(t)/dt)(dN(t+\tau)/d\tau)|\mathcal{H}(t+\tau)\big]\Big].$$

Given the history up to time $t+\tau$ (i.e., $\mathcal{H}(t+\tau)$) for $\tau > 0$, $N(t)$ of the first derivative can be viewed as a non-random count of observed points $t_i < t$. So, we can move the derivative $dN(t)/dt$ outside the inner expectation, leaving $\mathrm{E}\Big[\big(dN(t)/dt\big)\mathrm{E}\big(dN(t+\tau)/d\tau|\mathcal{H}(t+\tau)\big)\Big]$. We approximate the derivative numerically by using the central difference, such that $dN(t)/dt \approx (N(t+h) - N(t-h))/2h$ for $h > 0$. In accordance with the definition of the HP conditional intensity, the remaining inner expectation becomes a conditional intensity $\lambda^*(t+\tau)$, which simplifies the covariance function as $r_t(\tau) \approx \mathrm{E}\Big[\big[\big(N(t+h) - N(t-h)\big)/2h\big]\lambda^*(t+\tau)\Big] - \mathrm{E}\big[dN(t)/dt\big]\mathrm{E}[dN(t+\tau)/d\tau\big]$. Note that each expectation of the second term is a first-order intensity, such that

$$\mathrm{E}(dN(t)/dt) = \mathrm{E}\Big[\mathrm{E}\big[dN(t)/dt|\mathcal{H}(t)\big]\Big] = \mathrm{E}\big[\lambda^*(t)\big].$$

Thus, by replacing the expectations with first-order intensities, we can further simplify the second-order intensity to

$$r_t(\tau) \approx \mathrm{E}\Big[\big[\big(N(t+h) - N(t-h)\big)/2h\big]\lambda^*(t+\tau)\Big] - \mathrm{E}\big[\lambda^*(t)\big]\mathrm{E}\big[\lambda^*(t+\tau)\big]. \qquad (17)$$

We will then approximate the expectations through Monte Carlo integration. So, using random draws for the histories $\mathcal{H}(t)$ and $\mathcal{H}(t+\tau)$, we will compute the three expectations numerically. Our posterior simulation, therefore, involves drawing multiple HP realizations to create a set of histories. The following steps outline the procedure for posterior inference about the second-order intensity.

1. Draw multiple HP realizations over $(0, t+\tau)$ for given positive scalars t and τ, using the immigrant intensity and the excitation function specified at a MCMC iteration.

2. Plug each realization into $\{t_i\}$ of the functions, $\big[\big(N(t+h) - N(t-h)\big)/2h\big]\lambda^*(t+\tau)$, $\lambda^*(t)$, and $\lambda^*(t+\tau)$, where, given the history, $N(t) = |\{t_i : t_i < t\}|$.

3. Take the average of each function across realizations for Monte Carlo integration and substitute each approximate into the corresponding expectation of (17).

4. Repeat the above steps for each MCMC iteration to obtain posterior samples of the second-order intensity.

3.2 Model assessment and comparison

This section discusses a few methods for evaluating the performance of models. A histogram of data can be used to determine the goodness-of-fit of a model in density estimation by comparing it to the estimated function. Likewise, one can assess the goodness-of-fit of immigrant / offspring intensity models by mapping the histograms of observed immigrant / offspring points to the corresponding estimated functions. Specifically, the goodness-of-fit of immigrant intensity models can be determined by aligning the immigrant points' histogram with the posterior estimate of immigrant intensity normalized over the observation window, $\mu(t)/\int_0^T \mu(u)du$. The comparison with the histogram of offspring points requires some additional consideration. According to the cluster process representation, the offspring data of a HP realization is a collection of all descendants from multiple families; simulating children of a family is based on a PP with intensity of the excitation function supported on $(0, T - t_i)$, where $\{t_i\}$ is a parent of the family. Accordingly, offspring points are not sampled from a common distribution as the immigrant data, but rather from a distribution that varies according to the parent $\{t_i\}$. In order to accommodate such offspring data, we designed a mixture of offspring density functions with varying support, which we referred to as the *aggregate offspring density function*. Denoted by $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$ an observed point pattern, the aggregate offspring density function is defined as

$$q(x|\boldsymbol{t}, \tilde{\boldsymbol{y}}, h) = \frac{\sum_{k=1}^{\tilde{n}} H(T - t_{\tilde{y}_k}) \times \left(\frac{h(x)}{H(T-t_{\tilde{y}_k})} \mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x) \right)}{\sum_{k=1}^{\tilde{n}} \int_0^T h(s) \mathbf{1}_{(0,T-t_{\tilde{y}_k})}(s) ds} = \frac{\sum_{k=1}^{\tilde{n}} h(x) \mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x)}{\sum_{k=1}^{\tilde{n}} H(T - t_{\tilde{y}_k})}, \quad (18)$$

where $H(t)$ denotes the cumulative excitation function, given by $H(t) = \int_0^t h(s)ds$, and $\tilde{\boldsymbol{y}} = \{\tilde{y}_k\}$, $k = 1, 2, \ldots, \tilde{n}$, the set of unique values of the branching structure $\{y_1, \ldots, y_n\}$ with $\tilde{n} < n$. In the numerator, $\left(\frac{h(x)}{H(T-t_{\tilde{y}_k})}\mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x)\right)$ represents the offspring density (normalized excitation) function over $(0, T-t_{\tilde{y}_k})$. The other term $H(T-t_{\tilde{y}_k})$ indicates how many points are expected to be drawn from the offspring density function, corresponding to the expected cumulative intensity of the immigrant PP in the cluster process representation. Thus, the product of the terms in the numerator represents an intensity function for descendants of a family with a parent $t_{\tilde{y}_k}$, and we add them up for all $t_{\tilde{y}_k}$, which we assume are given, for all families. Consequently, normalizing the mixture over $(0, T)$ with the normalizing constant (also called the compensator) in the denominator provides a density function that is comparable to the offspring data.

The remainder of the section will present some numerical criteria that enable quantitative comparisons of models. The first is the total variation distance (TVD). Based on the target functions to which our posterior estimated cumulative distributions are matched, we define three TVDs:

- TVD^a: the immigrant cumulative distribution, $\mathrm{M}(t) = \int_0^t \left[\mu(u)/\int_0^T \mu(s)ds\right]du$;
- TVD^b: the offspring cumulative distribution, $\mathrm{G}(x) = \int_0^x g(u)du$; and
- TVD^c: the empirical distribution of offspring data.

Again, as the offspring data is a collection of descendants from all families, we applied the aggregate offspring density function to TVD^c by mapping the empirical distribution to the cumulative distribution $\mathrm{Q}(x) = \int_0^x g(u|\boldsymbol{t}, \tilde{\boldsymbol{y}}, h)du$, where h is specified by posterior samples of model parameters.

Additionally, we propose two other criteria, both based on the branching structure: the immigrant/offspring cluster size and the immigrant/offspring misclassification. The cluster sizes are defined as $n_I = |\{i : y_i = 0,\ \ i = 1, \ldots, n\}|$ for the immigrant size and

$n_O = |\{i : y_i \neq 0,\ \ i = 1, \ldots, n\}|$ for the offspring size, where n denotes the sample size. Using posterior samples of the branching structure, we can estimate the sizes, and then compare them with observed immigrant/offspring counts for evaluating the model. Lastly, we have misclassification evaluation tools as follows:

- $M_I = |\{i : y_i = 0,\ \ y_i^{true} \neq 0,\ \ i = 1, \ldots, n\}|$ for immigrant misclassification;
- $M_O = |\{i : y_i \neq 0,\ \ y_i^{true} = 0,\ \ i = 1, \ldots, n\}|$ for offspring misclassification;
- $R = (M_I + M_O)/n$ for the misclassification rate,

where y_i^{true}, $i = 1, \ldots, n$ denotes the known branching structure. Immigrant misclassification refers to the number of immigrants that are incorrectly classified by posterior samples of the branching structure. In a similar way, offspring misclassification is also explicable. The misclassification rate provides aggregate and standardized information about the misclassification criteria.

3.3 Prediction

Chen and Stindl (2018) presents a method for predicting future event counts of the renewal Hawkes process, which predicts the number of events from the censoring time T of the observation window to a future time T^*. This is a simulation-based approach that generates points falling within the prediction window (T, T^*) by simulating the process with an estimated intensity function.

In this paper, we make posterior predictions about the future counts of a general Hawkes process. The posterior inference is also based on simulation: we draw a realization of the point process with an intensity function determined by posterior samples at each MCMC iteration.

Notice that, unlike the simulation for the second-order intensity estimation, the prediction simulates the process over the prediction window, not the observation window,

assuming that the process up to T is already given by the observed point pattern. Accordingly, the observations serve as the history of the predicted HP conditional intensity function, not only to estimate the parameters of the intensity function.

To achieve HP simulation for the prediction, we applied *Ogata's modified thinning algorithm* (Ogata, 1981; Laub et al., 2021). This is a sequential approach that generates a candidate based on the latest time point in a sample (and updates the sample by accepting the candidate randomly with a probability). So, the simulation of the process over the prediction window can be obtained by resuming the algorithm with sampling from the last point observation until the new candidate is greater than T^*. The thinning algorithm requires the intensity function to be non-increasing in periods of no arrivals, which is not the case when the excitation function of the intensity is non-decreasing. For instance, our Erlang mixture model in Section 2.2.1 can generate a mode in the excitation function, so the HP excitation can occur with a margin of time. This results in the conditional intensity developing a mode rather than decreasing in time during the period of no arrivals. Due to the mode, $\lambda^*(t+\epsilon)$ with some tiny value $\epsilon > 0$ in the algorithm no longer serves as the upper bound for the conditional intensity $\lambda^*(t)$. As such, models that do not meet the non-increasing intensity requirement may need an alternative approach to finding the upper limit of the intensity function, such as discretizing the intensity with grid points. In Section 5, we will present predictions for an earthquake data set under the uniform-mixture-based models. Due to the fact that the excitation function of the models confines their representation to a non-increasing function, the predictions have been made without any modifications to the algorithm.

4 Simulation study

We conducted simulation studies with a range of synthetic data examples to empirically assess the performance of the proposed models. This section summarizes the simulation con-

figuration and results, with additional detail provided in Supplementary Material D. Supplementary Material D.1 presents examples that are intended to evaluate Erlang-mixture-based semiparametric and nonparametric models. Supplementary Material D.2 focuses on uniform-mixture-based semiparametric models.

Appendices D.1.1 and D.1.2 demonstrate the flexibility of the semiparametric approaches that model immigrant intensity or the excitation function using an Erlang mixture. Examples in the appendices feature irregular-shaped underlying functions (either unimodal or bimodal) for the immigrant intensity or the excitation function. The Erlang mixtures of the semiparametric models cover well such patterns that are difficult to capture with standard models. Additionally, the semiparametric models provide estimates that are in agreement with the true values for cluster sizes (n_I and n_O), as well as the branching ratio.

Supplementary Material D.1.3 contains three examples showing the performance of the nonparametric model based on Erlang mixtures. In the first example, a simple but typical form of HP conditional intensity is used for simulation, consisting of constant immigrant intensity and exponential offspring density. In other words, the synthetic data emulates one of the simplest point patterns one would expect to see in HP applications.

Figure S6 in Supplementary Material shows that the nonparametric model performs well in estimating immigrant intensity and offspring density functions. Particularly, it is positive that the nonparametric model captures the constant underlying immigrant intensity without being overfitted, despite the absence of a constant-shaped component in the mixture used to estimate immigrant intensity.

In the next two examples, we simulated HP point patterns with irregular-shaped underlying intensity functions to demonstrate the flexibility of the nonparametric model. Figures S8 and S9 illustrate that the model reproduces well the global patterns of the underlying functions and includes them within the 95% posterior interval estimates. There are some local discrepancies in the posterior estimates, such as the shift of the second mode to the

left in Figure S8. Such discrepancies are primarily due to unrepresentative data or/and model misclassification (M_I and M_O). For instance, in Figure S9, the histogram of the observed data and the aggregate offspring density estimates following it explain why the offspring density estimates (in the middle panel) are greater than the underlying function at the two modes. On the other hand, the estimated aggregate offspring density function in Figure S8 does not conform to the data histogram. However, there are a large number of misclassified offspring points in the example, which may cause confounding effects on the estimation of the offspring density function, leading to discrepancies.

Lastly, Supplementary Material D.2 illustrates semiparametric models based on uniform mixtures. Because the mixtures of the models specialize in representing decreasing functions, we considered examples in which the offspring density functions decreased over time but at different rates for each example, both of which are well captured by the models (Figure S10).

5 Real data analysis

Ogata (1988) provided a catalog of earthquakes with magnitudes of six or greater that occurred in Japan and its vicinity from 1885 through 1980 (over approximately $34,711$ days). We analyzed 458 point observations (258 main shocks and 200 aftershocks), obtained by removing 25 foreshocks from the total 483 observations in Ogata (1988) (see Figure 1 for the histogram of the data).

To show the model performance, we compared our semiparametric (Semipara) and nonparametric (Nonpara) models with the ETAS model (Para), which is commonly used for earthquake occurrence data analysis. As HP models in seismic applications assume decreasing excitation functions over time (e.g., Ogata, 1988; Kagan, 1991; Musmeci and Vere-Jones, 1992; Ogata, 1998), we chose the uniform DP mixture of Section 2.2.2 for modeling the offspring density function of Semipara and Nonpara. Immigrant intensity is

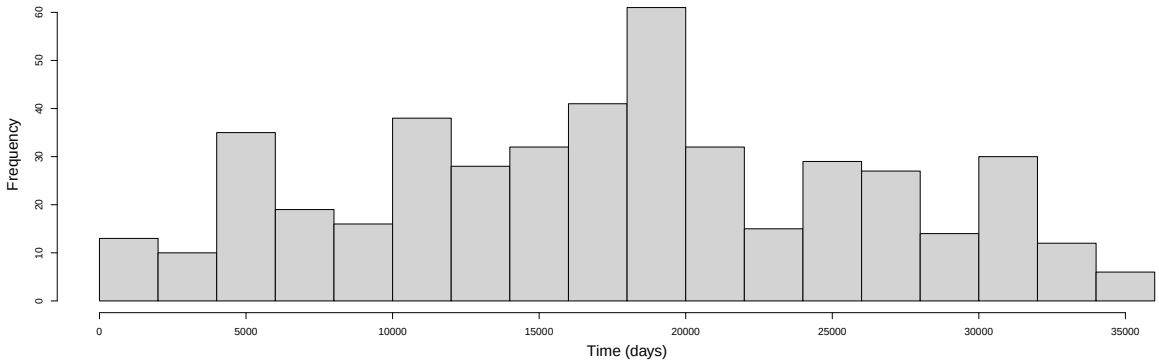

Figure 1: Histogram of earthquakes in Japan from 1885 through 1980.

	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
Para	0.897(0.291)	220(19)	238(19)	48(7)	86(14)	0.292(0.025)
Semipara	0.459(0.066)	250(25)	208(25)	54(8)	61(18)	0.251(0.029)
Nonpara	0.436(0.064)	261(24)	197(24)	57(8)	54(17)	0.242(0.027)

Table 1: Posterior means and standard deviations of the branching ratio, cluster sizes, and misclassification under each model.

treated as a constant in the semiparametric model, whereas the nonparametric model has a more flexible structure via an Erlang mixture.

Here is the set of priors used to specify the models: Exp(1) for both parameters p and c of the power-law density function in Para; Exp(0.0769) for a_{F_0}, Exp(0.0083) for β, and Ga$(4, 0.5)$ for α with $L = 85$ for Semipara; and, for Nonpara, Lo$(2, 1800)$ for θ_I, $L_I = 120$, Exp(0.1) for c_0, and Exp(0.0065) for b_{G_0}, along with the same priors for the uniform DP mixture as in the semiparametric model. We assigned Exp(153) and Ga$(2, 4)$ priors to the common parameters μ and γ of all models.

The uniform DP mixture of Semipara (and Nonpara) reduces the branching ratio es-

timate by almost half compared to Para (Table 1). This helps Semipara to obtain more relevant cluster size estimates: the posterior means of 250 and 208 are relatively close to the observed values, 258 and 200. On the other hand, the ETAS model produces a large offspring cluster of estimated size 238, which does not contain the observed value 208 in its 95% interval estimates. It indicates that Para's posterior mean of 0.897 for γ is overestimated. In contrast, the nonparametric model further improves the cluster size estimation by incorporating a flexible structure for immigrant intensity, which provides n_I (and n_O) estimates with the smallest biases. Nonpara also enhances the branching structure estimation with the lowest misclassification rate.

The time-rescaling theorem (also called the random time change theorem) is a graphical tool used to assess the validity of a model, based on transformed times, $(\Lambda(t_1), \ldots, \Lambda(t_n))$, $\Lambda(t) = \int_0^t \lambda^*(x)dx$, that form a Poisson process with unit rate (e.g., Daley et al., 2003; Laub et al., 2021). Therefore, the conditional intensity function $\lambda^*(\cdot)$ that describes the underlying point pattern perfectly allows the transformed time points to constitute a homogeneous Poisson process and so their inter-arrival times to follow a uniform distribution. Consequently, intensity estimates with a higher degree of accuracy should result in inter-arrival times whose distribution is more uniform, yielding a Q-Q plot with estimated quantiles that match theoretical (uniform) quantiles. In this criterion, our semi- and non-parametric models outperform the ETAS model, with posterior interval estimates covering the red line fully, whereas the ETAS model struggles to capture around 0.05 and 0.2.

The second row of Figure 2 shows immigrant intensity estimates. The purple line represents an intuitive and data-driven estimator $\tilde{\mu}$ for the immigrant intensity under a constant model assumption, that is, the immigrant count scaled over the observation window, $\tilde{\mu} = |\{t_i : y_i = 0\}|/T = 258/35000$. This can be used to assess the models Para and Semipara in immigrant intensity estimation by measuring how closely the immigrant estimates of the models match $\tilde{\mu}$. The ETAS model underestimates the data-driven value, while the

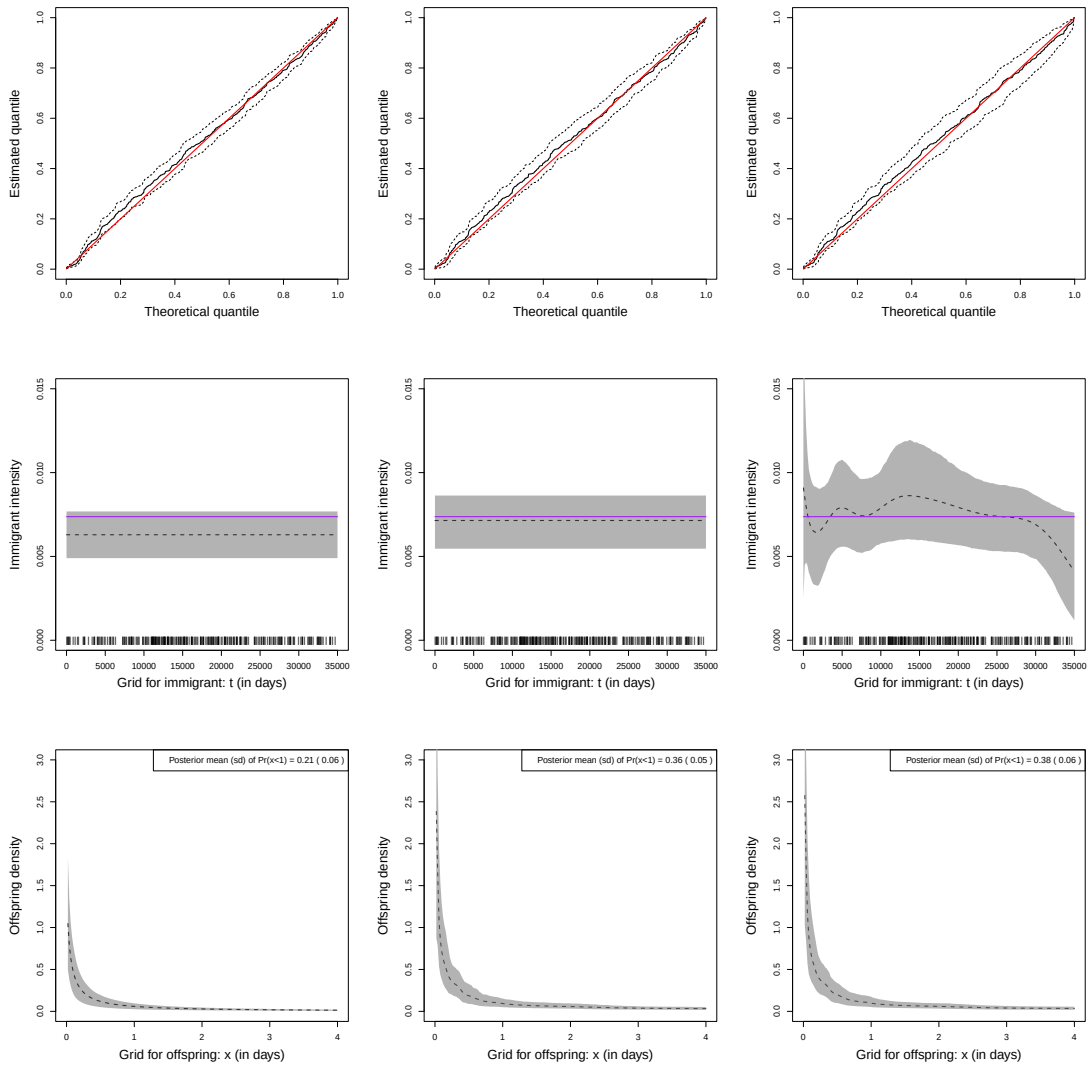

Figure 2: From left to right, parametric (Para), semiparametric (Semipara), and nonparametric (Nonpara) models. Q-Q plots in the first row display the results of the time-rescaling theorem. The next two rows demonstrate estimated functions for the immigrant intensity and the offspring density. Bars at the bottom of the panels in the second row indicate occurrence times of the main shock.

semiparametric model produces an immigrant estimate comparable to $\tilde{\mu}$. This result is consistent with the findings in Table 1, which demonstrates that the models have posterior means of 220 and 250 for immigrant cluster size (compared to the observed value of 258). The nonparametric model yields immigrant estimates that are not standard-shaped with multiple fluctuations. This intensity estimated function, in conjunction with the model comparison results in Table 1, casts doubt on the constant immigrant intensity assumption

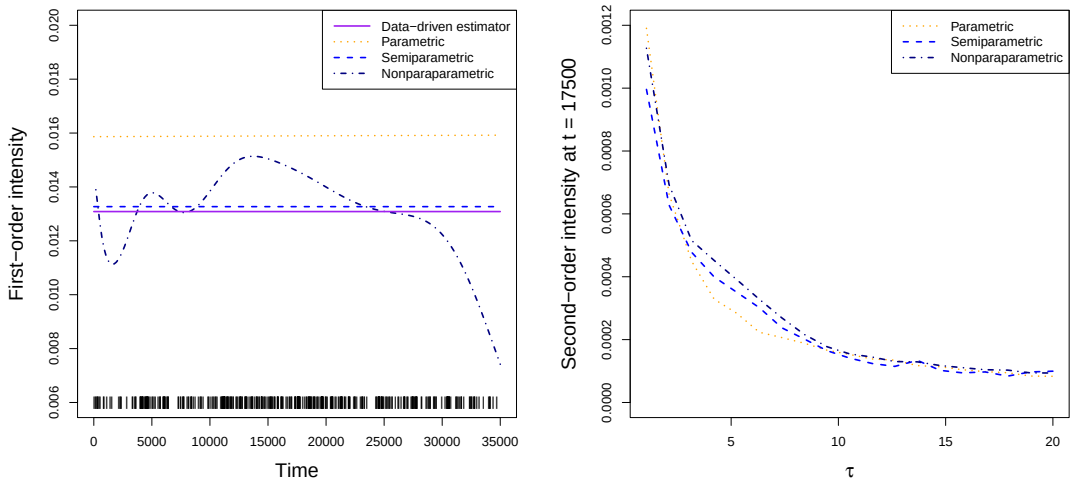

Figure 3: Posterior means for the first- and second-order intensities in the left and right panels. Bars at the bottom on the left indicate the whole data points.

of the parametric and semiparametric models.

Presented in the last row of the figure are offspring density estimates. Semipara and Nonpara result in more zero-concentrated estimated functions; under the two models, over one-third of the probability falls within the unit interval, whereas Para provides a relatively sparse distribution with a heavier tail.

The left panel of Figure 3 displays the posterior means for the first-order intensity (described in Section 3.1), as well as a data-driven estimator $\tilde{\lambda}(t)$ in purple. Based on the assumption of constant first-order intensity, $\tilde{\lambda}(t) = \tilde{\lambda}$ is defined by mapping $\int_0^T \tilde{\lambda}(u)du = n$ with the size of observed point pattern, $n = 458$, on the observation window $(0, T)$. The ETAS model has markedly larger first-order intensity estimates than those of the other models and the reference data-driven estimator. In particular, the fact that the estimates exceed $\tilde{\lambda}$ implies that the cumulative first-order intensity over $(0, T)$ is greater than n and therefore that the intensity in the parametric model is overestimated. The nonparametric model produces nonstandard-shaped first-order intensity estimates; they are similar in shape to the histogram of the whole data in Figure 1, which justifies the complex pattern of

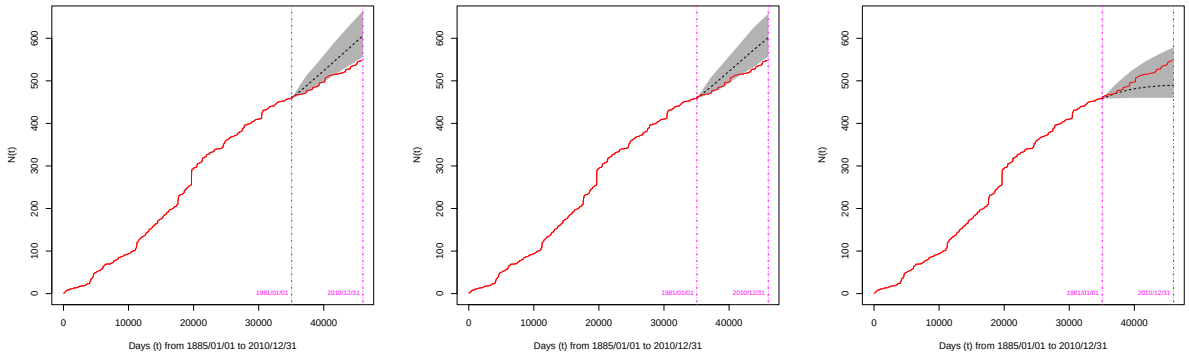

Figure 4: Posterior predicted mean (dashed curve) and 95% interval estimates (shaded area) for earthquakes in the prediction window of $(1981, 2010)$. From left to right, parametric semiparametric, and nonparametric models. The red line denotes the actual earthquake counts in the entire time window from 1885 to 2010.

the intensity estimates. The right panel shows the second-order intensity estimated function introduced in (17). As an example, we provide the intensity estimates at $t = 17,500$ (middle of the observation window) for τ ranging from 0 to 20, where they all have a common pattern with a similar decreasing rate.

Figure 4 depicts the prediction of earthquakes in the same region of Japan (geographical location specified in Ogata 1988) over the next 30 years (1981 to 2010). Using the method described in Section 3.3, we obtained the prediction results, for which Japan earthquake data from 1885 to 1980 were used to fit the models and to further update the predicted conditional intensity. To evaluate the predictions, we compare the results with the actual 30-year earthquake catalogs for the region, acquired from the USGS earthquake archive, available at `https://earthquake.usgs.gov/earthquakes/search/`. In Figure 4, both parametric and semiparametric models overpredict, with their prediction intervals failing to capture the true values at half of the prediction window. In contrast, the nonparametric model, despite being slightly underpredictive, has its prediction interval perfectly covering the true values throughout the time window.

6 Summary

We have developed a nonparametric Bayesian modeling framework for temporal Hawkes processes. The primary motivation is to enhance the scope for inference and prediction under this popular class of self-exciting point processes. We have proposed different nonparametric prior models for the immigrant intensity and for the excitation function (or offspring density), the two components of the HP conditional intensity. A key consideration for the model constructions has been to balance flexible shapes for the HP functions with computational efficiency in model implementation. Particularly important regarding the latter is the HP branching structure cluster representation. The modeling framework includes an option that supports general decreasing offspring densities. This option may be appealing for HP applications to earthquake modeling, since it incorporates the meaningful constraint of a decreasing offspring density and, at the same time, it allows for general tail behavior to be uncovered. We have designed posterior simulation algorithms to implement the different models, and developed methods for inference for the HP first-order and second-order intensities, for model assessment, and for prediction. Different aspects of the methodology have been explored through a comprehensive simulation study, and with the analysis of high-magnitude earthquake occurrences in Japan and its vicinity.

SUPPLEMENTARY MATERIAL

A. Prior specification: Strategies for specifying the priors of model parameters.

B. MCMC details: Details of MCMC posterior inference methods used.

C. Sensitivity analysis: Results of sensitivity analyses for semiparametric models for the excitation function.

D. Simulation study: Simulation study results for all models proposed.

Supplementary Material for "Nonparametric Bayesian Modeling and Inference for Hawkes Processes"

A. Prior specification

Here, we discuss approaches to prior specification for the different models for the Hawkes process (HP) conditional intensity developed in Section 2 of the paper.

A.1 Erlang mixture model for the immigrant intensity function

Consider the model for the immigrant intensity based on the weighted combination of Erlang densities (Section 2.1). The model is implemented with a prior on θ, hyperpriors for the gamma process parameters, c_0 and b_{G_0}, and with a specified value for the number of Erlang basis densities, L. The prior specification strategy discussed below builds from the approach in Kim and Kottas (2022).

A useful result is that, under the exponential G_0, the prior expectation for the immigrant intensity function (over the gamma process prior) is roughly constant at $1/b_{G_0}$. The integration of the function over $(0, T)$ (i.e., a cumulative intensity T/b_{G_0}) provides a reasonable approximation to the total immigrant intensity in the bounded interval. The number of observed immigrant points (or half of the total point observations) can be used as a proxy for the immigrant intensity, so we can specify an exponential prior for b_{G_0} by matching the value with the prior mean.

As for the precision parameter c_0, we take the Exp(0.1) prior with mean 10. For G variability around the centering function G_0, we prefer to have smaller values, and therefore consider a simple, but monotonically decreasing prior distribution. Empirically, the parameter choice, 0.1, is regarded as a conservative value, resulting in substantial prior-to-posterior learning towards 0.

The common scale parameter, θ, of the Erlang density basis functions is assigned the Lomax prior with shape 2 and scale d_θ. The prior choice with the shape parameter allows for infinite variance and thus a large effective support for θ. Like the prior specification for c_0, we choose the hyperparameter, d_θ, that can produce conservative values for θ, causing the posterior distribution to shift left from the prior distribution. Consider the

least flexible model of a single component (i.e., $L = 1$). Under the model, $(0, L\theta) = (0, \theta)$ becomes a rough guess about the effective support for the immigrant intensity function. Using the observation window as a proxy for the effective support, d_θ is chosen such that $\Pr(0 < \theta < T) \approx 0.999$. Such a choice is conservative because more flexible models of $L > 1$ would lead to smaller values for θ through the adaptive upper bound $L\theta$ of the support with the fixed proxy interval.

Using the effective support $(0, L\theta^*)$ and its proxy $(0, T)$, we set the number of mixture components, L, to the integer part of T/θ^*, where θ^* denotes the prior median for θ. In order to achieve a more conservative value for L, we recommend conducting sensitivity analysis based on the value as a lower bound. A larger L may improve estimation of immigrant intensity when the underlying immigrant intensity is deemed non-standard.

A.2 Erlang mixture model for the excitation function

Below is a strategy for specifying the parameters of the Erlang mixture model presented in Section 2.2.1. First, we need to choose the prior for θ and the number of mixture components L for the weighted combination of Erlang densities. Similar strategies described in the previous section can be applied to them. Then, the three hyperparameters $(\alpha_0, \eta, b_{F_0})$ involved in the mixture weights remain to be determined. α_0 and η govern the (gamma) prior distribution of the branching ratio γ by serving as shape and scale parameters, respectively; we exploit the stability condition $\gamma \in (0, 1)$ to set the two hyperparameters. b_{F_0} is the rate parameter of the exponential cumulative distribution $F_0(t) = 1 - \exp\{-b_{F_0}t\}$. Section A.2.1 defines the *asymptotic expected offspring density function*, with which the prior for b_{F_0} is established.

We take the Lomax prior for θ with shape 2 and scale d_θ, such that $\Pr(0 < \theta < T_O) \approx 0.999$, where $(0, T_O)$ is a proxy for the effective support, $(0, L\theta^*)$, of the excitation function. Based on background knowledge of the application the model is being used for, T_O, the upper bound for offspring points, can be chosen subjectively. But, more generally, conservative choices, such as the upper bound of the observation window or the difference between the last and first observed time points, $x_{n,1} \equiv t_n - t_1$, can be a viable alternative; we observed that the posterior estimated excitation function is robust to the choice (see Section C.1 for sensitivity analysis). Let θ^* be the prior median of θ. We set L to the integer part of T_O/θ^* by mapping the upper bound of the proxy, T_O, to that of the effective

support, $L\theta^*$.

The two hyperparameters (α_0, η) are set to values that make the prior distribution $\mathrm{Ga}(\alpha_0, \eta)$ for the branching ratio have the majority of probabilities within the unit interval (without being overly focused on a few values), such that $\Pr(0 < \gamma < 1) \approx 0.9$. This choice enables our prior probability model to produce stable HPs.

We assigned an exponential prior, $\mathrm{Exp}(a_{b_{F_0}})$, to b_{F_0}. The rate parameter was chosen using the asymptotic expected offspring density function, that is, an exponential density with the rate parameter $\phi = (1 - \exp\{-\theta/b_{F_0}\})/\theta$ (see Section A.2.1 for the derivation details). Under this density function (for $t - t_y$), the mean distance (between an offspring t and its parent t_y) is given by $1/\phi$, and the distance should be in $(0, T_O)$, the presumptive window for offspring. As a conservative choice, we map the mean distance to T_O, and solving the equation for b_{F_0} leads to $b_{F_0} = \theta/-\log(1-\theta/T_O)$. The rate parameter is then defined as $a_{b_{F_0}} = 1/\mathrm{E}(b_{F_0}) = 1/\theta/-\log(1-\theta/T_O)$ for a given θ, for which we use the prior median θ^*.

A.2.1 Asymptotic expected offspring density function

As a semiparametric model, we defined the excitation function as Erlang mixture $h(x) = \sum_{l=1}^{L} \nu_l \mathrm{ga}(x|l, \theta^{-1})$, where the weight ν_l is gamma distributed with rate η and shape $A_l = \alpha_0[F_0(l\theta) - F_0((l-1)\theta)]$, $l = 1, \ldots, L-1$ and $A_L = \alpha_0[1 - F_0((L-1)\theta)]$. Taking the expectation over the gamma prior for ν_l yields the expected excitation function $\mathrm{E}(h(x)) = \sum_{l=1}^{L} A_l \mathrm{ga}(x|l, \theta^{-1})/\eta$. By normalizing the function, we can derive the expected offspring density function as

$$\begin{aligned}
\mathrm{E}(g(x)) &= \sum_{l=1}^{L} A_l \mathrm{ga}(x|l, \theta^{-1})/\alpha_0 \\
&= \sum_{l=1}^{L-1} \Big[F_0(l\theta) - F_0((l-1)\theta)\Big] \mathrm{ga}(x|l, \theta^{-1}) + \Big[1 - F_0((L-1)\theta)\Big] \mathrm{ga}(x|L, \theta^{-1}).
\end{aligned}$$

With the exponential cumulative distribution function for F_0, the expected density is expressed as,

$$\begin{aligned}
&\sum_{l=1}^{L-1}\Big[\exp\{-l\theta/b_{F_0}\}(\exp\{\theta/b_{F_0}\}-1)\Big]\mathrm{ga}(x|l,\theta^{-1})+\exp\{-L\theta/b_{F_0}\}\exp\{\theta/b_{F_0}\}\mathrm{ga}(x|L,\theta^{-1})\\
&=\sum_{l=1}^{L}\exp\{-l\theta/b_{F_0}\}\exp\{\theta/b_{F_0}\}\mathrm{ga}(x|l,\theta^{-1})-\sum_{l=1}^{L-1}\exp\{-l\theta/b_{F_0}\}\mathrm{ga}(x|l,\theta^{-1})\\
&=\sum_{l=1}^{L}\mathrm{Poisson}(l-1|\exp\{-\theta/b_{F_0}\})\theta^{-1}\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\}\\
&-\sum_{l=1}^{L-1}\mathrm{Poisson}(l-1|\exp\{-\theta/b_{F_0}\})\theta^{-1}\exp\{-\theta/b_{F_0}\}\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\},
\end{aligned}$$

where $\mathrm{Poisson}(x|\lambda)$ denotes the Poisson probability mass function with mean λ. Asymptotically, this function becomes an exponential density,

$$\begin{aligned}
\lim_{L\to\infty}\mathrm{E}(g(x)) &= \theta^{-1}\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\}-\theta^{-1}\exp\{-\theta/b_{F_0}\}\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\}\\
&=\Big[(1-\exp\{-\theta/b_{F_0}\})/\theta\Big]\exp\{-x(1-\exp\{-\theta/b_{F_0}\})/\theta\},
\end{aligned}$$

which corresponds to an exponential density function with rate $(1-\exp\{-\theta/b_{F_0}\})/\theta$.

A.3 Uniform mixture models for the offspring density function

This section discusses how we specify the priors for the uniform mixture models introduced in Section 2.2.2. The models involve three key components: branching ratio, mixture weights, and location parameters for uniform density basis functions: $\gamma \sim \mathrm{Ga}(a_\gamma, b_\gamma)$, $\omega_l = \text{S-B}(\nu_l)$ or $\text{S-B}(\zeta)$, and $Z_l \sim F_0 = \mathrm{IG}(a_{F_0}, \beta)$, respectively. S-B denotes the stick-breaking mechanism with $\nu_l|\alpha \overset{i.i.d.}{\sim} \mathrm{Be}(1,\alpha)$ and $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$ for DP or $\zeta \sim \mathrm{Be}(a_\zeta, b_\zeta)$ for GW. Hence, for the implementation of the models, we need to choose the following hyperparameters, (a_γ, b_γ), (a_{F_0}, β), and (a_α, b_α) or (a_ζ, b_ζ), along with the number of mixture components L.

As in Section A.2, we set the hyperparameters for γ, considering the HP stability condition, so that $\Pr(0 < \gamma < 1) \approx 0.999$. For the centering function F_0, we use the inverse gamma distribution $\mathrm{IG}(a_{F_0}, \beta)$, which allows the location parameters Z_l to have the posterior full conditional in closed form, namely the piecewise truncated inverse gamma distribution. The two hyperparameters are assigned $\mathrm{Exp}(a_{a_{F_0}})$ and $\mathrm{Exp}(a_\beta)$ priors; and we

specify them using the expected offspring density function, given by $g^*(x) = \mathrm{E}(g(x)) = \mathrm{E}\big(\int \theta^{-1}\mathbf{1}_{[0,\theta)}(x)dF(\theta)\big) = \int \theta^{-1}\mathbf{1}_{[0,\theta)}(x)dF_0(\theta) = \int \theta^{-1}\mathbf{1}_{[0,\theta)}(x)\mathrm{IG}(\theta|a_{F_0},\beta)d\theta$. More precisely, to set the hyperparameters $a_{a_{F_0}}$ and a_β, we exploit the cumulative distribution function, $P(x) = \int_0^x g^*(u)du$, of the density and its mean, $\mathrm{E}(X) = \int xg^*(x)dx$ (see Section A.3.1 for details on their derivation). Denote by $x = t - t_y$ the distance between a time point t and its parent t_y, and by $x_i = t_i - t_{y_i}$ its observed version. We can determine $a_{a_{F_0}}$ and a_β by mapping the average of observed distances to $\mathrm{E}(X)$ and finding solutions to the approximation $P(T_O) \approx 0.999$, where T_O is an upper limit of a presumptive offspring window, discussed in Section A.2 (for a full description, refer to Section A.3.1). In many cases, limited information about the branching structure y is provided, to the extent of simply classifying points into immigrant and offspring groups, or no information is provided at all. Then, the parent t_y could be replaced with the closest immigrant predecessor of t or just the closest predecessor if no branching structure is available. For those who are concerned about the lack of universal choice for t_y and thus for $a_{a_{F_0}}$ and a_β, we will demonstrate in Section C.2 that the proposed models are robust to the choice of the hyperparameters in estimating the offspring density function.

Other key parameters are α for DP and ζ for GW. They modulate the discreteness of the distribution F and its variability around the centering function F_0. Moreover, under the modeling framework, they contribute to the smoothness of the target function $f(t|F)$, which is a stepwise function driven by uniform density kernels of the models. For example, a small α leads to mixture weights being concentrated on a few components rather than (almost) evenly distributed across all components. Compared to a large α, this results in relatively rough target densities. For α of DP, we empirically choose a gamma prior with effective support of 20 (e.g., a gamma prior with a 95 percentile of 20), which provides a good balance between model flexibility and smoothness in estimating continuous underlying functions. Likewise, in GW, a high value of $(1-\zeta)$ will result in a rough target function since the first few weights have almost all the probability, while a low value will reduce model flexibility due to relatively small discreteness and variability. Considering the trade-off between flexibility and smoothness, we specify a beta prior for ζ such that the prior density is centered at 0.5 and gets smaller towards both ends.

Finally, the truncation level L of the uniform mixture models is determined, considering the prior expectation of the partial sum of weights: $\mathrm{E}(\sum_{l=1}^{L}\omega_l|\alpha) = 1-\{\alpha/(\alpha+1)\}^L$ for DP

and $\mathrm{E}(\sum_{l=1}^{L}\omega_l|\zeta) = 1-\zeta^L$ for GW, where the expectation of the partial sum marginalized over α or ζ is approximately 1 under the selected L.

A.3.1 Mean and cumulative distribution function of expected offspring density

Below is a description of the cumulative distribution function $P(x)$ and mean $\mathrm{E}(X)$ derived from the expected offspring density, denoted as $g^*(x) = \mathrm{E}(g(x)) = \int \theta^{-1}\mathbf{1}_{[0,\theta)}(x)\mathrm{IG}(\theta|a_{F_0},\beta)d\theta$.

$$\begin{aligned}
P(x) &= \int_0^x g^*(u)du = \int_{\mathbb{R}^+}\Big(\int_0^x \theta^{-1}\mathbf{1}_{[0,\theta)}(u)du\Big)\mathrm{IG}(\theta|a_{F_0},\beta)d\theta \\
&= \int_x^\infty x\theta^{-1}\mathrm{IG}(\theta|a_{F_0},\beta)d\theta + \int_0^x \mathrm{IG}(\theta|a_{F_0},\beta)d\theta \\
&= \Big[\big(x\Gamma(a_{F_0}+1)\big)/\big(\beta\Gamma(a_{F_0})\big)\int_x^\infty \mathrm{IG}(\theta|a_{F_0}+1,\beta)d\theta\Big] + \int_0^x \mathrm{IG}(\theta|a_{F_0},\beta)d\theta.
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(X) &= \int x\Big(\int \theta^{-1}\mathbf{1}_{[0,\theta)}(x)\mathrm{IG}(\theta|a_{F_0},\beta)d\theta\Big)dx = \int\Big(\int x\theta^{-1}\mathbf{1}_{[0,\theta)}(x)dx\Big)\mathrm{IG}(\theta|a_{F_0},\beta)d\theta \\
&= \int (\theta/2)\beta^{a_{F_0}}/\Gamma(a_{F_0})\theta^{-a_{F_0}-1}\exp\{-\beta/\theta\}d\theta = \beta^{a_{F_0}}/\Gamma(a_{F_0})/2\int \theta^{-(a_{F_0}-1)-1}\exp\{-\beta/\theta\}d\theta \\
&= \beta^{a_{F_0}}/\Gamma(a_{F_0})/2\times\Gamma(a_{F_0}-1)/\beta^{(a_{F_0}-1)} = \big(\beta\Gamma(a_{F_0}-1)\big)/\big(2\Gamma(a_{F_0})\big).
\end{aligned}$$

Suppose that m is the average of observed distances x_i and T_O is the upper bound of a presumptive offspring window $(0, T_O)$. Substituting m for $\mathrm{E}(X)$, we can obtain $\beta = m2\Gamma(a_{F_0})/\Gamma(a_{F_0}-1)$ by solving the equation. With this β, we can determine a_{F_0} by finding a value that fulfills the approximation $P(T_O)\approx 0.999$. Denote these values by β^* and $a^*_{F_0}$. Our model is built using random a_{F_0} and β with $\mathrm{Exp}(a_{a_{F_0}})$ and $\mathrm{Exp}(a_\beta)$ priors, and we specify the hyperpriors by matching the values to the prior means, that is, $a^*_{F_0} = 1/a_{a_{F_0}}$ and $\beta^* = 1/a_\beta$.

B. MCMC details

In this section, we present MCMC posterior inference details for the models proposed in Section 2 of the main paper: the immigrant Erlang mixture model in Section B.1; the offspring Erlang mixture model in Section B.2; uniform-mixture-based models in Section B.3. Without loss of generality, we assume constant immigrant intensity in the offspring Erlang mixture model and the uniform-mixture-based models. Recall that, given the branching

structure, we can factorize the likelihood into the immigrant and offspring parts, as shown in (2). Therefore, the parameters of an immigrant intensity model and an offspring intensity/density model can be estimated separately regardless of their functional forms (unless the two models share common parameters). Inference details for the fully nonparametric model (Section 2.3) are omitted; however, since it is a concatenation of two semiparametric models, we can apply the inference methods described below for the semiparametric models to the fully nonparametric model.

B.1 Posterior simulation for the immigrant Erlang mixture model

This section focuses on the MCMC details for the parameters of an Erlang mixture for immigrant intensity. The model parameters for the excitation function (offspring intensity) can be treated separately using the factorization of the augmented HP likelihood, and, thus, the posterior inference details will vary based on the model.

As a result of the HP cluster representation, we can leverage the inference method for the NHPP model introduced in Kim and Kottas (2022) to estimate the immigrant intensity of an Erlang mixture. The Gibbs sampler underlies the posterior inference method, where the hyperparameters c_0 and b_{G_0} are updated with the Metropolis-Hastings (M-H) algorithm. The scale parameter, θ, of basis functions can also be updated using the M-H algorithm, or, to achieve better mixing in the posterior distribution, the Hamiltonian Monte Carlo (HMC) algorithm can be used (described later in this section).

The key feature of the MCMC method for this model is the closed-form posterior full conditionals for the mixture weights ν_l's. With the $\text{Ga}(c_0\theta/b_{G_0}, c_0)$ prior, induced by the gamma process prior on G, the posterior full conditional distribution of ν_l for $l = 1, \ldots, L$ is derived as follows,

$$\begin{aligned} p(\nu_l|\theta, c_0, b_{G_0}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{t}) &\propto \exp\Big(-\nu_l \int_0^T \text{Ga}(u|l, \theta^{-1})du\Big)\Big[\prod_{\{i:t_i\in I\}} \nu_{\xi_i}\Big] \exp\Big(-\nu_l c_0\Big)\nu_l^{c_0\theta/b_{G_0}-1} \\ &\propto \exp\Big(-\nu_l\Big(\int_0^T \text{Ga}(u|l, \theta^{-1})du + c_0\Big)\Big)\nu_l^{\left(\sum_{\{i:t_i\in I\}}\delta_l(\xi_i)\right)+c_0\theta/b_{G_0}-1}, \end{aligned}$$

which is proportional to a $\text{Ga}(n_l + c_0\theta/b_{G_0}, \int_0^T \text{Ga}(u|l, \theta^{-1})du + c_0)$ density, where $n_l = \sum_{\{i:t_i\in I\}} \delta_l(\xi_i) = |\{\xi_i = l : t_i \in I\}|$. $\delta_s(\cdot)$ is a delta function that has a point mass at s. As such, we can draw the posterior sample of ν_l from the updated gamma distribution.

With a discrete prior distribution with probability $\Pr(\xi_i = l) = \nu_l / \sum_{k=1}^{L} \nu_k$ for an immigrant point $t_i \in I$, the auxiliary variable has the posterior full conditional,

$$\Pr(\xi_i = l | \boldsymbol{\nu}, \theta, \boldsymbol{y}, \boldsymbol{t}) \propto \nu_l \mathrm{Ga}(t_i | l, \theta^{-1}), \;\; l = 1, \ldots, L.$$

Accordingly, the posterior sample of ξ_i can be drawn from a discrete distribution with the quantities for probabilities.

The branching structure, denoted by y_i for any point observation t_i in the immigrant process I, is another auxiliary variable, used for the augmented HP likelihood. We assigned to y_i the prior of $\delta_0(y_1) \prod_{i=2}^{n} \mathrm{Unif}(y_i \mid 0, 1, \ldots, i-1)$, where $\mathrm{Unif}(\cdot \mid 0, 1, \ldots, i-1)$ is a discrete uniform distribution on $\{0, 1, \ldots, i-1\}$. Then, the posterior full conditional for the auxiliary variable is given by

$$\Pr(y_i = k | \boldsymbol{\nu}, \theta, \boldsymbol{\alpha}, \boldsymbol{t}) = \begin{cases} \dfrac{\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t_i | l, \theta^{-1})}{\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t_i | l, \theta^{-1}) + \sum_{r=1}^{i-1} h(t_i - t_r)}, & k = 0; \\ \\ \dfrac{h(t_i - t_k | \boldsymbol{\alpha})}{\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t_i | l, \theta^{-1}) + \sum_{r=1}^{i-1} h(t_i - t_r | \boldsymbol{\alpha})}, & k = 1, \ldots, i-1. \end{cases}$$

Hamiltonian Monte Carlo (HMC) suppresses the local random walk behavior in the Metropolis algorithm using an additional parameter called a momentum variable φ, thus allowing it to move much more rapidly through the target distribution (in our case, the posterior distribution of θ). θ and φ are then updated together in a new Metropolis algorithm, in which φ largely determines the jumping distribution for θ. To update them, we use the leapfrog method; the two HMC parameters, the step size ϵ and the number of steps E, are randomized for more flexibility in the algorithm, such that $\epsilon \sim \mathrm{Unif}(0, 2\epsilon_0)$ and $E \sim$ discrete $\mathrm{Unif}(1, 2E_0)$ with $\epsilon_0 E_0 = 1$ (detailed in Gelman et al. 2013). Our choice of ϵ_0 and E_0 was based on ensuring that the acceptance probability of θ was close to the optimal value of 0.65. For brevity, we use the $\mathrm{N}(0, 1)$ prior for φ. The gradient of the

log-posterior density of θ, used to make a half-step of φ in the algorithm, is derived as

$$\begin{aligned}\frac{d\log p(\theta|c_0, b_{b_{G_0}}\boldsymbol{\nu}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{t})}{d\theta} &= d\log\Big\{\Big[\prod_{\{i:t_i\in I\}}\mathrm{Ga}(t_i|\xi_i,\theta^{-1})\Big]\exp\big(-\sum_{l=1}^{L}\nu_l\int_0^T\mathrm{Ga}(u|l,\theta^{-1})du\big)\\ &\quad\times\Big[\prod_{l=1}^{L}\mathrm{Ga}(\nu_l|c_0\theta/b_{G_0},c_0)\Big]\mathrm{Lo}(\theta|2,d_\theta)\Big\}\Big/d\theta\\ &= \sum_{\{i:t_i\in I\}}\Big(-\xi_i\theta^{-1}+t_i\theta^{-2}\Big)-\sum_{l=1}^{L}\nu_l\Big(-(T\theta^{-1})^l\theta^{-1}\exp(-T\theta^{-1})/\Gamma(l)\Big)\\ &\quad+\sum_{l=1}^{L}c_0/b_{G_0}\Big(\log(c_0)-\mathrm{digamma}(c_0\theta/b_{G_0})+\log(\nu_l)\Big)-(2+1)/(d_\theta+\theta),\end{aligned}$$

where $\mathrm{Lo}(\theta|a,b)$ indicates a Lomax density with shape a and scale b. $\Gamma(\cdot)$ is the gamma function, and digamma$(\cdot)$ is the derivative of the (natural) logarithm of the gamma function.

B.2 Posterior simulation for the offspring Erlang mixture model

The inference method for the immigrant Erlang mixture model described in the previous section can be applied to this offspring model with a few modifications. As with b_{G_0} of the immigrant Erlang mixture, hyperparameter b_{F_0} can be sampled using the M-H algorithm. Posterior sampling of θ can be undertaken using the M-H algorithm or the HMC algorithm; the HMC algorithm requires a slight adjustment to the gradient of the log posterior full conditional, while the approach to specifying the HMC parameters, $(\epsilon, E, \epsilon_0, E_0)$, is still applicable. Auxiliary variable ξ_i has as its full conditional distribution a discrete distribution with probabilities

$$\Pr(\xi_i = l|\boldsymbol{\nu},\theta,\boldsymbol{y},\boldsymbol{t}) \propto \nu_l\mathrm{Ga}(t_i|l,\theta^{-1}) \text{ for } l=1,\ldots,L \text{ and } \{i:t_i\in O\}.$$

The posterior full conditional for the branching structure, with the prior of a delta function for y_1 and discrete uniform distributions for $y_2,\ldots,y_n$, is given by

$$\Pr(y_i=k|\mu,\boldsymbol{\nu},\theta,\boldsymbol{t}) = \begin{cases}\dfrac{\mu}{\mu+\sum_{r=1}^{i-1}\sum_{l=1}^{L}\nu_l\mathrm{Ga}(t_i-t_r|l,\theta^{-1})}, & k=0;\\[2ex] \dfrac{\sum_{l=1}^{L}\nu_l\mathrm{Ga}(t_i-t_k|l,\theta^{-1})}{\mu+\sum_{r=1}^{i-1}\sum_{l=1}^{L}\nu_l\mathrm{Ga}(t_i-t_r|l,\theta^{-1})}, & k=1,\ldots,i-1.\end{cases}$$

In our model with the $\mathrm{Ga}(A_l,\eta)$ mixture weights, the posterior full conditional for ν_l,

$l = 1, \ldots, L$, is given in closed-form as follows

$$p(\nu_l|\theta, b_{F_0}, \boldsymbol{\xi}, \boldsymbol{y}, \boldsymbol{t}) \propto \exp\Big(-\sum_{t_i \in \boldsymbol{t}} \nu_l \int_0^{T-t_i} \mathrm{Ga}(u|l, \theta^{-1}) du\Big) \Big[\prod_{\{i:t_i \in O\}} \nu_{\xi_i}\Big] \exp\Big(-\nu_l \eta\Big) \nu_l^{A_l - 1}$$
$$\propto \exp\Big\{-\nu_l \Big[\Big(\sum_{t_i \in \boldsymbol{t}} \int_0^{T-t_i} \mathrm{Ga}(u|l, \theta^{-1}) du\Big) + \eta\Big]\Big\} \nu_l^{\left(\sum_{\{i:t_i \in O\}} \delta_l(\xi_i)\right) + A_l - 1},$$

which is proportional to a $\mathrm{Ga}(n_l + A_l, \big(\sum_{t_i \in \boldsymbol{t}} \int_0^{T-t_i} \mathrm{Ga}(u|l, \theta^{-1}) du\big) + \eta)$ density, where $n_l = \sum_{\{i:t_i \in O\}} \delta_l(\xi_i) = |\{\xi_i = l : t_i \in O\}|$. Therefore, ν_l can be updated using the gamma posterior full conditional distribution.

Lastly, we assign the $\mathrm{Exp}(a_\mu)$ prior to the constant immigrant intensity, μ, which yields $\mathrm{Ga}(n_I + 1, T + a_\mu)$ as its posterior full conditional. T is the upper limit of an observation window $(0, T)$. n_I indicates the immigrant cluster size, defined in Section 3.2.

B.3 Posterior simulation for uniform-mixture-based models

Described in this section is the inference method for a semiparametric model consisting of constant immigrant intensity, branching ratio, and offspring density of uniform DP mixtures. As in Section B.2, the constant immigrant intensity, with the $\mathrm{Exp}(a_\mu)$ prior, has a closed-form posterior full conditional, $\mathrm{Ga}(n_I + 1, T + a_\mu)$. For the branching ratio, with a conjugate $\mathrm{Ga}(a_\gamma, b_\gamma)$ prior, the posterior full conditional is derived as

$$p(\gamma|\boldsymbol{\omega}, \boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{t}) \propto \gamma^{n_O} \exp\Big(-\sum_{l=1}^{L} \gamma \omega_l K(Z_l)\Big) \gamma^{a_\gamma - 1} \exp\big(-\gamma b_\gamma\big)$$
$$\propto \gamma^{n_O + a_\gamma - 1} \exp\Big\{-\gamma\Big[\sum_{l=1}^{L} \gamma \omega_l K(Z_l) + b_\gamma\Big]\Big\}.$$

Therefore, we can update γ using $\mathrm{Ga}(n_O + a_\gamma, \sum_{l=1}^{L} \omega_l K(Z_l) + b_\gamma)$, where n_O denotes the offspring cluster size, defined in Section 3.2. The parameters of uniform DP mixtures for offspring density are inferred using the blocked Gibbs sampler. Firstly, given the $\mathrm{IG}(a_{F_0}, \beta)$ prior, the scale parameter Z_l of uniform mixtures has a piecewise truncated inverse gamma distribution as its full conditional. Let ξ_s^* for $s = 1, \ldots, L^*$ with $L^* \leq L$ be the distinct values of auxiliary variables $\{\xi_i : t_i \in O\}$. Denoted by "$-$" a set of variables $\{\boldsymbol{\omega}, \boldsymbol{\xi}, \gamma, \beta, \boldsymbol{y}, \boldsymbol{t}\}$, the piecewise truncated inverse gamma distribution is defined as

- If $l \notin \{\xi_1^*, \ldots, \xi_{L^*}^*\}$,

$$Z_l|- \overset{ind.}{\sim} \begin{cases} \mathrm{IG}\Big(a_{F_0}, \gamma\omega_l(\sum_{j=1}^{n}(T-t_j)) + \beta\Big), \;\; Z_l \in (T-t_1, \infty), \;\; \text{w/ prob. } \frac{c_0}{\sum_{k=0}^{n} c_k}; \\ \mathrm{IG}\Big(a_{F_0}, \gamma\omega_l(\sum_{j=r+1}^{n}(T-t_j)) + \beta\Big), \;\; Z_l \in (T-t_{r+1}, T-t_r], \;\; \text{w/ prob. } \frac{c_r}{\sum_{k=0}^{n} c_k}; \\ \mathrm{IG}\Big(a_{F_0}, \beta\Big), \;\; Z_l \in (0, T-t_n], \;\; \text{w/ prob. } \frac{c_n}{\sum_{k=0}^{n} c_k}; \end{cases}$$

$$c_0 = \frac{\beta^{a_{F_0}}}{\Gamma(a_{F_0})} \frac{\Gamma(a_{F_0})}{(\gamma\omega_l \sum_{k=1}^{n}(T-t_k) + \beta)^{a_{F_0}}} \Big[\int_{T-t_1}^{\infty} \mathrm{IG}\Big(s|a_{F_0}, \gamma\omega_l \sum_{k=1}^{n}(T-t_k) + \beta\Big) ds \Big],$$

$$c_r = \frac{\beta^{a_{F_0}}}{\Gamma(a_{F_0})} \frac{\Gamma(a_{F_0})}{(\gamma\omega_l \sum_{k=r+1}^{n}(T-t_k) + \beta)^{a_{F_0}}} \Big[\int_{T-t_{r+1}}^{T-t_r} \mathrm{IG}\Big(s|a_{F_0}, \gamma\omega_l \sum_{k=r+1}^{n}(T-t_k) + \beta\Big) ds \Big]$$

$$\times \exp(-\gamma\omega_l r), \quad r = 1, \ldots, n-1,$$

$$c_n = \frac{\beta^{a_{F_0}}}{\Gamma(a_{F_0})} \frac{\Gamma(a_{F_0})}{\beta^{a_{F_0}}} \Big[\int_{0}^{T-t_n} \mathrm{IG}\Big(s|a_{F_0}, \beta\Big) ds \Big] \exp(-\gamma\omega_l n).$$

- If $l \in \{\xi_1^*, \ldots, \xi_{L^*}^*\}$,

$$Z_l|- \overset{ind.}{\sim} \begin{cases} \mathrm{IG}\Big(a_{F_0} + n_l, \gamma\omega_l(\sum_{j=1}^{n}(T-t_j)) + \beta\Big), \;\; Z_l \in (T-t_1, \infty), \;\; \text{w/ prob. } \frac{c_0}{\sum_{k=0}^{n} c_k}; \\ \mathrm{IG}\Big(a_{F_0} + n_l, \gamma\omega_l(\sum_{j=r+1}^{n}(T-t_j)) + \beta\Big), \;\; Z_l \in (T-t_{r+1}, T-t_r], \;\; \text{w/ prob. } \frac{c_r}{\sum_{k=0}^{n} c_k}; \\ \mathrm{IG}\Big(a_{F_0} + n_l, \beta\Big), \;\; Z_l \in (0, T-t_n], \;\; \text{w/ prob. } \frac{c_n}{\sum_{k=0}^{n} c_k}; \end{cases}$$

$$c_0 = \frac{\beta^{a_{F_0}}}{\Gamma(a_{F_0})} \frac{\Gamma(a_{F_0}+n_l)}{(\gamma\omega_l \sum_{k=1}^{n}(T-t_k) + \beta)^{a_{F_0}+n_l}} \Big[\int_{b_0}^{\infty} \mathrm{IG}\Big(s|a_{F_0}+n_l, \gamma\omega_l \sum_{k=1}^{n}(T-t_k) + \beta\Big) ds \Big],$$

$$c_r = \frac{\beta^{a_{F_0}}}{\Gamma(a_{F_0})} \frac{\Gamma(a_{F_0}+n_l)}{(\gamma\omega_l \sum_{k=r+1}^{n}(T-t_k) + \beta)^{a_{F_0}+n_l}} \Big[\int_{b_r}^{T-t_r} \mathrm{IG}\Big(s|a_{F_0}+n_l, \gamma\omega_l \sum_{k=r+1}^{n}(T-t_k) + \beta\Big) ds \Big]$$

$$\times \exp(-\gamma\omega_l r), \quad r = 1, \ldots, n-1,$$

$$c_n = \frac{\beta^{a_{F_0}}}{\Gamma(a_{F_0})} \frac{\Gamma(a_{F_0}+n_l)}{\beta^{a_{F_0}+n_l}} \Big[\int_{b_n}^{T-t_n} \mathrm{IG}\Big(s|a_{F_0}+n_l, \beta\Big) ds \Big] \exp(-\gamma\omega_l n),$$

where $n_l = |\{\xi_j : \xi_j = l, \;\; t_j \in O\}|$, $b_0 = T - t_1$, $b_r = \min(T - t_r, \max(T - t_{r+1}, \max(t_j - t_{y_j})))$, and $b_n = \min(T - t_n, \max(t_j - t_{y_j}))$.

The $\mathrm{Exp}(a_\beta)$ prior for β provides ready prior-to-posterior updating with the posterior full conditional distribution $\mathrm{Ga}(a_{F_0}L + 1, \sum_{l=1}^{L} Z_l^{-1} + a_\beta)$. The posterior sample of a_{F_0} is drawn using the M-H algorithm with a log-normal proposal distribution and the posterior full conditional of $p(a_{F_0}|\beta, \boldsymbol{Z}) \propto \Big[\prod_{l=1}^{L} \beta^{a_{F_0}}/\Gamma(a_{F_0}) Z_l^{a_{F_0}}\Big] \exp(-a_{F_0} a_{a_{F_0}})$. The posterior full conditional for the auxiliary variable ξ_i is represented as a discrete distribution with

probabilities $\Pr(\xi_i = l|\boldsymbol{\omega}, \boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{t}) \propto \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_{y_i})$ for $l = 1, \ldots, L$, where $1_{(a,b)}(x)$ denotes an indicator function: 1 if $x \in (a, b)$ or 0 otherwise.

Mixture weights are updated using the stick-breaking method with the posterior samples of ν_l, $l = 1, \ldots, L-1$, for DP / ζ for GW. Under the DP prior model, the posterior full conditionals for ν_l's are derived as

$$p(\boldsymbol{\nu}|\boldsymbol{\xi}, \boldsymbol{Z}, \gamma, \alpha) \propto \exp\Big\{ -\gamma\Big(\nu_1 K(Z_1) + \sum_{l=2}^{L-1} \nu_l \big(\prod_{r=1}^{l-1}(1-\nu_r)\big) K(Z_l) + \big(\prod_{r=1}^{L-1}(1-\nu_r)\big) K(Z_L)\Big)\Big\}$$
$$\times \nu_1^{n_1} \Big[\prod_{l=2}^{L-1} \Big(\nu_l (\prod_{r=1}^{l-1}(1-\nu_r))\Big)^{n_l}\Big] \Big(\prod_{r=1}^{L-1}(1-\nu_r)\Big)^{n_L} \prod_{l=1}^{L-1}(1-\nu_l)^{\alpha-1},$$

We draw the posterior sample of ν_l from the distribution using the slice sampling method. Using an auxiliary variable for the exponential term in the method allows us to sample ν_l from a truncated beta distribution. Under the GW prior model, the posterior full conditional for ζ is derived as

$$p(\zeta|\boldsymbol{\xi}, \boldsymbol{Z}, \gamma) \propto \exp\Big\{-\gamma\Big(\sum_{l=1}^{L-1} K(Z_l)(1-\zeta)\zeta^{l-1} + K(Z_L)\zeta^{L-1}\Big)\Big\} \zeta^{\sum_{l=1}^{L}(l-1)n_l + a_\zeta - 1} (1-\zeta)^{\sum_{l=1}^{L-1} n_l + b_\zeta - 1}.$$

We use the M-H algorithm to sample ζ with a beta proposal distribution $\mathrm{Be}(\rho(\sum_{l=1}^{L}(l-1)n_l + a_\zeta), \rho(\sum_{l=1}^{L-1} n_l + b_\zeta))$, which is constructed by using part of the posterior full conditional with a tuning parameter ρ.

As a conjugate prior, $\mathrm{Ga}(a_\alpha, b_\alpha)$ results in a gamma posterior full conditional for the precision parameter α of DP. With the generalized Dirichlet distribution for $p(\boldsymbol{\omega}|\alpha)$, induced by the stick-breaking mechanism with $\nu_l \sim \mathrm{Be}(1, \alpha)$, the gamma posterior full conditional is given by

$$p(\alpha|\omega_L) \propto \alpha^{L-1} \omega_L^\alpha \alpha^{a_\alpha - 1} \exp(-\alpha b_\alpha) = \alpha^{a_\alpha + L - 2} \exp\Big(-\alpha\big(b_\alpha - \log(\omega_L)\big)\Big).$$

With $\log(\omega_L) = \log \prod_{r=1}^{L-1}(1-\nu_r)$, α is updated via $\mathrm{Ga}\Big(a_\alpha + L - 1, b_\alpha - \sum_{r=1}^{L-1} \log(1-\nu_r)\Big)$.

Finally, given the prior of $\delta_0(y_1) \prod_{i=2}^{n} \mathrm{Unif}(y_i \mid 0, 1, \ldots, i-1)$, the branching structure is updated using the following discrete distribution,

$$\Pr(y_i = k|\mu, \boldsymbol{\omega}, \boldsymbol{Z}, \gamma, \boldsymbol{t}) = \begin{cases} \dfrac{\mu}{\mu + \gamma \sum_{r=1}^{i-1} \sum_{l=1}^{L} \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_r)}, & k = 0; \\ \\ \dfrac{\gamma \sum_{l=1}^{L} \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_k)}{\mu + \gamma \sum_{r=1}^{i-1} \sum_{l=1}^{L} \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_r)}, & k = 1, \ldots, i-1. \end{cases}$$

C. Sensitivity analyses

C.1 Erlang mixture for the excitation function: choice of T_O

In the semiparametric model employing an Erlang mixture for the excitation function, a selection of the interval $(0, T_O)$ that will serve as a proxy for the effective support of the modeled excitation function must be undertaken prior to the subsequent specification of θ and b_{F_0}. But, in real applications, there might be little information available about the distance $t - t_i$ and its presumptive upper bound T_O. So, we designed the following sensitivity analysis to see the effect of T_O choice on posterior inference. The analysis is based on the Weibull and Weibull mixture examples provided in Section D.1.2, but with a different choice of $(0, T_O)$; we have plugged the observation window of $(0, 10000)$ into the interval, which is a conservative choice and is much larger than the pre-specified favorable intervals, $(0, 6)$ and $(0, 15)$ for each example.

The posterior estimates (summarized in Table S1) with the extreme choice of T_O are comparable to those of the simulation studies with favorable bounds given in Table 4. According to Figure S1, the models capture well the underlying densities, incorporating them within their uncertainty bounds, even though the prior models fail to capture the underlying functions due to the extreme value of T_O, as opposed to the simulation studies shown in Figure S4. These results support the claim that this semiparametric modeling approach is robust to T_O choice.

Example	μ	γ	n_I	n_O	M_I	M_O	R
Weibull	0.011(0.001)	0.8(0.04)	105(4)	429(4)	8(3)	7(2)	0.028(0.006)
Weibull mixture	0.011(0.001)	0.787(0.041)	105(5)	395(5)	20(5)	16(2)	0.071(0.010)

Table S1: Posterior means and standard deviations of (μ,γ), cluster sizes, misclassifications, and TVDs in each example.

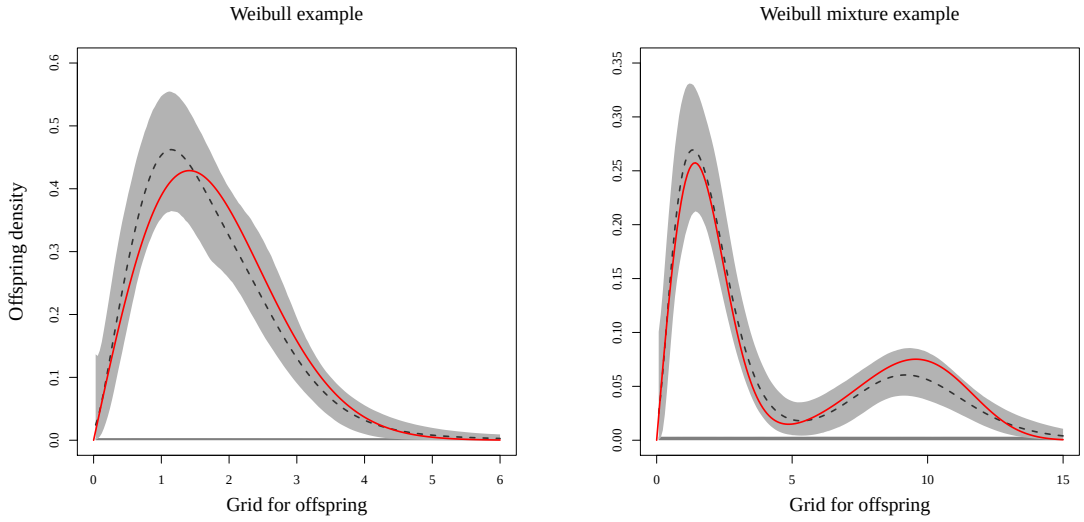

Figure S1: Posterior means (dashed) and 95% uncertainty bounds (light gray) of the offspring density functions, with the true densities (red). The use of the observation window made prior interval estimates (dark gray) failed to cover the true functions, leaving them at the bottom of the panels.

C.2 Uniform mixture-based models: prior choices for α and (a_{F_0}, β)

The following is a summary of the sensitivity analysis results under the uniform mixture-based models. In particular, we examined the sensitivity of the uniform DP mixture to the prior for α. Since the branching structure of the underlying HP intensity may be unknown in real application, we also performed sensitivity analysis to determine if the models are sensitive to the priors for a_{F_0} and β, which require such information for prior specification. These analyses will continue to use the observed point patterns, the Weibull and the power-law examples. The following results are based on the uniform DP mixture semiparametric model, but the GW mixture-based model also produced similar results in a_{F_0} and β sensitivity analysis.

Table S2 and Figure S2 illustrate the implications of prior choices for α. As shown in the table, such a choice does not impact the estimation of quantitative standards from model parameters μ and γ to TVD's. On the other hand, the prior choice affects the posterior estimated functions in terms of their uncertainty and smoothness. Note that α governs the flexibility and smoothness of the prior model: the larger α, the less flexible but smoother the prior model. Compared to our recommended choice, $\text{Ga}(4, 0.5)$ with mean 8, the other prior, $\text{Ga}(2, 1.5)$ with mean $2/1.5$ has a larger posterior uncertainty with rougher estimated functions due to α estimate's sensitivity to its prior.

Table S3 and Figure S3 represent the robustness of the models to the prior choices for parameters of the inverse gamma centering distribution. A prior that failed to capture the underlying density function was considered and compared to our recommended choice of prior. The posterior estimated functions retrieve the underlying density function with uncertainty bounds that fully cover the density regardless of the priors. As can be seen in the table, the quantitative results also support the robustness, with no significant differences between the estimates from each of the priors.

Prior for α	μ	γ	n_I	n_O	M_I
$\text{Ga}(2, 1.5)$	0.009(0.001)	0.799(0.043)	93(2)	377(2)	7(2)
$\text{Ga}(4, 0.5)$	0.009(0.001)	0.798(0.042)	92(2)	378(2)	7(2)
	M_O	R	TVD[b]	TVD[c]	
$\text{Ga}(2, 1.5)$	6(1)	0.028(0.005)	0.062(0.020)	0.065(0.020)	
$\text{Ga}(4, 0.5)$	6(1)	0.028(0.005)	0.061(0.020)	0.066(0.020)	

Table S2: Sensitivity to the prior for α. Weibull example. Posterior means and standard deviations of (μ,γ), cluster sizes, misclassifications, and TVDs for the two different prior choices for α.

Prior for (a_{F_0},β)	μ	γ	n_I	n_O	M_I
Exp(0.333)Exp(0.284)	0.021(0.002)	0.806(0.040)	103(6)	431(6)	11(3)
Exp(0.125)Exp(0.00357)	0.021(0.002)	0.803(0.039)	102(6)	432(6)	11(3)
	M_O	R	TVD[b]	TVD[c]	
Exp(0.333)Exp(0.284)	13(4)	0.045(0.009)	0.059(0.022)	0.055(0.018)	
Exp(0.125)Exp(0.00357)	13(4)	0.045(0.009)	0.057(0.021)	0.053(0.017)	

Table S3: Sensitivity to the prior for (a_{F_0},β). Power-law example. Posterior means and standard deviations of (μ,γ), cluster sizes, misclassifications, and TVDs for the two different prior choices for a_{F_0} and β under the uniform DP mixture.

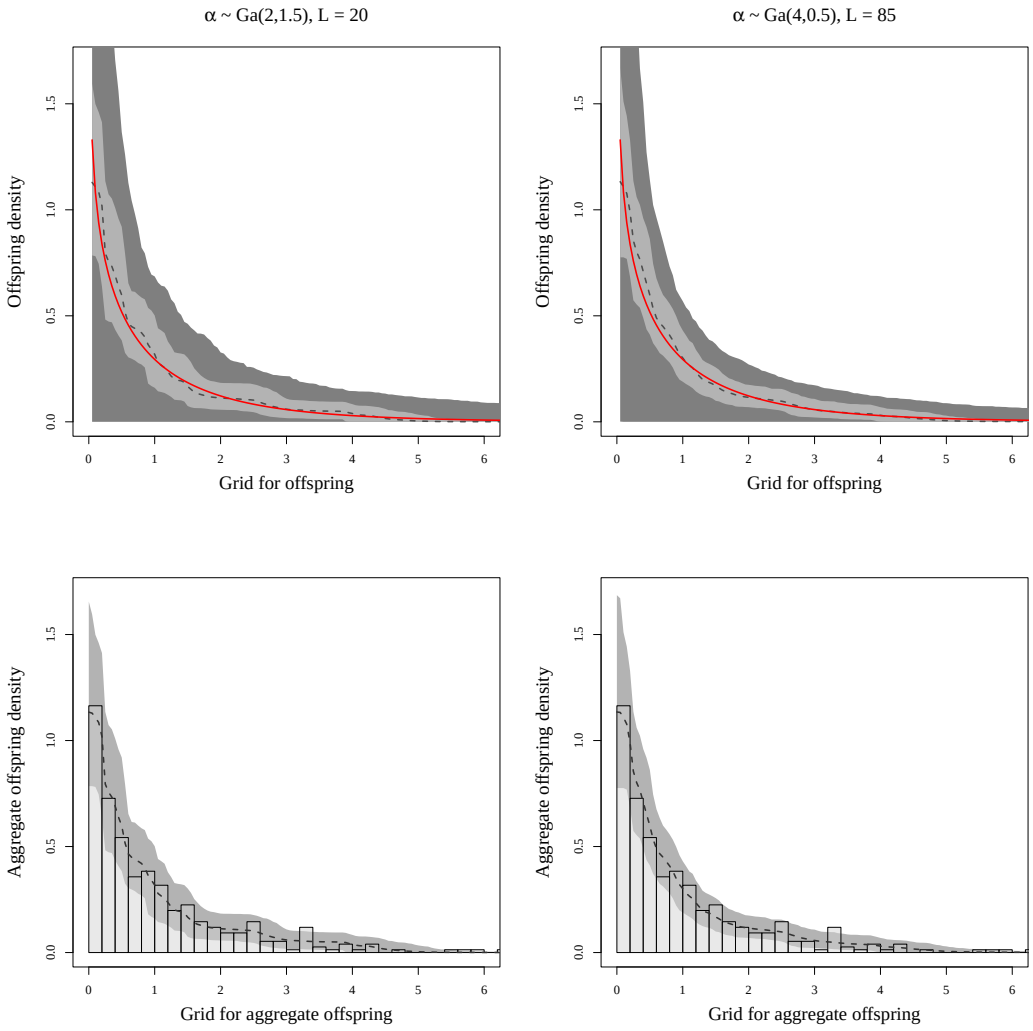

Figure S2: Sensitivity to the prior for α: Ga$(2, 1.5)$ (left) and Ga$(4, 0.5)$ (right). Weibull example. Posterior means (dashed) and 95% uncertainty bounds (light gray) of the offspring density (top) and aggregate offspring density (bottom) functions, as well as 95% prior uncertainty bounds (dark gray). The red lines indicate the underlying function, and the histograms represent the distribution of distances between offspring and their parents.

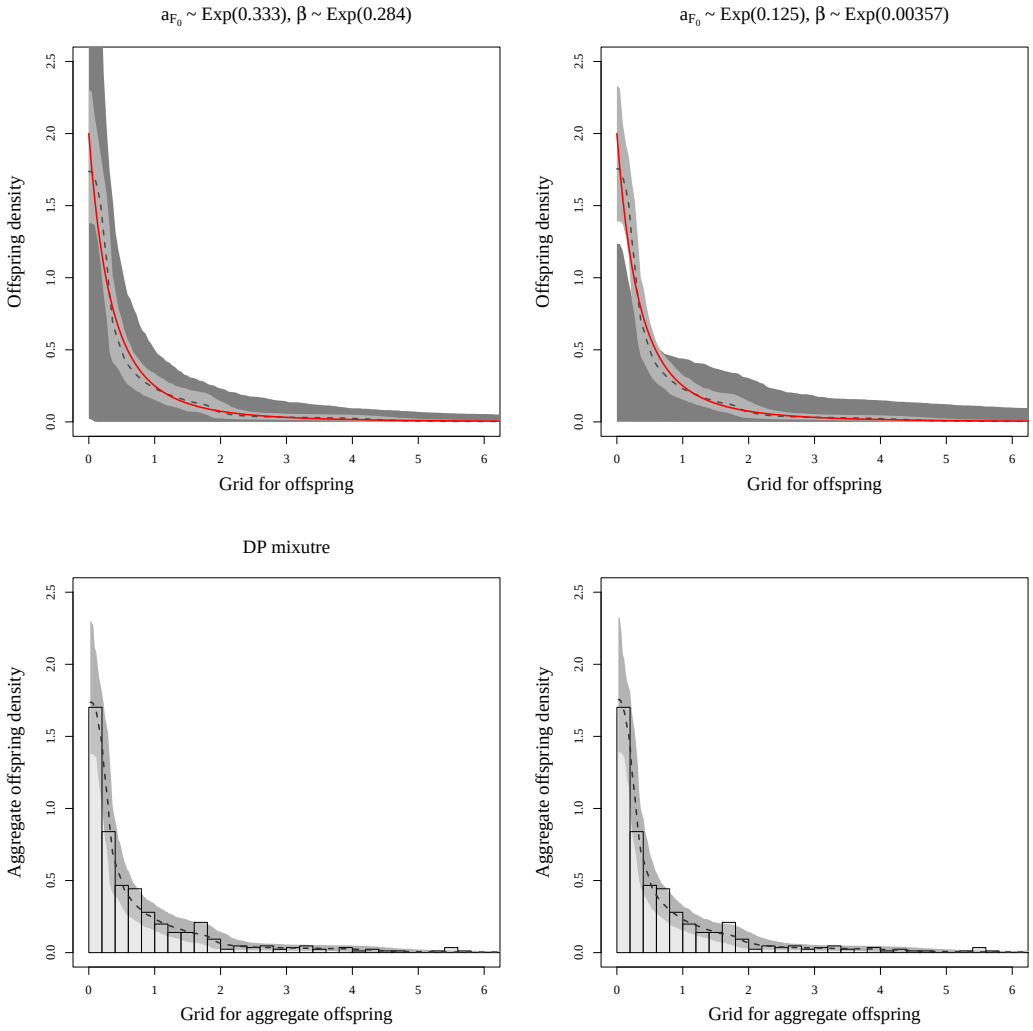

Figure S3: Sensitivity to the prior for (a_{F_0},β): Exp(0.25)Exp(0.167) (left) and Exp(0.125)Exp(0.00357) (right). Power-law example. Posterior means (dashed) and 95% uncertainty bounds (light gray) of the offspring density (top) and aggregate offspring density (bottom) functions, as well as 95% prior uncertainty bounds (dark gray). The red lines indicate the underlying density function, and the histograms represent the distribution of distances between offspring and their parents.

D. Simulation study

This section examines the performance of our models by comparing estimated intensity/density with the underlying functions used for simulations. Sections D.1.1 and D.1.2 discuss the results of simulations for the semiparametric models that are based on Erlang mixtures for modeling immigrant intensity and excitation function, respectively. In Section D.1.3, the nonparametric model is analyzed with simulations in a variety of settings, and its results are compared with those of the semiparametric models. Section D.2 demonstrates decreasing offspring density examples for the uniform-mixture-based semiparametric models described in Section 2.2.2 of the paper.

D.1 Erlang-mixture-based models

D.1.1 Semiparametric model with immigrant Erlang mixture

We generated 504 points (397 for immigrants and 107 for offspring) from a HP with intensity $\lambda^*(t) = 400[0.6\text{We}(t|1.5, 2000) + 0.4\text{We}(t|7, 8000)] + 0.2\sum_{t_i<t}\text{Exp}(t - t_i|2)$, $t \in (0, 10000)$. Background intensity, which consists of a mixture of two Weibull densities, results in an irregular shape; this is intended to demonstrate the flexibility of the semiparametric model with the Erlang mixture for immigrant intensity.

For this study, we applied an exponential density function to the offspring density of the semiparametric model. That is, $h(t - t_j) = \gamma\text{Exp}(t - t_j|\alpha)$. We are using the same type of density as in the underlying offspring density function, which enables the Erlang mixture of the model to estimate immigrant intensity without interference from confounding effects caused by misspecifying the offspring density.

Following the prior specification in Section A, we set the priors as follows: the $\text{Lo}(2, 500)$ prior for θ with $L = 80$; $\text{Exp}(0.1)$ for c_0; $\text{Exp}(0.0252)$ for b_{G_0}, $\text{Ga}(2, 4)$ for γ. For simplicity, the model parameter α of the offspring density is assigned the $\text{Exp}(1)$ prior.

These prior choices provide prior model uncertainty that covers the underlying immigrant intensity and even the entire panel of the figure for the intensity function (see left of Figure S4). In the following examples, we observed similar wide prior model uncertainties for semiparametric/nonparametric model with the Erlang mixture immigrant intensity and priors that complied with our prior specification strategy. We will no longer superimpose prior uncertainties over panels showing immigrant intensity estimates.

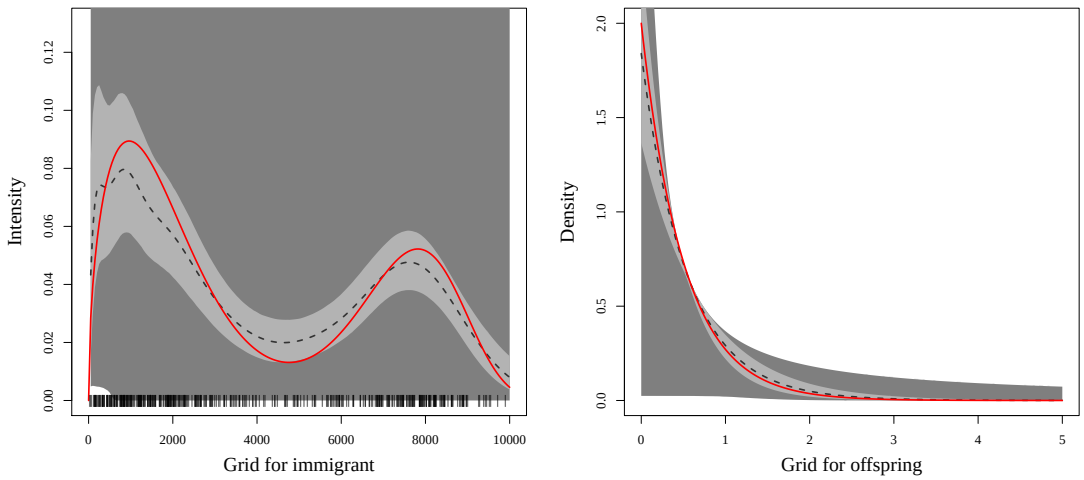

Figure S4: Weibull mixture immigrant and exponential offspring example. Semiparametric model with Erlang mixture immigrant intensity. The left panel displays the posterior mean (dashed) and the posterior 95% interval estimates (light gray) for the immigrant intensity function with: the prior 95% uncertainty bounds (dark gray); the underlying immigrant intensity (red solid); and the point pattern (bar) at the bottom. Similarly, the right panel demonstrates the prior uncertainty bounds (dark gray), the posterior mean (dashed), and the interval estimates (light gray) with the underlying offspring density function (red solid).

In Figure S4, the posterior mean of the immigrant intensity function retrieves the bimodality of the underlying intensity, with the interval estimates encompassing the true values of the underlying function. A particular focus of this section is on the flexibility of mixture modeling for immigrant intensity, with the offspring density function set to an exponential density function in accordance with the underlying function. As a result, the estimated offspring densities are nearly identical to the true functional values. The posterior means of the cluster sizes n_I and n_O are 392 and 112 (with the common posterior standard deviation of 6), which correspond to the true values of 397 and 107, respectively. The posterior means of the misclassification quantities (shown in Section 3.2 of the paper) are given by $M_I = 22$, $M_O = 27$, and $R = 0.098$. Branching ratio γ has the posterior mean of 0.225 and the posterior 95% interval estimates of (0.180,0.275), which contains true $\gamma = 0.2$. The inferences are made on the basis of $20,000$ posterior samples taken after discarding $20,000$ burn-in steps.

	μ	γ	Cluster size		Misclassification		
			n_I	n_O	M_I	M_O	R
Ex1	0.011(0.001)	0.798(0.039)	107(3)	427(3)	8(3)	6(1)	0.026(0.006)
Ex2	0.011(0.001)	0.786(0.041)	106(5)	394(5)	20(5)	15(1)	0.071(0.010)

Table S4: Weibull (top) and Weibull mixture (bottom) examples. Semiparametric model with Erlang mixture offspring intensity. Posterior means and standard deviations for parameters (μ, γ), cluster sizes, and misclassifications.

D.1.2 Semiparametric model with offspring Erlang mixture

We will study the performance of the semiparametric model incorporating an Erlang mixture as the excitation function using two synthetic data sets whose underlying offspring density functions are either the unimodal Weibull or the two-component Weibull mixture. These simulations emulate plausible scenarios in which secondary events are activated after some idle time from their parent events or are intensified twice at two different points in time.

The HP intensity function of the first example is given by

$$\lambda^*(t) = 0.01 + 0.8 \sum_{t_i<t} \mathrm{We}(t - t_i|2, 2), \quad t \in (0, 10000).$$

We sampled 534 time points (105 for immigrants and 429 for offspring) from the intensity function. The priors we used for model fitting are : Exp(37.5) for μ, Lo$(2, 0.3)$ for θ with $L = 50$, Exp(0.168) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

The second example has the intensity function defined as

$$\lambda^*(t) = 0.01 + 0.8 \sum_{t_i<t} [0.6\mathrm{We}(t - t_i|2, 2) + 0.4\mathrm{We}(t - t_i|5, 10)], \quad t \in (0, 10000),$$

from which we generated 500 points (101 for immigrants 399 for offspring). We placed priors Exp(38.6) on μ, Lo$(2, 0.7)$ on θ with $L = 50$, Exp(0.0673) on b_{F_0}, and fixed hyperparameters $(2, 4)$ on (α_0, η) for model fitting.

Table S4 provides posterior estimates of the immigrant intensity, the branching ratio, the cluster sizes, and the misclassification. The posterior means of μ and γ are close to the true values of $(0.01, 0.8)$. The estimated cluster sizes are comparable to the true sizes: immigrant/offspring sizes (105/429) for the Weibull example and (101/399) for the Weibull

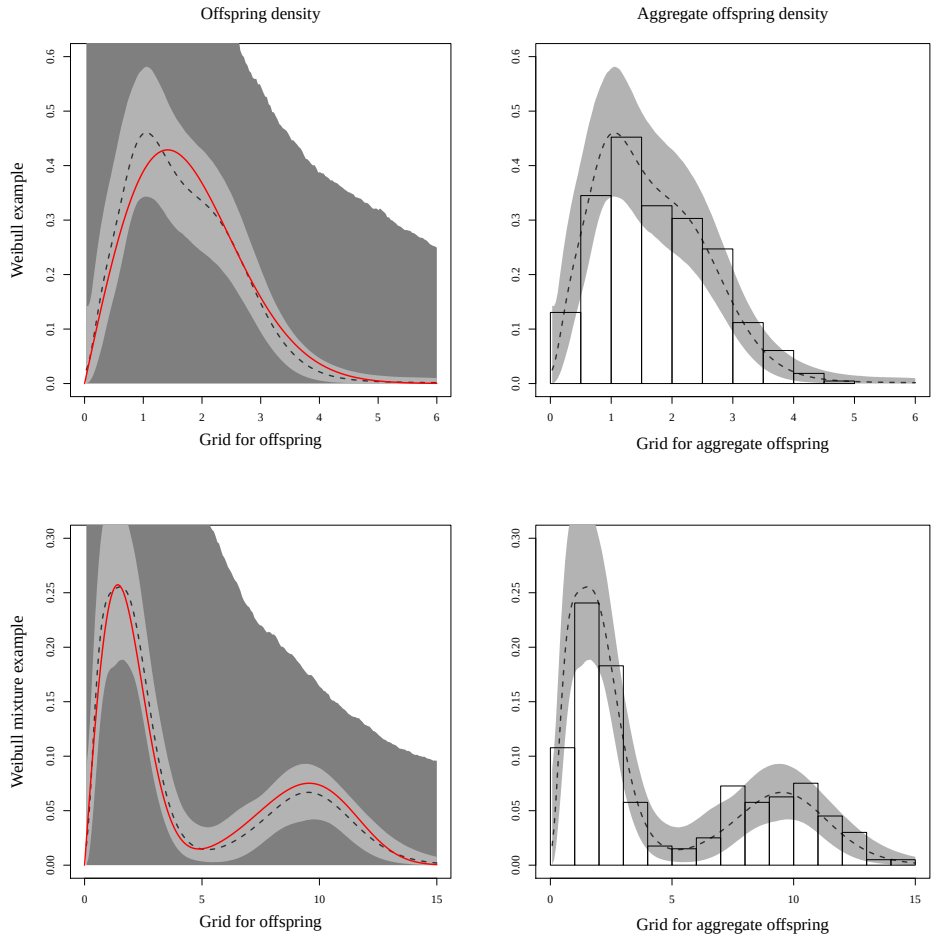

Figure S5: Weibull (top) and Weibull mixture (bottom) examples. Semiparametric model with Erlang mixture offspring intensity. The left panels present the posterior means (dashed) and posterior 95% interval estimates (light gray) for offspring density functions with: prior uncertainty bounds (dark gray) and underlying density functions (red solid). The right panels display the posterior means and interval estimates for aggregate offspring density functions as well as data histograms.

mixture example. As illustrated in the left column of Figure S5, the estimated offspring density functions adequately account for the non-standard-shaped underlying functions, which corroborates the flexibility of the semiparametric model. In the right column are shown the differences between the aggregate offspring density functions and the data histograms. They are useful in understanding the local discrepancy between the estimated and underlying offspring density functions. For instance, the aggregate offspring density superimposed on the data histogram in the Weibull example confirms that the over and underestimations at $x = 1$ and 1.8 are due to the non-representative data set, rather than

an artifact of the model. The results above are based on $20,000$ posterior samples, obtained by $20,000$ burn-in from a total of $40,000$ samples.

D.1.3 Fully nonparametric model

The section involves three examples simulated using the following HP intensities:

- $\lambda^*(t) = 0.02 + 0.8 \sum_{t_i<t} \text{Exp}(t - t_i|1),\ t \in (0, 5000)$
- $\lambda^*(t) = 200\text{We}(t|3, 5000) + 0.6 \sum_{t_i<t} [0.6\text{We}(t - t_i|2, 2) + 0.4\text{We}(t - t_i|5, 10)],\ t \in (0, 10000)$
- $\lambda^*(t) = 200[0.6\text{We}(t|1.5, 2000) + 0.4\text{We}(t|7, 7500)] + 0.6 \sum_{t_i<t} [0.6\text{We}(t - t_i|2, 2) + 0.4\text{We}(t - t_i|5, 10)],$ $t \in (0, 10000)$

Each of the HPs generates 534 (105 for immigrants/429 for offspring), 500 (202/298), and 542 (206/336) point observations, respectively. Inferences below are based on: $30,000$ posterior samples (after 10,000 burn-in) in the first two examples; and $10,000$ posterior samples (after 30,000 burn-in) in the last example.

The HP realization of the first example uses a conventional choice for the conditional intensity, that is, a combination of constant immigrant intensity and an exponential offspring density function (e.g., Hawkes, 1971; Adamopoulos, 1976), which is still used in many applications to account for temporal changes of the point process (e.g., Mohler (2014) in criminology, Rizoiu et al. (2017) in social science). In other words, the data set represents one of the simplest point patterns that one might think would be observed in HP applications at first glance. We fitted the synthetic data to the semiparametric and nonparametric models based on Erlang mixtures and compared them to a parametric model consisting of immigrant intensity and exponential offspring density as in the conventional intensity function. As the parametric model is identical in form to the underlying function, estimates from the model should serve as the gold standard for evaluating our proposed models in the comparison.

The prior for each model is as follows: (parametric model) Exp(18.7) for μ, Ga$(1, 5)$ for γ, and Exp(1) for rate parameter of the exponential offspring density; (semiparametric model) Exp(18.7) for μ, Lo$(2, 0.5)$ for θ with $L = 50$, Exp(0.101) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η); (nonparametric model) Lo$(2, 250)$ for θ_I with $L_I = 50$,

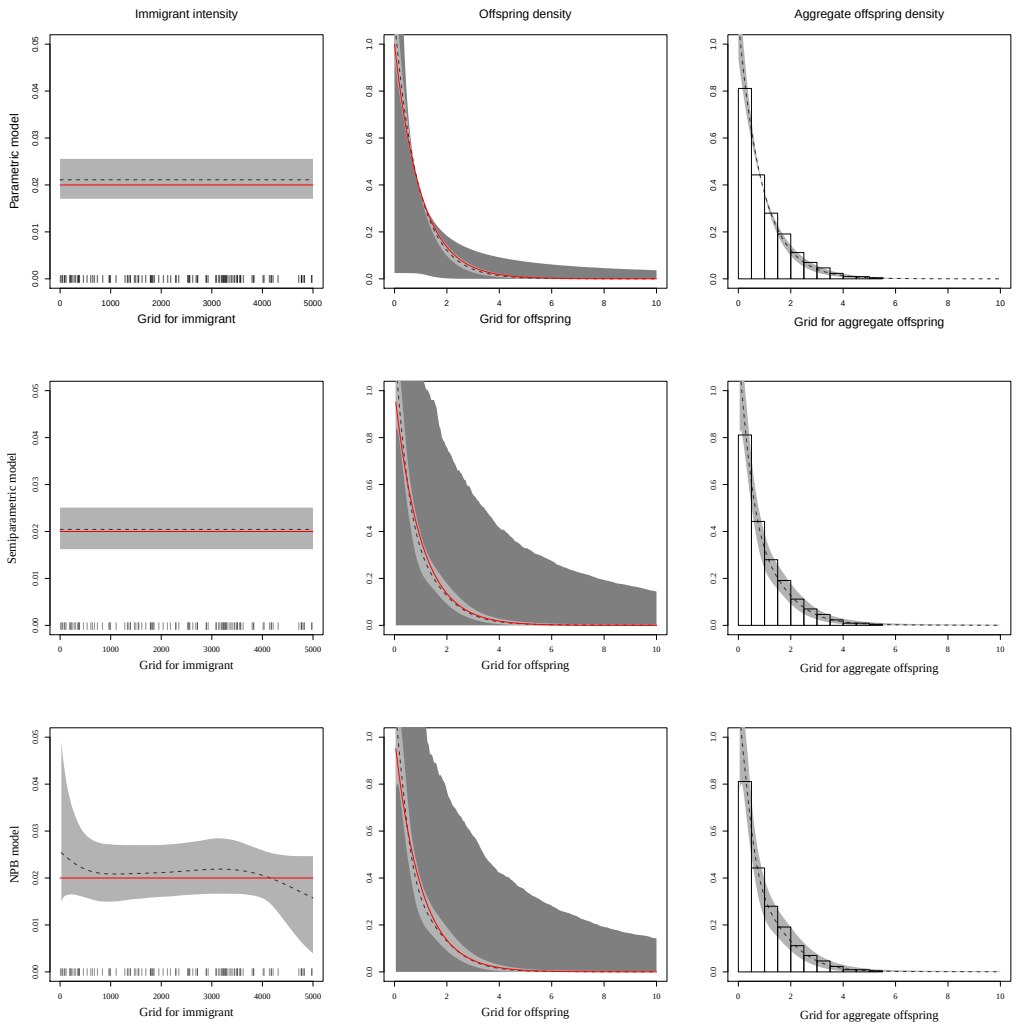

Figure S6: Constant immigrant and exponential offspring example. Parametric (top row), semiparametric (middle row), and nonparametric (bottom row) models. The left column represents immigrant intensity estimates. The middle and right columns display offspring density and aggregate offspring density estimates.

Exp(0.1) for c_0, Exp(0.0534) for b_{G_0}, Lo$(2, 0.5)$ for θ_O with $L_O = 50$, Exp(0.101) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

Figure S6 illustrates that our models perform well when it comes to intensity/density estimation. The Erlang mixtures in our models lead to relatively large posterior interval estimates compared to the parametric model, which is to be expected given the prior flexibility of the mixture modeling. Although an Erlang mixture consists of decreasing- and unimodal-shaped density functions, the Erlang-mixture-based immigrant intensity of the NPB model captures well the constant underlying function without overfitting.

Ex1	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
Para	0.798(0.039)	105(4)	429(4)	10(3)	10(2)	0.036(0.006)
Semi	0.808(0.040)	101(5)	433(5)	9(3)	13(4)	0.041(0.008)
NPB	0.805(0.039)	103(5)	431(5)	10(3)	12(3)	0.040(0.008)
	TVD^a	TVD^b	TVD^c			
Para	0	0.038(0.025)	0.052(0.021)			
Semi	0	0.053(0.024)	0.055(0.020)			
NPB	0.037(0.029)	0.056(0.024)	0.056(0.021)			
Ex2	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
Semi	0.604(0.042)	198(12)	302(12)	56(9)	60(6)	0.231(0.017)
NPB	0.603(0.042)	198(12)	302(12)	56(9)	60(6)	0.231(0.017)
	TVD^a	TVD^b	TVD^c			
Semi	0.035(0.020)	0.082(0.029)	0.093(0.029)			
NPB	0.047(0.020)	0.080(0.028)	0.090(0.028)			
Ex3	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
NPB	0.622(0.039)	205(10)	337(10)	53(8)	55(4)	0.199(0.014)
	TVD^a	TVD^b	TVD^c			
NPB	0.064(0.023)	0.072(0.025)	0.080(0.026)			

Table S5: Posterior means and standard deviations for γ, cluster sizes, misclassifications, and TVDs (Ex1: constant immigrant and exponential offspring, Ex2: Weibull immigrant and Weibull mixture offspring, and Ex3: Weibull mixtures for both immigrant and offspring).

Quantitative results are also comparable; all models produce the branching ratio and the cluster sizes estimates that are agree with the true values of $\gamma = 0.2$, $n_I = 105$, and $n_O = 429$ (Table S5). In both parametric and semiparametric models, the constant immigrant intensity results in a uniform density over $(0, T)$ regardless of the intensity estimate, so both the underlying and estimated density functions provide 0 for TVD^a.

Figure S7 shows the estimated first- and second-order intensities along with their corresponding true values, for which the analytical formulas can be found in Hawkes (1971). All

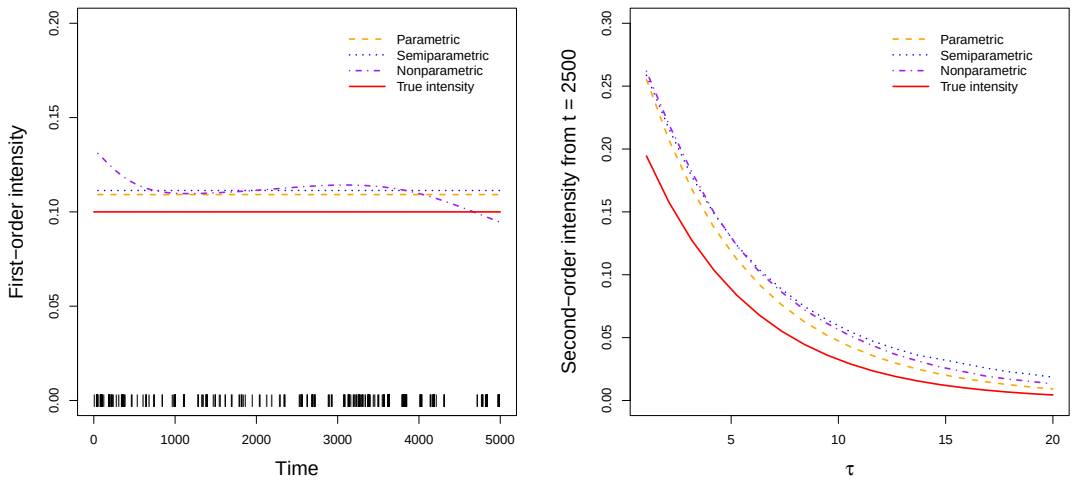

Figure S7: Constant immigrant and exponential offspring example. First-order (left) and second-order (right) intensity estimates from each model. Bars at the bottom of the left panel indicate the observed point pattern.

models exhibit posterior means that exceed the true function values, but their 95% posterior interval estimates (not shown) adequately cover the true intensities. Overestimation of the parametric model at similar levels to those of the other models indicates that the unfavorable results are due to the data, not an artifact of our models.

In the next example, we simulated a HP point pattern with an irregular-shaped intensity function comprised of unimodal immigrant intensity and bimodal excitation function. Semi- and non-parametric models were applied to capture the non-standard-shaped intensity function. The semiparametric method models immigrant intensity using a parametric function proportional to a Weibull density, as in the underlying function.

We set the model parameters to the following priors and values: (semiparametric model) Exp(0.0018), Exp(0.1), and Exp(0.0001) for (μ, a_ψ, b_ψ), Lo$(2, 0.7)$ for θ with $L = 50$, Exp(0.0673) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η); (nonparametric model) Lo$(2, 500)$ for θ_I with $L_I = 50$, Exp(0.1) for c_0, and Exp(0.025) for b_{G_0}, Lo$(2, 0.7)$ for θ with $L = 50$, Exp(0.0673) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

Both models well reproduce the global patterns of the underlying functions and include them within their 95% posterior interval estimates (Figure S8). The second mode of the underlying excitation function has a slight leftward shift in the posterior estimates. As can be seen in the last panel of the figure, the estimated aggregate offspring density function

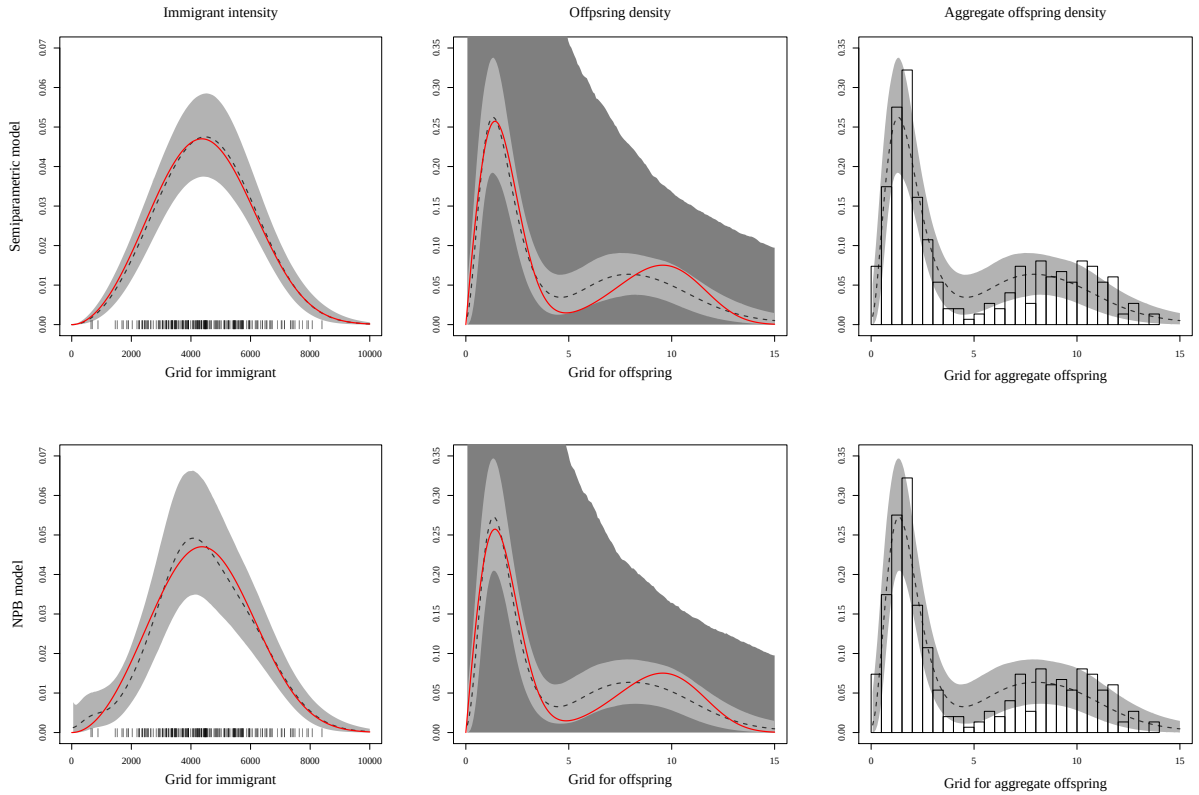

Figure S8: Weibull immigrant and Weibull mixture offspring example. Semiparametric (top) and nonparametric (bottom) models. Estimates for the immigrant intensity (left), offspring density (middle), and aggregate offspring density (right) functions.

does not follow the data histogram and is still leftward shifted at the second mode. This indicates that the deviation is not due to the use of unrepresentative data. Instead, in Table S5, the higher misclassification rate ($R = 0.231$) compared to the other examples provides a clue to the cause of the shift. Specifically, $M_O = 60$ ($\approx 20\%$) out of $n_O = 302$ points are misclassified as offspring, which are supposed to be immigrants. A large number of misclassified offspring points may cause confounding effects. Also, 242 correctly classified offspring points may not be sufficient to represent the two modes accurately.

The intensity function of the last example is more difficult to capture, as the immigrant intensity function is also bimodal. This simulation study is intended to demonstrate the flexibility of the nonparametric model. So, we fit the Erlang-mixture-based NPB model to the data, with the priors: $\text{Lo}(2, 500)$ for θ_I with $L_I = 50$, $\text{Exp}(0.1)$ for c_0, $\text{Exp}(0.0271)$ for b_{G_0}, $\text{Lo}(2, 0.7)$ for θ_O with $L_O = 50$, $\text{Exp}(0.0673)$ for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

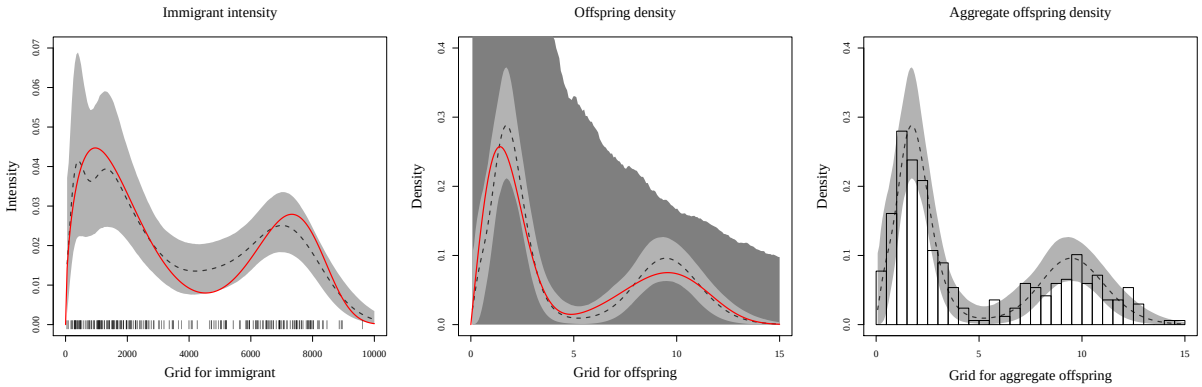

Figure S9: Example of Weibull mixtures for both immigrant and offspring. Nonparametric model. Estimates of the immigrant intensity (left), offspring density (middle), and aggregate offspring density (right) functions.

The nonparametric model captures well the global pattern of the two bimodalities, but with some local discrepancies (Figure S9). Data representativeness and model misclassification (M_I and M_O) are the primary causes of such differences; for example, the histogram of observed data in the last panel can explain why the density estimates are greater than the underlying function at the two modes in the offspring density estimation. In immigrant intensity estimation, only 152 points are used on average, due to $M_I = 53$ and $n_I = 205$ (see Table S5), which may be insufficient to retrieve the bimodal underlying function, whereas the observed 206 immigrant points well support the bimodality of the function (not shown).

D.2 Uniform-mixture-based models

We present two examples of decreasing offspring density functions to illustrate the proposed models, given in Section 2.2.2 of the paper. To focus on the offspring density estimation, we assume constant immigrant for the data-generating intensities and for the models. Again, the models can be extended to fully nonparametric models by replacing the constant immigrant with the Erlang mixture of Section 2.1 of the paper. We employed power-law (i.e., Lomax) and Weibull (with shape < 1) densities for the offspring density functions in the following simulations. These two densities have polynomial heavy tails, and especially, the power-law density is considered to be the underlying offspring density of HPs for earthquake data, namely aftershocks. We also carried out a simulation with an exponential offspring density function, as in the first example of Section D.1.3 and found that the mod-

Example	Model	μ	γ	Cluster size		
				n_I	n_O	M_I
Power-law	DP	0.021(0.002)	0.806(0.040)	103(6)	431(6)	11(3)
	GW	0.020(0.002)	0.807(0.041)	102(6)	432(6)	10(3)
Weibull	DP	0.009(0.001)	0.798(0.042)	92(2)	378(2)	7(2)
	GW	0.009(0.001)	0.801(0.042)	92(3)	378(3)	7(2)
		Misclassification		TVD[b]	TVD[c]	
		M_O	R			
Power-law	DP	13(4)	0.045(0.009)	0.059(0.022)	0.055(0.018)	
	GW	14(4)	0.045(0.009)	0.058(0.022)	0.054(0.017)	
Weibull	DP	6(1)	0.028(0.005)	0.061(0.020)	0.066(0.020)	
	GW	6(1)	0.028(0.005)	0.056(0.017)	0.060(0.018)	

Table S6: Posterior means and standard deviations of (μ,γ), cluster sizes, misclassifications, and TVDs. DP and GW refer to DP- and GW-based semiparametric models, respectively.

els captured the underlying density function well (not shown). We generated 534 points (105 for immigrants/429 for offspring) and 470 points (92 for immigrants/378 for offspring) individually from each of the underlying HP intensity functions defined as

- $\lambda^*(t) = 0.02 + 0.8 \sum_{t_i<t} \text{Lo}(t - t_i|2, 1)$ for $t \in (0, 5000)$; and
- $\lambda^*(t) = 0.01 + 0.8 \sum_{t_i<t} \text{We}(t - t_i|0.5, 2)$ for $t \in (0, 10000)$.

For the first example, we assigned priors Exp(18.7) to μ, Ga$(2, 4)$ to γ, Exp(0.333) to a_{F_0}, Exp(0.284) to β, Ga$(4, 0.5)$ to α, and $L = 85$ for the DP-based model; and the same priors to the common parameters, (μ,a_{F_0},γ,β), and Be$(3, 3)$ to ζ with $L = 50$ for the GW-based model. Priors for the second example are Exp(42.6) for μ, Ga$(2, 4)$ for γ, Exp(0.2) for a_{F_0}, Exp(0.116) for β, Ga$(4, 0.5)$ for α, and $L = 85$ for the DP-based model; and, with the same priors for the common parameters, Be$(3, 3)$ for ζ with $L = 50$ for the GW-based model.

Quantitative model comparison of Table S6 shows that the two models have consistent results with nearly identical estimated values across all criteria. Posterior estimates for (μ, γ, n_I, n_O) are aligned with the true values: $(0.02, 0.8, 105, 429)$ for the power-law example and $(0.01, 0.8, 92, 378)$ for the Weibull example.

Both models perform well in estimating offspring densities, providing posterior uncer-

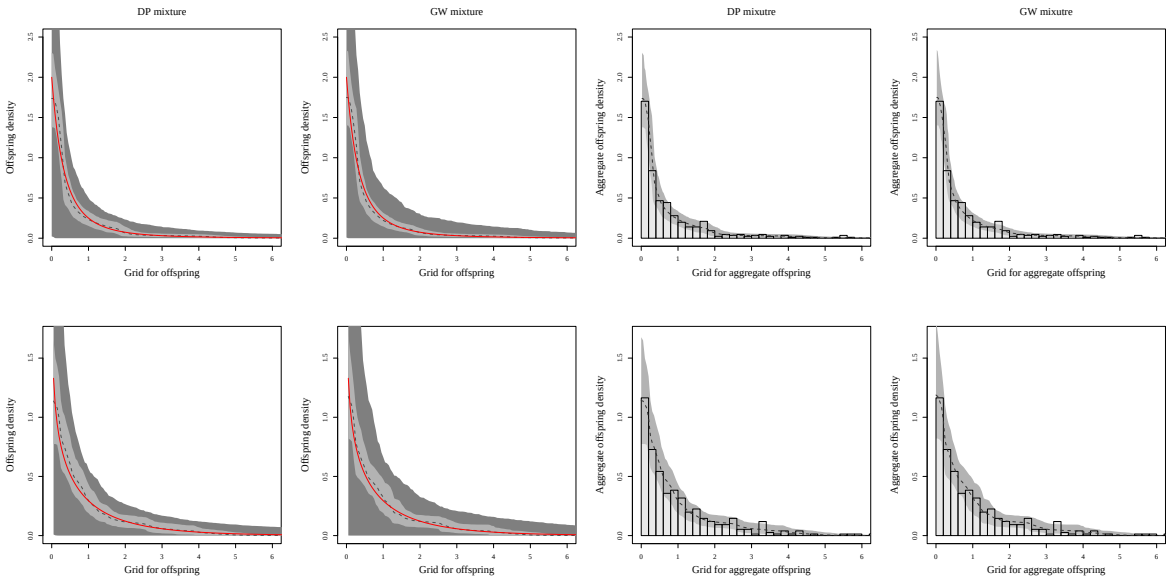

Figure S10: Power-law (top) and Weibull (bottom) offspring density examples. The first two columns show posterior point (dashed) and 95% interval (light-gray) estimates of the offspring density function with prior 95% uncertainty bounds (dark-gray) and underlying functions (red). The last two columns display posterior point (dashed) and interval (light-gray) estimates of the aggregate offspring density function with the histogram of all distances between offspring and their parent.

tainty bounds that include the underlying functions (see, Figure S10). Discrepancies in the estimation can be explained by the aggregate offspring density estimates that follow the histogram of the data. For instance, underestimation at $x = 0$ in the power-law example stems from a lack of small-valued data rather than a model artifact.

All results are based on our prior specification strategy, and Section C.2 presents some results of sensitivity analysis. We used $1,000$ posterior samples for the analysis, obtained by discarding $5,000$ burn-in and thinning every 5th step for DP and 15th step for GW. Empirically, for the same length of MCMC chains, the GW-based model tends to have smaller effective sizes of posterior samples, namely the offspring density posterior samples, due to the poor mixing of the weight parameter ζ. Thus, we used that longer chain of total $20,000$ MCMC iterations for GW, whereas $10,000$ MCMC iterations were used for DP.

In terms of computing time, despite different MCMC iterations in the two models, there is no substantial difference between the models (≈ 6 hours for each model in both examples) as the DP takes the larger mixing components $L = 85$ to balance the model flexibility and

smoothness, whereas the GW-based model works well with $L = 50$.

References

Adamopoulos, L. (1976), "Cluster models for earthquakes: Regional comparisons," *Journal of the International Association for Mathematical Geology*, 8, 463–475.

Adams, R. P., Murray, I., and MacKay, D. J. (2009), "Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*.

Balderama, E., Schoenberg, F. P., Murray, E., and Rundel, P. W. (2012), "Application of branching models in the study of invasive species," *Journal of the American Statistical Association*, 107, 467–476.

Butzer, P. (1954), "On the extensions of Bernstein polynomials to the infinite interval," *Proceedings of the American Mathematical Society*, 5, 547–553.

Chen, F. and Stindl, T. (2018), "Direct likelihood evaluation for the renewal Hawkes process," *Journal of Computational and Graphical Statistics*, 27, 119–131.

Da Fonseca, J. and Zaatour, R. (2014), "Hawkes process: Fast calibration, application to trade clustering, and diffusive limit," *Journal of Futures Markets*, 34, 548–579.

Daley, D. J., Vere-Jones, D., et al. (2003), *An introduction to the theory of point processes: volume I: elementary theory and methods*, Springer.

Donnet, S., Rivoirard, V., and Rousseau, J. (2020), "Nonparametric Bayesian estimation for multivariate Hawkes processes," *The Annals of Statistics*, 48, 2698–2727.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Aki, V., and Rubin, D. B. (2013), *Bayesian data analysis (3rd ed.)*, Chapman and Hall/CRC.

Hardiman, S. J., Bercot, N., and Bouchaud, J.-P. (2013), "Critical reflexivity in financial markets: a Hawkes process analysis," *The European Physical Journal B*, 86, 1–9.

Hawkes, A. G. (1971), "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, 58, 83–90.

Hawkes, A. G. and Oakes, D. (1974), "A cluster process representation of a self-exciting process," *Journal of Applied Probability*, 11, 493–503.

Kagan, Y. Y. (1991), "Likelihood analysis of earthquake catalogues," *Geophysical journal international*, 106, 135–148.

Kalbfleisch, J. D. (1978), "Non-parametric Bayesian analysis of survival time data," *Journal of the Royal Statistical Society. Series B*, 40, 214–221.

Kim, H. and Kottas, A. (2022), "Erlang mixture modeling for Poisson process intensities," *Statistics and Computing*, 32, 3.

Laub, P. J., Lee, Y., and Taimre, T. (2021), *The elements of Hawkes processes*, Springer.

Lee, S. C. and Lin, X. S. (2010), "Modeling and evaluating insurance losses via mixtures of Erlang distributions," *North American Actuarial Journal*, 14, 107–130.

Li, Y., Lee, J., and Kottas, A. (2023), "Bayesian nonparametric Erlang mixture modeling for survival analysis," *Computational Statistics and Data Analysis*, To appear.

Mohler, G. (2014), "Marked point process hotspot maps for homicide and gun crime prediction in Chicago," *International Journal of Forecasting*, 30, 491–497.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011), "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, 106, 100–108.

Musmeci, F. and Vere-Jones, D. (1992), "A space-time clustering model for historical earthquakes," *Annals of the Institute of Statistical Mathematics*, 44, 1–11.

Ogata, Y. (1981), "On Lewis' simulation method for point processes," *IEEE transactions on information theory*, 27, 23–31.

— (1988), "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical association*, 83, 9–27.

— (1998), "Space-time point-process models for earthquake occurrences," *Annals of the Institute of Statistical Mathematics*, 50, 379–402.

Rambaldi, M., Pennesi, P., and Lillo, F. (2015), "Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach," *Physical Review E*, 91, 012819.

Rasmussen, J. G. (2013), "Bayesian inference for Hawkes processes," *Methodology and Computing in Applied Probability*, 15, 623–642.

Reinhart, A. (2018), "A review of self-exciting spatio-temporal point processes and their applications," *Statistical Science*, 33, 299–318.

Rizoiu, M.-A., Lee, Y., Mishra, S., and Xie, L. (2017), "A tutorial on hawkes processes for events in social media," *arXiv preprint arXiv:1708.06401*.

Taddy, M. A. and Kottas, A. (2012), "Mixture modeling for marked Poisson processes," *Bayesian Analysis*, 7, 335–362.

Veen, A. and Schoenberg, F. P. (2008), "Estimation of space–time branching process models in seismology using an em–type algorithm," *Journal of the American Statistical Association*, 103, 614–624.

Xiao, S., Kottas, A., Sansó, B., and Kim, H. (2021), "Nonparametric Bayesian modeling and estimation for renewal processes," *Technometrics*, 63, 100–115.

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. (2019), "Efficient Non-parametric Bayesian Hawkes Processes," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, URL `https://doi.org/10.24963/ijcai.2019/597`.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. (2020), "Efficient Inference for Nonparametric Hawkes Processes Using Auxiliary Latent Variables," *Journal of Machine Learning Research*, 21, 1–31, URL `http://jmlr.org/papers/v21/19-930.html`.

Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. (2021), "Efficient EM-variational inference for nonparametric Hawkes process," *Statistics and Computing*, 31, 1–11.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002), "Stochastic declustering of space-time earthquake occurrences," *Journal of the American Statistical Association*, 97, 369–380.

supplementary