Interval-censored Hawkes processes

Marian-Andrei Rizoiu   Marian-Andrei.Rizoiu@uts.edu.au
University of Technology Sydney
Ultimo NSW 2007, Australia

Alexander Soen   alexander.soen@anu.edu.au
The Australian National University
Canberra ACT 2601, Australia

Shidi Li   shidi.li@anu.edu.au
The Australian National University
Canberra ACT 2601, Australia

Pio Calderon   piogabrielle.b.calderon@student.uts.edu.au
University of Technology Sydney
Ultimo NSW 2007, Australia

Leanne J. Dong   leanne.dong@concordia.ca
Concordia University
Montréal, Canada

Aditya Krishna Menon   adityakmenon@google.com
Google Research

Lexing Xie   lexing.xie@anu.edu.au
The Australian National University
Canberra ACT 2601, Australia

Editor: Mohammad Emtiyaz Khan

Abstract

Interval-censored data solely records the aggregated counts of events during specific time intervals – such as the number of patients admitted to the hospital or the volume of vehicles passing traffic loop detectors – and not the exact occurrence time of the events. It is currently not understood how to fit the Hawkes point processes to this kind of data. Its typical loss function (the point process log-likelihood) cannot be computed without exact event times. Furthermore, it does not have the independent increments property to use the Poisson likelihood. This work builds a novel point process, a set of tools, and approximations for fitting Hawkes processes within interval-censored data scenarios. First, we define the Mean Behavior Poisson process (MBPP), a novel Poisson process with a direct parameter correspondence to the popular self-exciting Hawkes process. We fit MBPP in the interval-censored setting using an interval-censored Poisson log-likelihood (IC-LL). We use the parameter equivalence to uncover the parameters of the associated Hawkes process. Second, we introduce two novel exogenous functions to distinguish the exogenous from the endogenous events. We propose the multi-impulse exogenous function – for when the exogenous events are observed as event time – and the latent homogeneous Poisson process exogenous function – for when the exogenous events are presented as interval-censored volumes. Third, we provide several approximation methods to estimate the intensity and compensator function of MBPP when no analytical solution exists. Fourth and finally, we

connect the interval-censored loss of MBPP to a broader class of Bregman divergence-based functions. Using the connection, we show that the popularity estimation algorithm Hawkes Intensity Process (HIP) (Rizoiu et al., 2017b) is a particular case of the MBPP. We verify our models through empirical testing on synthetic data and real-world data. We find that our MBPP outperforms HIP on real-world datasets for the task of popularity prediction. This work makes it possible to efficiently fit the Hawkes process to interval-censored data.

Keywords: Hawkes process, Interval-censored, Mean Behavior Poisson process, Bregman divergence, popularity prediction, multi-impulse exogenous function, latent homogeneous Poisson process exogenous function

1. Introduction

Point processes are a class of well-understood mathematical instruments to model the occurrence of events in time (and optionally space). The *Hawkes process* (Hawkes, 1971) is a particular type of point process that can model *self-excitation* – i.e., the occurrence of one event increases the likelihood of future events. The Hawkes process was successfully deployed for applications where the events are individually observable – such as earthquakes (Ogata, 1988; Cox and Isham, 1980; Daley and Vere-Jones, 2003), financial transactions (Bowsher, 2007; Hardiman et al., 2013), or social media postings (Zhao et al., 2015; Kobayashi and Lambiotte, 2016; Mishra et al., 2016). However, in many scenarios, one does not know the exact historical event times, but rather only the counts of events in specific time intervals. Such *interval-censored* (Bernoulli, 1760) scenarios could be due to data availability – say, for epidemics when the exact infection time for each individual is unknown and the number of hospital admissions per day is known – or due to data privacy – e.g., e-commerce websites do not expose who bought a particular item, but they will report rankings of most bought items. Due to its lack of independent increments property (see Section 2.2), the Hawkes process cannot be directly fitted in the interval-censored scenario. This work circumvents this shortcoming by building a novel point process, a set of tools, and approximations for fitting Hawkes processes within interval-censored data scenario.

There is prior work that deals with interval-censored data and Hawkes processes. Here we highlight three families of approaches, and we present an in-depth review in Section 9.1. The first family is the integer-valued auto-regressive (INAR) models (Kirchner, 2016, 2017; Manolakis et al., 2019). These have been shown to yield approximations of the Hawkes model parameters after appropriate scaling. However, these works do not expose the connection between the Hawkes process's event-time and interval-censored formulations. The second family of approaches refers to interval-censored data as panel count data (Sun and Zhao, 2013; Zhu et al., 2014; Ding et al., 2018; Moreno et al., 2020). These works typically concentrate on the non-parametric estimation of the intensity of Poisson processes from interval-censored data. As far as we are aware, they do not specifically consider the Hawkes process and are not directly link to the current work apart from the presentation of the data. The third family is the Hawkes Intensity Processes (HIP) (Rizoiu et al., 2017b). The HIP parameters are fitted by minimizing the squared error of the observed counts to the expected intensity of a Hawkes process. However, HIP has several shortcomings (detailed in Section 2.4), e.g., its fitting objective does not relate to the likelihood of a particular process, and it does not establish a link to the parameters of the underlying Hawkes process.
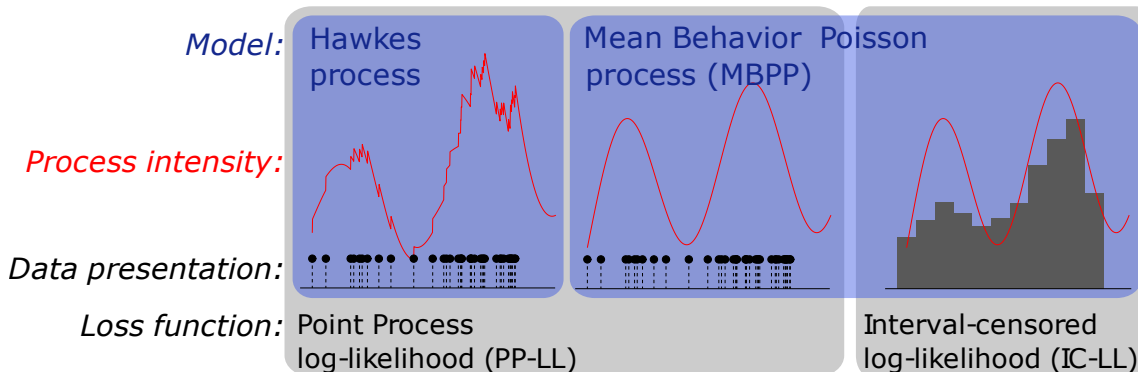
Figure 1: Illustration of the applicability of the Hawkes process and MBPP in event time (left gray bubble) and interval censored settings (right gray bubble). The Hawkes process is usable only when the data is in event time format, which can then be fitted using the standard point process log-likelihood (PP-LL). On the other hand, the MBPP can be used with both event time and interval-censored data by switching between PP-LL and the interval-censored log-likelihood (IC-LL).

To our knowledge, no prior work has proposed a generalized treatment of Hawkes processes across event-time and interval-censored data.

In this paper, we build a new point process and a set of tools to operate with Hawkes processes on both event time and interval-censored data. First, we introduce the Mean Behavior Poisson (MBPP) – a non-homogeneous Poisson process, with a deterministic intensity function equal to the expectation of a Hawkes process intensity, taken over all possible realizations. There is a one-to-one parameter equivalence between MBPP and the Hawkes process. Next, we propose two approximations for the event intensity and compensator (the integral of the intensity) of MBPP, and we use them in an *interval-censored log-likelihood* (IC-LL) which allows fitting the process on interval-censored data. We show that IC-LL can be interpreted as a Bregman divergence, and we use it to draw a connection between the MBPP and the HIP (Rizoiu et al., 2017b) models. Last, we empirically show in Section 7 that MBPP allows one to retrieve the Hawkes model parameters for both interval-censored and event time data. Fig. 1 illustrates the relation between Hawkes and MBPP. When a realization is observed as event times, both Hawkes and MBPP can successfully fit the model using a *point process likelihood* (PP-LL). When the same realization is observed interval-censored, only MBPP can be used to retrieve the original parameters using the *interval-censored likelihood* (IC-LL) (see Section 4).

We also study scenarios in which the exogenous events in the Hawkes process are separable from the endogenously generated events. This is often the case in real-world applications — such as our popularity prediction exercise in Section 8 where the shares and tweets driving the views of YouTube videos are observed separately from the views process. In this setup, the exogenous events enter the process according to an unknown intensity function, and they can be observed either as event times or interval-censored. We construct two novel exogenous functions – the multi-impulse function for when the exogenous events are event times and the latent homogeneous Poisson process function for when the exogenous

events are interval-censored – and a new fitting procedure that only accounts for endogenous events. Finally, we show that on real-world data our approaches outperform HIP (Rizoiu et al., 2017b), the current state of the art in popularity prediction.

The main contributions of this work are as follows:

C1: we show how to construct a non-homogeneous Mean Behavior Poisson process (MBPP) that approximates the mean behavior of the underlying Hawkes process (Section 3). The likelihood of this process allows the use of a broader class of Bregman divergence based loss functions (Section 5).

C2: we prove that the MBPP defines a causal, linear, continuous-time, time-invariant system with the impulse response computed as an infinite sum of convolutions (Theorem 2). This allows finding both a closed-form solution for the MBPP intensity (where it exists) and a numerical approximation (where it does not exist).

C3: we propose a second approximation for the MBPP intensity (Section 4.3), which is numerically more efficient, can exploit information about interval lengths and be used for predicting future event volumes (i.e., for forecasting).

C4: we introduce a set of tools (exogenous functions, loss functions) that allow fitting the parameters of a Hawkes process using MBPP, when either the exogenous or the endogenous events (or both) are interval-censored (Section 6).

C5: on a real-world dataset containing information about tweets and YouTube views associated with YouTube videos, we show our approach outperforms HIP (Rizoiu et al., 2017b), the current state-of-the art in popularity prediction (Section 8).

2. Background and notations

In this section, we provide an overview of the different mathematical objects used in this paper. First, in Section 2.1, we introduce point processes and Poisson processes using random measures, and we discuss processes with independent increments. Second, in Section 2.2, we introduce the Hawkes self-exciting point process, and we show that it does not have the independent increments property. Third, in Section 2.3, we discuss the parameter estimation of point processes. Last, in Section 2.4, we present the Hawkes intensity process — a Hawkes-based interval-censored process and the previous work closest to the present paper.

In this paper, we use the following elementary notations — for more specialized notations concerning point processes, please refer to Table 1. We denote $[\![p]\!] = 1$ if p is true and $[\![p]\!] = 0$ if p is false, following notation in Knuth (1992). The Laplace transform and inverse Laplace transform are defined as $\mathcal{L}\{\cdot\}$ and $\mathcal{L}^{-1}\{\cdot\}$ respectively. The Dirac delta function $\delta(x)$ is defined by $\int_a^b \delta(x)f(x)\,dx = f(0)$ if f is continuous and $a < 0 < b$. The convolution of functions f and g is defined as $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)\,d\tau$.

2.1 Random measures, Point processes and Poisson processes

In this section, we briefly introduce point processes and Poisson processes using the notion of random measures. For a more complete (yet accessible) introduction of point processes using random measures, we refer to (Grandell, 1977). We opted for this rigorous definition of point processes for completeness reasons. For a more intuitive and streamlined introduction

Table 1: Notation used in this paper. Throughout this paper, we denote observed quantities using sans-serif font face.

Symbol	Meaning	Defined/Introduced
$\lambda(t)$	Hawkes Process Intensity	Eq. (3)
$\xi(t)$	Mean Behavior Poisson Intensity	Eqs. (6) and (9)
$\Lambda(t)$	Hawkes Process Compensator	Eq. (2)
$\Xi(t)$	Mean Behavior Poisson Compensator	Eq. (10)
$N(t)$	Hawkes Counting Process	Eq. (1)
$M(t)$	Mean Behavior Poisson Counting Process	Eq. (18)
$\mathsf{C}(x,y]$	Observed counts in interval $(x,y]$	Eqs. (8) and (17)
$\mathsf{S}(x,y]$	Observed immigrant counts in interval $(x,y]$	Eq. (41)
$\mathsf{F}(x,y]$	Observed offspring counts in interval $(x,y]$	Eq. (51)
$B_\varphi(x)$	Bregman Divergence	Eq. (28)
$s_{m\delta}(t)$	Multi-impulse exogenous function	Eq. (39)
$s_\blacksquare(t)$	Latent Homogeneous Poisson process exog. func.	Eq. (43)
Ψ_T	Set of immigrant events up to time T	Eq. (50)
Υ_T	Set of offspring events up to time T	Eq. (50)

using counting processes and intensity functions, we refer the reader to (Rizoiu et al., 2017a; Laub et al., 2015).

Point processes as random measures. Let $(X, \mathcal{B})$ be a measurable state space and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $M = M(X)$ be the space of all σ-finite measures on the sigma algebra $\mathcal{B}$. Let $\mathcal{M}$ denote the smallest σ-algebra on M with respect to which the function $\mu \mapsto \mu(B)$ is measurable for all $B \in \mathcal{B}$. A *random measure* (Benes and Rataj, 2007) on X is a measurable map $\eta : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{M})$. Let $N \subset M$ be the space of all σ-finite counting measures on $\mathcal{B}$. Furthermore, let $\mathcal{N}$ denote the restriction of $\mathcal{M}$ when only considering σ-finite counting measures on $\mathcal{B}$, i.e., $\mathcal{N} = \{S \cap \mathcal{N} \mid S \in \mathcal{M}\}$. A *point process* (Benes and Rataj, 2007) on X is a measurable map $\hat{\eta} : (\Omega, \mathcal{F}, \mathbb{P}) \to (N, \mathcal{N})$. A temporal point process is a random measure on the real line taking values in the non-negative integers or infinity (Brillinger et al., 2002).

More intuitively, temporal point processes are a family of stochastic processes $(t_n)_{n \in \mathbb{N}_+}$, in which each random variable t_n represents the occurrence time for an event (Klebaner, 2012). Such processes can also be specified as a *counting process* $(N(t))_{t \geq 0}$, where the random variable $N(t)$ represents the number of events that have occurred up to and including time t (Ross, 1995, pg. 59).

The Poisson process. Let Λ be a positive measure such that $\Lambda(A) < \infty$ for every bounded set A. In the literature of martingale treatments of point processes (Jacobsen, 2006), Λ is often denoted as the *compensator* – we also use this term throughout this paper. The Poisson process is a point process which is solely defined by the compensator. In particular, the Poisson process is often characterised by (and can be defined by) the evaluation of the probability on disjoint sets. That is for every collection of disjoint, bounded Borel measurable sets $A_1, \cdots, A_k$, for a Poisson process, the $N(A_i)$ are independent Poisson

random variables with rate $\Lambda(A_i)$:

$$\mathbb{P}(N(A_i) = n_i, i = 1, \cdots, k) = \prod_{i=1}^{k} \frac{\Lambda(A_i)^{n_i}}{n_i!} e^{-\Lambda(A_i)}. \tag{1}$$

Next, we use the notion of counting processes to define the Poisson process, i.e., specifying the compensator Λ. Suppose that $\Lambda(A) = \lambda l(A)$, where $l(A)$ is the Lebesgue measure of A and $\lambda > 0$ a positive constant. We define the *homogeneous Poisson process* $(N(t))_{t \geq 0}$ as

1. $N(0) = 0$ almost surely

2. for any $s < t$, $N(t) - N(s) \sim \text{Poisson}(\lambda(t - s))$

3. for any $s < t \leq s' < t'$, $N(t) - N(s) \perp\!\!\!\perp N(t') - N(s')$,

where $X \perp\!\!\!\perp Y$ denotes that X and Y are independent.

We say a Poisson process is non-homogeneous when λ is no longer a constant, but a non-negative function. Here, the compensator is redefined as $\Lambda(A) = \int_A \lambda(x) \, dx$, and the *non-homogeneous Poisson process* (NHPP) is

(a) $N(0) = 0$ almost surely

(b) for any $s < t$, $N(t) - N(s) \sim \text{Poisson}(\Lambda(s, t))$

(c) for any $s < t \leq s' < t'$, $N(t) - N(s) \perp\!\!\!\perp N(t') - N(s')$.

Associated with the counting process are two additional sequences of interest. The first is the sequence of *event histories* $(\mathcal{H}(t))_{t \geq 0}$, where each random variable $\mathcal{H}(t)$ records the times of all events that have occurred up to and including time t. Strictly speaking $\mathcal{H}(t)$ is a filtration, that is, an increasing sequence of σ-algebras (Laub et al., 2015). The second is the sequence of *conditional intensities* $(\lambda(t \mid \mathcal{H}(t^-)))_{t \geq 0}$ — denoted for simplicity as $\lambda(t)$ —, where each random variable $\lambda(t)$ measures the expected rate of events occurring in an infinitesimal window around the time t. Formally, (Cox and Lewis, 1972, Equation 3.1),

$$\lambda(t) \doteq \lim_{h \to 0^+} \frac{1}{h} \cdot \mathbb{E}\left[N(t^- + h) - N(t^-) \mid \mathcal{H}(t^-)\right]$$
$$= \lim_{h \to 0} \frac{1}{h} \cdot \mathbb{P}\left(N(t^- + h) - N(t^-) = 1 \mid \mathcal{H}(t^-)\right),$$

where the second equality holds for *simple* processes wherein multiple events cannot occur at the same time, that is the jump size $N(t) - N(t^-)$ at each time t can be either 0 or 1. Finally, we can generalize the definition of the compensator $\Lambda(s, t)$ as the integral of the intensity between two moments of time:

$$\Lambda(s, t) \doteq \int_s^t \lambda(\tau) d\tau, \tag{2}$$

where $0 \leq s < t$. We further shorthand $\Lambda(t) \doteq \Lambda(0, t)$.

Processes with independent increments and Markov processes. The process $(N_t)_{t \geq 0}$ is said to have the *independent increments* property when, for any $u \leq s \leq t$, the random variables $N_t - N_s$ and N_u are independent. The process $N = (N_t)_{t \geq 0}$ is said to be a *Markov process* with respect to $\mathcal{H}(t)$ if $\mathbb{E}[X_s \mid \mathcal{H}(t)] = \mathbb{E}[X_s \mid X_t]$ whenever $s \geq t$. In simple terms, this means that the future of a Markov process is independent of its past. Moreover, if N is a Markov process and conditional stationary, i.e., the conditional distribution is invariant to time translation, it results that it is a homogeneous Markov process. That is, the transition probability function $\mathbb{P}(N(t + s) \leq y \mid N(t) = x)$ is independent to t. Processes with independent increments are Markov processes. However, the converse is not true — not every Markov process obeys the property of independent increments. One such example is the Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930), a Markov process with the "return to mean" property – intuitively, the process is dragged towards zero when it is far away from zero. As a result, the increments are not independent and even negatively correlated. Given the definition in Eq. (1), the Poisson process has the independent increments property and, therefore, is a Markov process.

2.2 Hawkes process

The *Hawkes process* is a temporal point process with the property of *self-excitation*. Intuitively, self-excitation means that the occurrence of one event increases the likelihood of further events in the near future. The process was widely applied in analyzing social media (Kobayashi and Lambiotte, 2016; Cao et al., 2017; Zhang et al., 2019), earthquake aftershocks (Ogata, 1988; Helmstetter and Sornette, 2002a), nuclear physics (Snyder and Miller, 1991), neuronal activity (Apostolopoulou et al., 2019), online advertising (Parmar et al., 2017) and finance (Bacry et al., 2015).

The Hawkes process is a doubly stochastic point process – i.e., the event intensity function is also stochastic and depends upon the previous values of the process itself – (Hawkes, 1971):

$$
\begin{aligned}
\lambda(t) &= s(t) + \int_0^{t^-} \phi(t - s) \, \mathrm{d}N(s) \\
&= \underbrace{s(t)}_{\text{exogenous}} + \underbrace{\sum_{t_i < t} \phi(t - t_i)}_{\text{endogenous}},
\end{aligned}
\tag{3}
$$

where $s(\cdot)$ is the (deterministic) *background intensity* function which controls the arrival of external events into the system — dubbed as *immigrants*; $s(\cdot)$ is also known as the exogenous intensity. $\phi(\cdot)$ is the *kernel* function which controls the arrival of *offspring* — events spawned through self-excitement — by any one given event. The arrival of offspring generated by any previous event is controlled by the second term of Eq. (3), also known as the endogenous intensity. $\{t_i\}$ denotes a realisation of the Hawkes process, with the immigrants and the offspring are collectively denoted as events. In the rest of this paper, we use t_i interchangeably to denote both events and event times.

The Hawkes process is said to have an exponential kernel when $\phi(\tau) = \kappa\theta e^{-\theta\tau}$, with κ the *excitation level* and $\theta > 0$ the *decay rate*; similarly, the Hawkes process has a power-law kernel when $\phi(\tau) = \kappa(\tau + c)^{-(1+\theta)}$, with c a time-warping parameter to keep the kernel

bounded when τ approaches zero, and κ and θ have similar meanings as for the exponential kernel. Note that by the endogenous summation in Eq. (3), λ depends on the history of events that have occurred up to time t. Thus given a fixed history $\mathcal{H}(t^-)$, the intensity before time t is deterministically calculated by Eq. (3). For times greater of equal to t, the intensity is a random variable.

Relation to the Poisson process and processes with independent increments. The Hawkes process is not a Poisson process but rather a non-Markovian extension of the Poisson process. A Hawkes process can also be viewed as a cluster of Poisson processes (Hawkes and Oakes, 1974), in which each event spawns offspring following a non-homogeneous Poisson process of intensity $\phi(t)$. The link between events and their offspring induces the *branching structure* (see (Rizoiu et al., 2017a) for more details). Because the Hawkes process's intensity function depends on the previous events, it follows that the Hawkes process is not a Markov process and, as a result, does not have the independent increments property (unlike the Poisson point process). In particular, the number of jumps in a given interval is thus a function of the number of jumps that happened earlier. We will leverage this result in Section 4 for fitting Hawkes processes in the interval-censored setting.

2.3 Parameter estimation

We are given a sequence $\mathcal{H}(T) \doteq \{t_n \mid t_n \leq T\}$ of observed events up to an arbitrary maximum time $T > 0$, usually $T \doteq \max_n t_n$. We try to find the set of parameters θ of a given point process which represent best the sequence. As shown by Daley and Vere-Jones (2003), this amounts to minimizing the negative log-likelihood with respect to θ, which in turn maximizes the probability of events occurrence at the observed times Ω_T:

$$\mathcal{L}(\theta; T) \doteq - \sum_{t_n \in \Omega_T} \log \lambda(t_n; \theta) + \int_0^T \lambda(u; \theta) \, \mathrm{d}u. \tag{4}$$

As we can see, the log-likelihood requires observing the timing of events, which may not be the case in real application scenarios in which only volumes of events are observed (see Section 4). We denote the standard negative log-likelihood loss in Eq. (4) as the *point process log-likelihood* (PP-LL) loss function for the remainder of the paper.

2.4 Hawkes intensity process

Rizoiu et al. (2017b) proposed the Hawkes intensity process (HIP), a Hawkes-derived process that operates in the interval-censored setting – i.e., the set of observation times $O = \{o_i \mid i = 0, \ldots, m, o_i \leq T\}$ partitions the time line into non-overlapping segments and only the number of events in each segment is observed (see Section 4.1 for more details). The HIP fitting objective is

$$\min_\theta \sum_{i=1}^m (\mathsf{C}(o_{i-1}, o_i] - \xi(o_i; \theta))^2, \tag{5}$$

where $\mathsf{C}(o_{i-1}, o_i]$ are the observed number of events having occurred in the ith interval $(o_{i-1}, o_i]$ and $\xi(t) \doteq \mathbb{E}\left[\lambda(t)\right]$ is the *expected intensity* of the Hawkes process.

The key ingredient of HIP is being able to efficiently compute ξ: Theorem 2.1 of (Rizoiu et al., 2017b) provides a integral equation for $\xi(t)$, namely

$$
\begin{aligned}
\xi(t) &= s(t) + \int_0^t \phi(s) \cdot \xi(t-s)\,\mathrm{d}s \\
&= s(t) + \int_0^t \phi(t-s) \cdot \xi(s)\,\mathrm{d}s.
\end{aligned}
\tag{6}
$$

For completeness, we include a derivation of this result in Appendix A. To optimize Eq. (5), one must (approximately) solve this integral equation to compute each $\xi(t;\theta)$. When $s(t) = \mu_0 + (\mu_1 - \mu_0) \cdot e^{-\gamma \cdot t}$, and ϕ corresponds to the exponential kernel, Dassios and Zhao (2013) showed that

$$
\xi(t) = \mu_1 \cdot e^{-(\gamma-\alpha)\cdot t} + \frac{\mu_0 \cdot \gamma}{\gamma - \alpha} \cdot (1 - e^{-(\gamma-\alpha)\cdot t}).
$$

For more general kernels, such as the power-law kernel, the closed-form solution for ξ often does not exist. Rizoiu et al. (2017b) proposed to approximately solve Eq. (6) by discretising time using the pre-defined observation endpoints $\{o_i\}_{i=0}^m$, yielding

$$
\hat{\xi}[o_i; \theta] = \mu[o_i] + \sum_{s=0}^{i-1} \phi(o_i - o_s) \cdot \hat{\xi}[o; \theta],
\tag{7}
$$

where the bracket notation $\hat{\xi}[\cdot]$ is used to explicate the discretization. One can now use Eq. (7) to compute each $\hat{\xi}[\cdot; \theta]$, beginning with the base case $\hat{\xi}[0; \theta] = a$. Equipped with this, one solves the approximation to the HIP objective:

$$
\min_\theta \sum_{i=1}^m (\mathsf{C}(o_{i-1}, o_i] - \hat{\xi}[o_i; \theta])^2.
\tag{8}
$$

Rizoiu et al. (2017b) showed that the solution to Eq. (8) has good empirical performance. Conceptually, however, the approach has a few limitations:

L1: There is a mismatch between the objective and the observed data; the observations are of event *counts*, while the objective measures closeness to the event *rates*. Intuitively, one should work with $\mathbb{E}[N(t)]$ rather than $\mathbb{E}[\lambda(t)]$.

L2: The HIP objective is not derived as the likelihood of a particular process, raising uncertainty as to precisely what sense it approximates the Hawkes likelihood. In the limiting case where each observation interval is made infinitesimally small, for example, it is unclear whether the solution to Eq. (5) approaches that of the standard Hawkes likelihood.

L3: It relies on $\hat{\xi}$, which is derived from an approximation to the underlying integral equation. This approximation is of unclear quality, and cannot handle intervals of unequal length.

3. The Mean Behavior Poisson process (MBPP)

In this section, we first introduce the Mean Behavior Poisson process (MBPP) – a Poisson process whose intensity function is the expected event intensity of a Hawkes process of
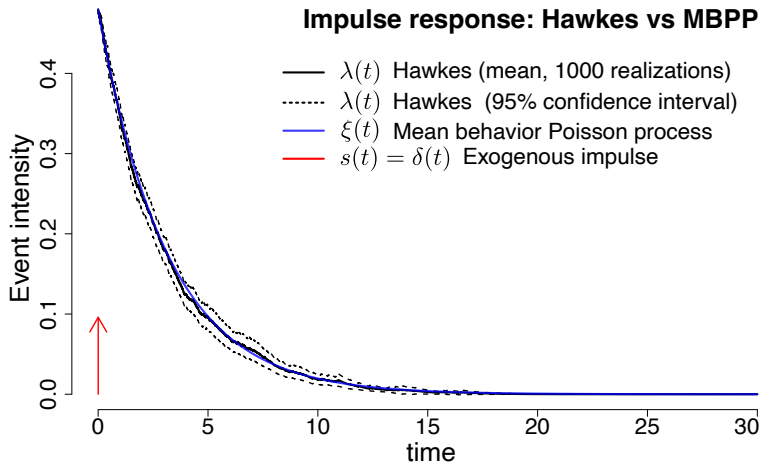
Figure 2: **Event intensity in Hawkes and MBPP** as a result of a single exogenous event (shown in red). We generate 1000 Hawkes realizations using an exponential kernel (Eq. (13)) and the parameters $\kappa = 0.9, \theta = 1.15$. We show the mean and the 95% confidence interval of the Hawkes intensity $\lambda(t)$ (in black), and the (deterministic) MBPP intensity $\xi(t)$ (in blue). We observe that the mean Hawkes intensity over the 1000 realizations follows closely the MBPP intensity, empirically showing the relation between the two processes.

the same parameters (Section 3.1). Next, we show that its compensator $\Xi(t)$ follows the same self-consistent equation as its intensity $\xi(t)$ (Section 3.2). Finally, we develop two methods to solve the self-consistent equation of MBPP for arbitrary exogenous functions (Section 3.3).

3.1 Mean Behavior Poisson: model definition

The intensity of a Hawkes process (defined in Eq. (3)) is a stochastic function – i.e., at the same future time t the conditional intensity of two Hawkes realizations with the same parameters can take very different values depending on the history of each realization. Here we introduce a new process — dubbed the *Mean Behavior Poisson process* (MBPP) — which captures the expected behavior of all the Hawkes realizations defined by the same set of parameters. Formally, given a Hawkes process we define MBPP as the non-homogenous Poisson process with a deterministic intensity $\xi(t)$ — the expected Hawkes intensity over all possible realizations (Eq. (6)):

$$\xi(t) \doteq \mathbb{E}_{\mathcal{H}_t}[\lambda(t)] = s(t) + \int_0^t \xi(\tau)\phi(t-\tau)\,\mathrm{d}\tau$$
$$= s(t) + (\xi * \phi)(t) \tag{9}$$

where $\lambda(t)$ denotes the event intensity in the Hawkes process (Eq. (3)); $s(t)$ denotes exogenous intensity function; $\phi(t)$ denotes the kernel function of the Hawkes process and $*$ is the convolution operator. For the remainder of the paper, when convenient, we suppress the functional dependence on parameters θ for intensity functions $(\lambda(t), \xi(t))$ and compensators $(\Lambda(t), \Xi(t))$.

10

We observe that the same objects fully define the MBPP intensity $\xi(t)$ as well as the Hawkes intensity $\lambda(t)$; that is, the exogenous intensity $s(t)$ and the kernel function $\phi(t)$. We, therefore, say that Hawkes processes have a one-to-one correspondence to MBPP. In Fig. 2 we visually denote the relation between the two processes by plotting the event intensity in Hawkes and in MBPP. We sample 10,000 realizations of Hawkes, and we plot its mean event intensity and its 95% confidence interval. We also plot the intensity of the mean behavior Poisson, which visually overlaps closely with the mean Hawkes intensity.

MBPP and HIP. Visibly, MBPP shares the same equation with HIP (Rizoiu et al., 2017b) (see Eq. (6)). However, their usages are different: HIP discretizes Eq. (6) and aims to fit volumes of events, while Eq. (9) defines the intensity of a novel point process that, as far as we are aware, has not been studied in the literature. In the next section, we study an important quantity of MBPP – its *compensator* – and we show that it satisfies a nearly identical integral equation to the intensity equation.

3.2 The compensator of MBPP

We define $\Xi(t)$ as the expectation of the Hawkes compensator (see Eq. (2)) over all possible realizations:

$$\Xi(t) = \mathbb{E}_{\mathcal{H}_t}\left[\Lambda(t)\right]$$

We show that the $\Xi(t)$ is the compensator of the MBPP using the Fubini theorem and passing the expectation inside the integral since $\lambda(t)$ is bounded (see Klebaner (2012, Theorem 8.4)):

$$
\begin{aligned}
\Xi(t) &= \int_0^t \mathbb{E}_{\mathcal{H}_t}\left[\lambda(z)\right]\,\mathrm{d}z \overset{cf.(9)}{=} \int_0^t \xi(z)\,\mathrm{d}z \\
&= \int_0^t \left[s(z) + \int_0^z \xi(z-y)\phi(y)\,\mathrm{d}y\right]\,\mathrm{d}z \\
&= \int_0^t s(z)dz + \int_{z=0}^t \int_{y=0}^z \xi(z-y)\phi(y)\,\mathrm{d}y\,\mathrm{d}z \\
&\overset{(a)}{=} \int_0^t s(z)dz + \int_{y=0}^t \int_{z=y}^t \xi(z-y)\phi(y)\,\mathrm{d}z\,\mathrm{d}y \\
&= \int_0^t s(z)dz + \int_0^t \phi(y)\int_y^t \xi(z-y)\,\mathrm{d}z\,\mathrm{d}y \\
&\overset{(b)}{=} \int_0^t s(z)dz + \int_0^t \phi(y)\int_0^{t-y} \xi(x)\,\mathrm{d}x\,\mathrm{d}y \\
&\overset{(c)}{=} S(t) + \int_0^t \Xi(t-y)\phi(y)\,\mathrm{d}y \\
&= S(t) + (\Xi * \phi)(t) \qquad\qquad\qquad\qquad\qquad (10)
\end{aligned}
$$

where (a) is obtained by reversing the order of the integrals and computing the new integration boundaries; (b) is obtained by performing the change of variable $x = z - y$; in (c) we apply $\Xi(t-y) \doteq \int_0^{t-y} \xi(x)\,\mathrm{d}x$; and $S(t) \doteq \int_0^t s(z)\,\mathrm{d}z$ is the cumulative immigrant intensity up to time t.

11

A cursory observation of Eq. (10) shows a distinct resemblance with Eq. (9): MBPP's compensator $\Xi(t)$ follows a very similar self-consistent equation as its intensity $\xi(t)$, with the compensator's value at time t being a function of its values at every previous time $\tau < t$, decayed by the corresponding kernel function value. The similarity in the functional form of the two equations implies that similar methods can be applied to deduce their analytical solutions, detailed in the next section.

Computational requirements. Both MBPP's compensator and Hawkes process intensity functions are quadratic to compute. The Hawkes process intensity function (unraveling the summation in Eq. (3)) is quadratic in the number of events in the history $\mathcal{H}_t$. The MBPP's compensator function is quadratic in the number of interval-censored intervals (see Section 4). However, calculating MBPP's compensator function is more computationally effective than computing the intensity function of a Hawkes process. Typically, the number of observational intervals is significantly smaller than the number of events.

3.3 Solving the Mean Behavior Poisson equation

To fit the MBPP (for both the event time and interval-censored setups), we need to have its intensity available in a tractable form. In this section, we propose two methods to solve Eq. (9), for arbitrary $s(t)$ and $\phi(t)$ functions. This type of equation is known as a Volterra integral equation of the second kind (Arfken, 1985), which admits a solution only in particular cases. The first method is based on the Laplace transform and has the downside of applying only to functions for which the transform and its inverse exist. The second method is a novel solution to the Volterra equation of the second kind, which employs notions of distribution theory first to compute the system's impulse response and afterward the solution.

Laplace transform. One natural approach to solving the MBPP equation is using the Laplace transform. Let $\mathcal{L}\{\cdot\}$ and $\mathcal{L}^{-1}\{\cdot\}$ denote the Laplace and the inverse of the Laplace transform respectively.

Theorem 1 *Given the functions $s(t)$ and $\phi(t)$, then*

$$\xi(t) = \mathcal{L}^{-1}\left\{\frac{\mathcal{L}\{s(t)\}(\omega)}{1 - \mathcal{L}\{\phi(t)\}(\omega)}\right\}(t) \tag{11}$$

is a closed-form solution for Eq. (9) if the Laplace and inverse Laplace transform exist.

The derivation of Eq. (11) is shown in Appendix A.

This Laplace transform-based method has several drawbacks. First, it requires that the Laplace transform exists for both $s(t)$ and $\phi(t)$. This requirement is not always fulfilled by the exogenous intensity functions $s(t)$ that we construct in Definition 12. These involve Dirac delta functions, for which the Laplace transform does not exist in the classical sense. Note that, while there are extensions for the Laplace transform to generalized functions (we refer the curious reader to van Dijk (2013, Ch. 8)), their direct application to solving Eq. (9) is not obvious, and they are outside the scope of this work.

Second, the method requires that the inverse Laplace function exists, which does not hold when ϕ involves a power-law or a Rayleigh time decay function. This is particularly limiting since the power-law kernel ($\phi(t) = \kappa(t + c)^{-(1+\theta)}$) has been shown to be the best

performing in applications involving social media data (Rizoiu et al., 2017b; Mishra et al., 2016); the Rayleigh function ($\phi(t) = \kappa \frac{t}{\theta^2} e^{-\frac{t^2}{2\theta^2}}$) is widely used in epidemics modeling as it allows for a period of increasing intensity corresponding to disease incubation (Unwin et al., 2021).

Lastly, this method involves re-computing Eq. (11) every time that $s(t)$ changes – with all the difficulties stated above occurring at each recomputation – rendering the method difficult to apply when the $s(t)$ changes often, or when it is not known before-hand.

The impulse response solution. We propose a novel method to solve the MBPP equation in Eq. (9), which deals with the shortcomings of the Laplace transform-based method. We first show that the Eq. (9) defines a linear, continuous-time, time-invariant system (Phillips et al., 2003), and we discuss a method to derive its impulse response solution (Theorem 2). Next, we show how to use the impulse response to easily compute the solution for any arbitrary $s(t)$ (Corollary 3).

Theorem 2 *Let H be a Hawkes process whose behavior is defined by the kernel function $\phi(t)$ satisfying $\int_0^\infty \phi(t)\mathrm{dt} < 1$, and let M be the mean behavior Poisson process (MBPP) associated (Eq. (9)) with H, with the event intensity denoted by $\xi(t)$, and defined in Eq. (9). Then the intensity of M defines a causal, linear, continuous-time, time-invariant system of input $s(t)$ and output $\xi(t)$. The system is completely characterized by its impulse response defined by*

$$E(t) = \delta(t) + h(t), \ \ where \ h(t) = \sum_{n=1}^{\infty} \phi^{\otimes n}(t)$$

where $\otimes n$ signifies n times convolution.

Proof We structure the proof into two parts. In the first part, we show that MBPP is an LTI system, and in the second part, we derive its impulse response.

For the first part of the proof, Rizoiu et al. (2017b, Corollary 2.2) have shown that Eq. (9) defines a causal, linear, continuous-time, time-invariant system of input $s(t)$ and output $\xi(t)$. For completeness reasons, we outline this proof here. It is easy to see that linearity holds by multiplying both sides of Eq. (9) by the same constant. The time-invariance property states that the response to a time-delayed input is identical and similarly time-delayed: if $s(t) \to \xi(t)$ then $s(t - t_0) \to \xi(t - t_0)$. Using Eq. (9) we compute the response for $s(t - t_0)$ as follows:

$$\xi(t - t_0) = s(t - t_0) + \int_0^t \xi(\tau)\phi(t - t_0 - \tau)\,\mathrm{d}\tau$$

$$\overset{(a)}{=} s(t') + \int_0^{t'+t_0} \xi(\tau)\phi(t' - \tau)\,\mathrm{d}\tau$$

$$\overset{(b)}{=} s(t') + \int_0^{t'} \xi(\tau)\phi(t' - \tau)\,\mathrm{d}\tau + \underbrace{\int_{t'}^{t'+t_0} \xi(\tau)\phi(t' - \tau)\,\mathrm{d}\tau}_{=0}$$

$$\overset{(c)}{=} s(t') + \int_0^{t'} \xi(\tau)\phi(t' - \tau)\,\mathrm{d}\tau$$

13

where at (a) we make a change of variable $t' = t - t_0$; at line (b) we write the integral into two parts, i.e., $(0, t)$ and $(t', t' + t_0)$; at line (c) we observe that $\phi(t)$ is a causal function, i.e., $\phi(t) = 0$ for $t < 0$, or $\phi(t' - \tau) = 0$ for $\tau > t'$, and the integral term from t' to $t' + t_0$ disappears. $\xi(t)$ is also a causal function, which together with the linearity and time-invariance properties renders the MBPP intensity system a causal, linear time-invariant (LTI) system.

In the second part of this proof, we concentrate on the impulse response function corresponding to the MBPP system. The impulse response of a dynamic system is its output when presented with a unit impulse input signal. The impulse response completely characterizes the system, as the response to a linear combination of time-delayed impulses is a linear combination of time-delayed impulse responses. For continuous time dynamic systems, the impulse is denoted by the Dirac delta function $\delta(t)$. We denote as $E(t)$ the impulse response of MBPP, with $E(t) = \delta(t) + h(t)$. From Eq. (9), it follows that

$$
\begin{aligned}
E(t) &= \delta(t) + \int_0^t E(\tau)\phi(t - \tau)\,\mathrm{d}\tau \\
&= \delta(t) + (E * \phi)(t) \\
&\overset{(a)}{=} \delta(t) + ((\delta + h) * \phi)(t) \\
&= \delta(t) + (\delta * \phi)(t) + (h * \phi)(t) \\
&\overset{(b)}{=} \delta(t) + \underbrace{\phi(t) + (h * \phi)(t)}_{h(t)}
\end{aligned}
$$

where (a) is obtained by substituting $E(t)$ with $\delta(t) + h(t)$; (b) is obtained by noticing that the Dirac δ function is the neutral element for the convolution operation. Consequently and by dropping the time notation, we obtain

$$
\begin{aligned}
h &= \phi + h * \phi \\
&= \phi + (\phi + h * \phi) * \phi \\
&= \phi + \phi^{\otimes 2} + h * \phi^{\otimes 2} \\
&= \phi + \phi^{\otimes 2} + \phi^{\otimes 3} + h * \phi^{\otimes 3} \\
&= \ldots \\
&= \sum_{n=1}^{\infty} \phi^{\otimes n}.
\end{aligned}
$$

For the above equation to hold, the infinite sum must converge. That is, we require that $\lim_{n \to \infty} \phi^{\otimes n} = 0$. Observe that

$$
\begin{aligned}
\lim_{n \to \infty} \int_0^{\infty} \phi^{\otimes n}(t)\mathrm{d}t &\overset{(a)}{\leq} \lim_{n \to \infty} \left(\int_0^{\infty} \phi(t)\mathrm{d}t \right)^n \\
&\overset{(b)}{=} 0,
\end{aligned}
$$

where for (a) we use Young's convolution inequality on $\phi(t)$ under the L^1 norm $\int_0^{\infty}(\cdot)\mathrm{d}t$, and for (b) we use $\int_0^{\infty} \phi(t)\mathrm{d}t < 1$. Since $\lim_{n \to \infty} \int_0^{\infty} \phi^{\otimes n}(t)\mathrm{d}t \geq 0$, we have $\lim_{n \to \infty} \int_0^{\infty} \phi^{\otimes n}(t)\mathrm{d}t = 0$.

By L^1 convergence, we then have $\lim_{n\to\infty} \phi^{\otimes n} = 0$. ■

Corollary 3 *Let M be a mean behavior Poisson process associated with a Hawkes process of kernel $\phi(t)$, and let $E(t) = \delta(t) + h(t)$ be its impulse response function. Let $s(t)$ be an arbitrary generalized function denoting an external stimuli. Then $\xi(t) = (E * s)(t)$ is a solution to the MBPP equation (Eq. (9)).*

Proof

$$\xi = E * s = (\delta + h) * s = \delta * s + h * s$$
$$\overset{(a)}{=} s + E * \phi * s = s + (E * s) * \phi$$
$$= s + \xi * \phi$$

where (a) is obtained by replacing $h = E * \phi$, knowing that $E = \delta + h$ and that E is the impulse response, i.e. a solution to $E = \delta + E * \phi$. ■

For example. Suppose
$$s(t) = \kappa\theta + (u_0 - \kappa\theta)e^{-\theta t} \tag{12}$$
and $\phi(t)$ is exponentially time-decaying. These particular forms for the $s(t)$ and $\phi(t)$ were proposed by Dassios and Zhao (2013), for which they also find a closed-form solution. We define the exponential kernel as
$$\phi(t) = \kappa\theta e^{-\theta t} \cdot [\![t > 0]\!]. \tag{13}$$

Following Theorem 2 we first compute the impulse response as $E(t) = \delta(t) + h(t)$, with $h \in L^1(\mathbb{R})$ given by $h = \sum_{n=1}^{\infty} \phi^{\otimes n}$. For $t \geq 0$ and ϕ defined in Eq. (13) we have
$$\phi^{\otimes 2} = (\phi * \phi)(t)$$
$$= \int_0^t \kappa\theta e^{-\theta\tau} \kappa\theta e^{-\theta(t-\tau)} \, d\tau$$
$$= \kappa^2\theta^2 e^{-\theta t} t$$

and one can easily show that
$$\phi^{\otimes n}(t) = \kappa^n \theta^n e^{-\theta t} \frac{t^{n-1}}{(n-1)!} \tag{14}$$

By summing up, we compute $h(t)$ as
$$h(t) = \sum_{n=1}^{\infty} \phi^{\otimes n}(t) = e^{-\theta t} \kappa\theta \sum_{n=0}^{\infty} \frac{(\kappa\theta t)^n}{n!}$$
$$= e^{-\theta t} \kappa\theta e^{k\theta t}$$
$$= \kappa\theta e^{(\kappa-1)\theta t}, \text{ for } t \geq 0$$

15

Since $\phi(t < 0) = 0$, we have $h(t < 0) = 0$. It follows that the impulse response for MBPP with the exponential kernel defined in Eq. (13) is

$$E(t) = \delta(t) + \kappa\theta e^{(\kappa-1)\theta t} \cdot [\![t > 0]\!] \tag{15}$$

From Corollary 3 it immediately follows that the solution for Eq. (9) with $s(t) = \kappa\theta + (u_0 - \kappa\theta)e^{-\theta t}$ is

$$\begin{aligned}
\xi(t) &= (E * s)(t) = s(t) + (h * s)(t) \\
&= \frac{\kappa\theta}{1-\kappa}\left(1 - e^{-(1-\kappa)\theta t}\right) + u_0 e^{-(1-\kappa)\theta t}
\end{aligned} \tag{16}$$

The full details of the derivation (and derivations of other exogenous functions) are shown in Appendix B.1.

Note that the infinite convolutions sum has a closed-form solution only for particular kernel functions (such as in Eq. (14)). For all the other kernels, one can apply an approximation based on finite sums. For example, knowing that $\lim_{n\to\infty} \phi^{\otimes n} = 0$ when $\int_0^\infty \phi(t)\mathrm{d}t < 1$ (see above), we can use the approximation $h \approx \sum_{n=1}^{T} \phi^{\otimes n}$, where T controls the approximation error and the computation speed. Note also that an analytical solution for the convolution $h * h$ does not exist for most kernels, requiring numerical solutions.
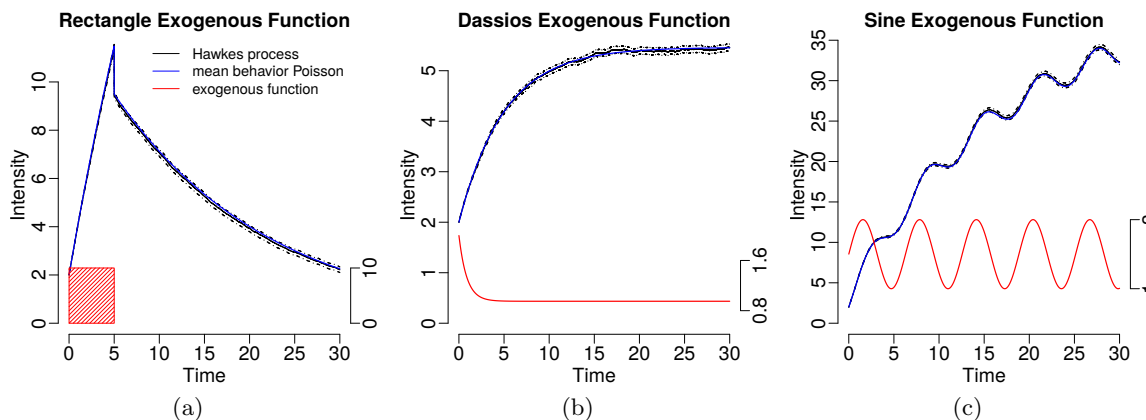

Figure 3: The intensity of 10,000 simulated realizations of a Hawkes process with exogenous function as a rectangle response, Dassios response (Dassios and Zhao, 2013), and sine response. Here the red part is the exogenous function both in the Hawkes process and MBPP. The blue line is the MBPP intensity, the solid black line is the mean intensity of the corresponding Hawkes process, and the black dashed lines indicate the 95% confidence interval.

4. Interval-censored point processes

In this section, we discuss fitting MBPP in the interval-censored setting (detailed in Section 4.1), also known in some areas of the literature as panel count data (Sun and Zhao,

2013; Ding et al., 2018; Moreno et al., 2020). The Hawkes process generally cannot be fit in the interval-censored setting as it does not have the *independent increment* property. Given the correspondence in parameters between MBPP and the Hawkes process, we can instead fit an MBPP using the interval-censored log-likelihood (IC-LL) as the MBPP does have the independent increment property. Finally, we present a method for numerically calculating the MBPP compensator function, which is needed to calculate the IC-LL loss function.

4.1 Interval-censored Hawkes processes

Suppose that we have the set of observation times $O = \{o_i \mid i = 0, \ldots, m, \, o_i \leq T\}$ which partitions into non-overlapping segments the time extent that the point process is observed $t \in (0, T]$. Without loss of generality, we assume that the time points in O are ordered by index, $o_0 = 0$, and $o_m = T$. Further, for each segment defined by the half-open interval of consecutive points $(o_{i-1}, o_i]$ we observe the number of events which occurs in the interval, denoted by $\mathsf{C}(o_{i-1}, o_i]$. We define this setting as the *interval-censored setting*. Fig. 4 illustrates the interval-censored setting for a simulated Hawkes process realization. Panel A shows the realization as event times; panel D shows the same Hawkes process realization as interval-censored.

If we want to estimate the parameters of a Hawkes process in this setting, we cannot use the standard point process log-likelihood in Eq. (4) as it is defined for event times. Instead, we aim to maximize the joint probability of the observed event volumes in each interval:

$$\max_{\theta} \mathbb{P}(N(o_{i-1}, o_i] = \mathsf{C}(o_{i-1}, o_i], i \in \{1, \ldots, m\}), \tag{17}$$

where $N(o_{i-1}, o_i] = N(o_i) - N(o_{i-1})$ is defined as the number of events in a realization on the half-open interval and θ are the Hawkes process's parameters.

However, there are several complications for computing Eq. (17) for Hawkes processes. First, unlike the homogeneous and homogeneous Poisson processes, the Hawkes process does not have the independent increments property. As a result, we cannot evaluate Eq. (17) as we did in Eq. (1), i.e., the product of disjoint Poisson distributions with the expected number of events set to the point process compensator. An alternative method for breaking down the Hawkes process is to decompose the process into the event count of each generation of offspring, formally known as a Galton-Watson branching process (Hawkes and Oakes, 1974). This provides a decomposition of Borel distributions (Borel, 1942). Kong et al. (2020b) derive the closed-form solution for the case where the offspring distribution is evaluated at $t \to \infty$. However, there is no known analytical solution for computing the Borel distributions at finite times. A Monte-Carlo solution is available, but the computation cost does not scale well (O'Brien et al., 2020). Further, there is no known analytic or numeric solution for computing the conditional factorization of the joint probability of all offspring generation Borel distributions. Thus, there is no closed-form solution for a Hawkes process likelihood function in the interval-censored setting in the general case.

Using the MBPP equivalence. Given the above difficulties, we propose in this section a solution based on the MBPP. In a nutshell, we replace the Hawkes process with an MBPP of equivalent parameters and fit the MBPP parameters in the interval-censored setting. Finally, we use the obtained MBPP parameters to approximate the Hawkes parameters. Our approach has two sources of information loss (that we study on synthetic
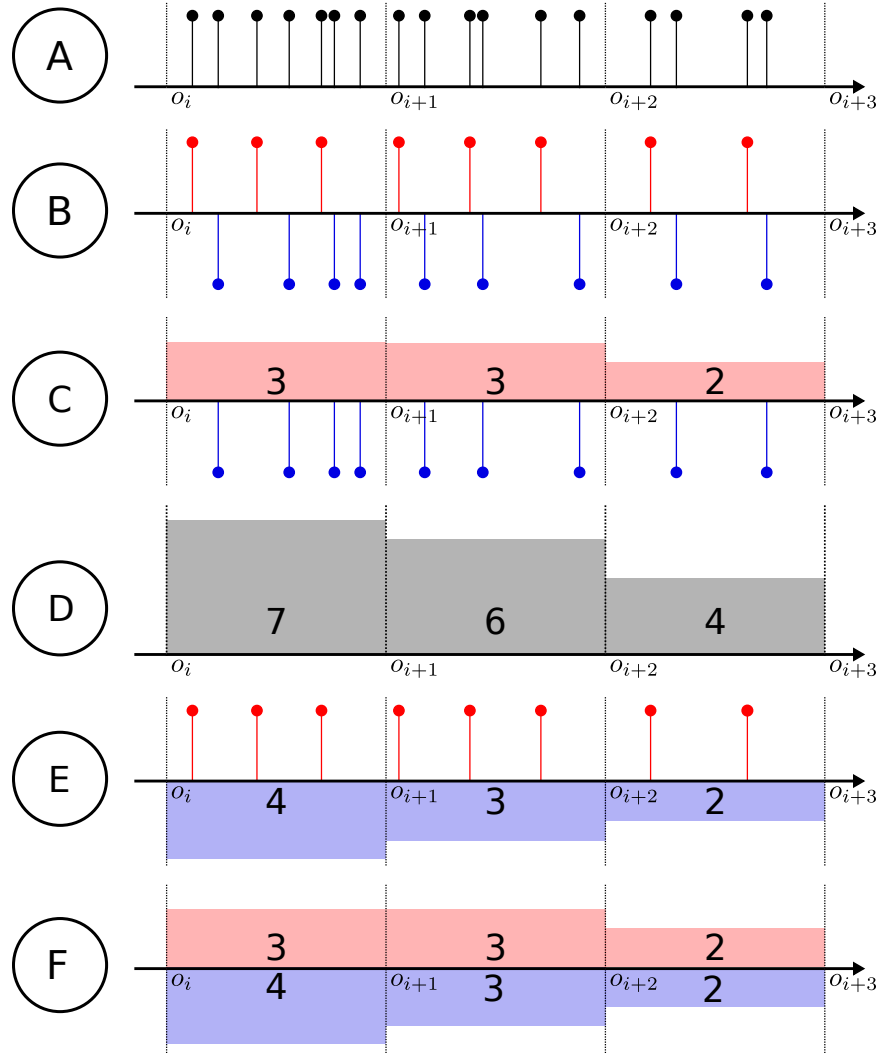
Figure 4: **The same point process realization is shown in six scenarios of data presentation – from event times, to separable event times and interval censored data.** Each row corresponds to a setting in Table 2. Red components correspond to exogenous events; and blue components correspond to endogenous events. Black color indicates that we cannot distinguish between exogenous and endogenous. o_i denote the observation times based on which the data is interval-censored. The lollipops markers represent event times, while the shaded areas represent event counts within intervals.

data in Section 7), which prevent obtaining the exact original Hawke parameters. First, the information contained in the interval-censored version of the data – when the observation periods are not infinitesimal in size – will necessarily have less information than the original event time data. Second, while having parameter equivalence, MBPP and Hawkes are different models – for example, MBPP is not a self-exciting model. As a result, the fitted MBPP parameters can only be approximations of the true Hawkes parameters.

4.2 Negative Poisson log-likelihood: Interval-censored loss

Unlike the Hawkes process, the MBPP introduced in Section 3 is an non-homogeneous Poisson process. Thus the MBPP has the independent increment property, allowing Eq. (17) to be evaluated as the likelihood of disjoint Poisson distributions (Daley and Vere-Jones, 2003, Chapter 2.4). The corresponding negative log-likelihood function of the MBPP is dubbed as the *interval-censored log-likelihood (IC-LL)*. The IC-LL was used with non-homogeneous Poisson processes by prior literature (Ding et al., 2018). In this work, we use IC-LL as a building block to estimate the parameters of Hawkes models (via MBPP) and to build new loss functions via the Bregman divergence interpretation (see Section 5.1).

The Poisson distributions which compose the negative Poisson log-likelihood are defined with respect to the MBPP compensator function $M(t)$, as per Eq. (1). Similar to the counts $N(t)$, define $M(o_{i-1}, o_i] = M(o_i) - M(o_{i-1})$. Note that $M(t)$ is stochastic, as it is the counting process of the MBPP. Using this joint distribution, we can calculate the negative log-likelihood for an MBPP by decomposing over each observation interval.

$$
\begin{aligned}
\mathscr{L}(\theta) &= -\log \mathbb{P}(M(o_{i-1}, o_i] = \mathsf{C}(o_{i-1}, o_i], \, i = 1, \ldots, m) \\
&= -\log \prod_{i=1}^{m} \mathbb{P}(M(o_{i-1}, o_i] = \mathsf{C}(o_{i-1}, o_i]) \\
&= -\log \prod_{i=1}^{m} \left(\frac{\Xi(o_{i-1}, o_i]^{\mathsf{C}(o_{i-1}, o_i]}}{\mathsf{C}(o_{i-1}, o_i]!} \exp(-\Xi(o_{i-1}, o_i]) \right) \\
&= -\sum_{i=1}^{m} \left(\mathsf{C}(o_{i-1}, o_i] \cdot \log \Xi(o_{i-1}, o_i] - \log(\mathsf{C}(o_{i-1}, o_i]!) - \Xi(o_{i-1}, o_i] \right) \\
&= \sum_{i=1}^{m} \Xi(o_{i-1}, o_i] - \sum_{i=1}^{m} \mathsf{C}(o_{i-1}, o_i] \log \Xi(o_{i-1}, o_i] + \sum_{i=1}^{m} \log(\mathsf{C}(o_{i-1}, o_i]!).
\end{aligned}
\tag{18}
$$

Notably, as we are aiming to optimise a parameter set θ of MBPP, the constant which only depends on the event counts $\sum_{i=1}^{m} \log(\mathsf{C}(o_{i-1}, o_i]!)$ can be ignored. As such, we further rewrite the above loss function to the following equivalent loss function when minimised.

Proposition 4 *The interval-censored log-likelihood (IC-LL) for an MBPP with corresponding compensator Ξ under interval-censored data is*

$$
\mathscr{L}_{\text{IC−LL}}(\theta) = \sum_{i=1}^{m} \Xi(o_{i-1}, o_i; \theta] - \sum_{i=1}^{m} \mathsf{C}(o_{i-1}, o_i] \log \Xi(o_{i-1}, o_i; \theta].
\tag{19}
$$

Note that partial similar results exist in literature. For example, (Ding et al., 2018, Sec. 2.1) derive a similar log-likelihood of a multivariate non-homogeneous point process. In this work, we further link IC-LL with the Bregman divergence (see Section 5.2), and we show how other loss functions can be obtained by varying the generator function.

4.3 Calculating the negative Poisson log-likelihood

The loss functions proposed above both require the computation of the compensator $\Xi(o_{i-1}, o_i]$, as defined in Eq. (10). From the definition of the compensator over an interval, we have

$$
\begin{aligned}
\Xi(o_{i-1}, o_i] &\doteq \int_{o_{i-1}}^{o_i} \xi(s)\, ds \\
&= \int_0^{o_i} \xi(s)\, ds - \int_0^{o_{i-1}} \xi(s)\, ds \\
&= \Xi(o_i) - \Xi(o_{i-1}).
\end{aligned}
\tag{20}
$$

We explore the two cases where we either have an analytical solution for the MBPP compensator or when we have to use numeric approximation to evaluate it.

Analytical solution. As we have foreshadowed in Section 3.2, the compensator of MBPP follows the same functional form as the MBPP intensity function ξ. Thus, we can use the same analytical methods for solving the compensator evaluated at a point. Specifically, we can apply Corollary 3 to derive an analytical solution:

Corollary 5 (to Theorem 2) *Let $\phi(t)$ be the kernel of a Hawkes process and $E(t) = \delta(t) + h(t)$ be its impulse response function, then for $S(t)$ a closed-form solution for the MBPP compensator over a half-open interval $(x, y]$ is*

$$
\Xi(x, y] = (E * S)(y) - (E * S)(x).
\tag{21}
$$

Proof Follows immediately from application of Eq. (20) and Corollary 3. ∎

Numerical approximation. The cases in which the compensator $\Xi(o_{i-1}, o_i]$ cannot be solved analytically are generally due to the lack of a closed-form solution for the h term in Theorem 2. While we can approximate h using a finite sum of numerical convolutions (see discussion in Section 3.3), here we consider a numeric lower bound approximation of the MBPP compensator – see discussion at the end of this section for additional advantages. Specifically, we approximate the MBPP counting process with a set of approximation points $\{d_j\}_{j=0}^D$, where $d_0 = 0$, $d_{j-1} < d_j$, and $0 < d_j < T$ for $j \in \{1, \ldots, D\}$. Consider the following lower-bound on the MBPP counting process using D approximation points:

$$
M(t) \geq M_D^-(t) \doteq \sum_{j=1}^D M(d_{j-1}, d_j] \cdot [\![d_j < t \leq d_D]\!].
\tag{22}
$$

Considering that the event count $M(t)$ is a monotonically increasing function, the lower bound $M_D^-(t)$ is intuitively a piecewise constant function, which gets updated at each d_j, and is constant in between approximation points. For equidistant approximation points

$$
d_j = \frac{T}{D} i,
$$

it can be shown that we can become arbitrarily accurate to $M(t)$ as the number of approximation points increases.

Proposition 6 *As $D \to \infty$, then $M_D^-(t) \to M(t)$.*

Using this lower bound counting process, we can approximate the compensator of MBPP, which follows a self-consistent equation (as per Eq. (10)). We consider the following formulation of the compensator which uses the equivalence between compensator and expected counts $\Xi(t) = \mathbb{E}[M(t)]$ for non-homogeneous Poisson processes:

$$
\begin{aligned}
\Xi(t) &= S(t) + \int_0^t \Xi(t - y)\phi(y)\mathrm{d}y \\
&= S(t) + \int_0^t \Xi(y)\phi(t - y)\mathrm{d}y \\
&= S(t) + \int_0^t \mathbb{E}[M(y)]\phi(t - y)\mathrm{d}y.
\end{aligned}
\tag{23}
$$

We now define the following lower-bound compensator using the counting process approximation $M_D^-(t)$:

$$
\begin{aligned}
\Xi^-(t) &\doteq S(t) + \int_0^t \mathbb{E}[M_D^-(y)]\phi(t - y)\mathrm{d}y \\
&= S(t) + \int_0^t \mathbb{E}\left[\sum_{j=1}^D M(d_{j-1}, d_j] \cdot [\![d_j < y \le d_D]\!]\right]\phi(t - y)\mathrm{d}y \\
&= S(t) + \sum_{j=1}^D \mathbb{E}\left[M(d_{j-1}, d_j]\right]\int_0^t [\![d_j < y \le d_D]\!]\phi(t - y)\mathrm{d}y \\
&\overset{(a)}{=} S(t) + \sum_{j=1}^{\bar{D}(t)} \mathbb{E}\left[M(d_{j-1}, d_j]\right]\int_{d_j}^{\min(t, d_D)}\phi(t - y)\mathrm{d}y,
\end{aligned}
\tag{24}
$$

where $\bar{D}(t) = \max\{i \in \{1, \ldots, D\} : d_i < t\}$. Step (a) follows from simultaneously changing the integral upper-bound and the summation bound with respect to the Iverson brackets.

An upper-bound approximation equivalent is given by

$$
\Xi^+(t) = S(t) + \sum_{j=1}^{\bar{D}(t)+1} \mathbb{E}\left[M(d_{j-1}, d_j]\right]\int_{d_{j-1}}^{\min(t, d_D)}\phi(t - y)\mathrm{d}y.
\tag{25}
$$

The negative Poisson log-likelihood loss function requires to compute the compensator over intervals $\Xi(o_{i-1}, o_i]$, which we approximate using the compensator lower-bound Ξ_D^-:

$$
\Xi(o_{i-1}, o_i] \approx \Xi_D^-(o_{i-1}, o_i] = \Xi_D^-(o_i) - \Xi_D^-(o_{i-1}).
\tag{26}
$$

By using Eq. (24), we can compute this approximation as follows,

Proposition 7 *The numeric approximation of the compensator over an interval* $\Xi(o_{i-1}, o_i] \approx \Xi_{\bar{D}}^{\bar{=}}(o_{i-1}, o_i]$ *can be computed as*

$$
\begin{aligned}
\Xi_{\bar{D}}^{\bar{=}}(o_{i-1}, o_i] = S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{o_{i-1}-d_j}^{o_i-d_j} \phi(y) \mathrm{d}y \\
+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_0^{o_{i-1}-d_j} \phi(y) \mathrm{d}y.
\end{aligned}
\tag{27}
$$

The proof of Proposition 7 can be found in Appendix A.

Leveraging Proposition 7 has several advantages. First, it allows approximating the compensator's value for functional forms that do not allow for closed-form solutions of the infinite convolution proposed in Section 3.3. Second, it allows leveraging observed history (the number of observed events in past time intervals) to compute future compensator values. We use the latter property in our real-world experiments in Section 8.

5. Bregman generalization of IC-LL

Previously in Section 4, we introduced the NPLL loss function for fitting in interval-censored data settings. This section shows that the IC-LL loss function can be interpreted as a Bregman divergence– a generalized notion of distances. We consider alternative loss functions that arise when changing the convex generator function of the Bregman divergence. This leads to a generalization and correction of the HIP loss function (Rizoiu et al., 2017b). In particular, we present specific conditions in which the generalization simplifies to the HIP loss function.

5.1 Bregman divergences

We first begin with a brief introduction of Bregman divergences. Bregman divergences are a widespread tool for designing and analyzing algorithms within machine learning. Bregman divergences have been used to construct generalizations of algorithms to use different definitions of distances, i.e., clustering (Banerjee et al., 2005), boosting (Collins et al., 2002), and computational geometry (Nielsen et al., 2007). Similarly, we will use Bregman divergences to generalize IC-LL by interpreting the loss function as the generalized Kullback-Leibler divergence (KL-divergence) between compensator values and interval-censored event counts.

A Bregman divergence is a generalized notion of distances between points. Suppose we are given a strictly convex generator $\varphi : S \to \mathbb{R}$, where $\mathrm{dom}(\varphi) = S$ is a convex set $S \subseteq \mathbb{R}^d$ and $\varphi(\cdot)$ is differentiable on the relative interior $\mathrm{ri}(S)$. Then the Bregman divergence $B_\varphi : S \times \mathrm{ri}(S) \to [0, \infty)$ is defined as (Banerjee et al., 2005):

$$
B_\varphi(\mathbf{x}, \mathbf{y}) \doteq \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^\top \nabla \varphi(\mathbf{y}).
\tag{28}
$$

A wide selection of popular divergences can be obtained by picking the appropriate generator function. In particular, we consider two specific convex generators for Bregman divergences.

KL-divergence. The KL-divergence can be obtained by considering the generator function

$$\varphi(\mathbf{x}) \doteq \sum_{i=1}^{d} x_i \log x_i, \tag{29}$$

yielding, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^+$,

$$\mathrm{KL}(\mathbf{x}, \mathbf{y}) \doteq B_\varphi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i} - \mathbf{1}^\top (\mathbf{x} - \mathbf{y}). \tag{30}$$

Squared loss. The squared loss function is obtained if we consider a squared Euclidean norm generator function

$$\hat{\varphi}(\mathbf{x}) \doteq \|\mathbf{x}\|^2, \tag{31}$$

yielding

$$\mathrm{SSE}(\mathbf{x}, \mathbf{y}) \doteq B_{\hat{\varphi}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2. \tag{32}$$

5.2 Connections to IC-LL

Given a single observation sequence, the IC-LL loss function is equivalent to the generalized KL-divergence with an additional constant factor. As a result, IC-LL is equivalent to the KL-divergence under minimization. First, let us define $\mathbf{C} \doteq (\mathsf{C}(o_0, o_1], \ldots, \mathsf{C}(o_{m-1}, o_m])$ and $\mathbf{\Xi}(\theta) \doteq (\Xi(o_0, o_1], \ldots, \Xi(o_{m-1}, o_m])$ as an observed volume of interval-censored events and corresponding interval compensator values using parameters θ, respectively.

Starting from a simple observation, we express the IC-LL using the KL-divergence.

Proposition 8 *The IC-LL loss function is the given by*

$$\mathscr{L}_{\mathrm{IC-LL}}(\theta) = \mathrm{KL}(\mathbf{C}, \mathbf{\Xi}(\theta)) + \Gamma(\mathbf{C}), \tag{33}$$

where $\Gamma(\mathbf{C})$ is a constant only dependent on the observed events. Furthermore,

$$\arg\min_\theta \mathscr{L}_{\mathrm{IC-LL}}(\theta) = \arg\min_\theta \mathrm{KL}(\mathbf{C}, \mathbf{\Xi}(\theta)). \tag{34}$$

The KL-divergence term in Eq. (33) can be rewritten as a Bregman divergence as follows:

$$B_\varphi(\mathbf{C}, \mathbf{\Xi}(\theta)) + \Gamma(\mathbf{C}), \tag{35}$$

where φ is given by Eq. (29). Thus the natural question arises of what happens when we switch the choice of Bregman divergence.

We now consider switching the KL-divergence generator function to that of the squared loss function given by Eq. (31). This gives a sum of the squared error loss function:

$$\begin{aligned} \mathscr{L}_{\mathrm{SSE}} &= \mathrm{SSE}(\mathbf{c}, \mathbf{\Xi}) + \Gamma(\mathbf{C}) \\ &= \sum_{i=1}^{m} (\mathsf{C}(o_{i-1}, o_i] - \Xi(o_{i-1}, o_i; \theta])^2 + \Gamma(\mathbf{C}). \end{aligned} \tag{36}$$

This SSE loss function is actually a generalized version of the HIP loss function (Rizoiu et al., 2017b); differing in only that the intensity function ξ is used instead of Ξ and the additional constant $\Gamma(\mathbf{C})$.

A notable connection is the bijection between (regular) Bregman divergences and (regular) exponential families (Banerjee et al., 2005, Theorem 6). In particular, the choice of Bregman divergence dictates what distribution the data is assumed to be sampled from for each observational interval. For example, by choosing the KL-divergence Eq. (33), each observational interval is assumed to be Poisson distributed, giving the IC-LL loss function. By considering SSE divergence Eq. (36), we are assuming that the event volume is Gaussian distributed; with constant standard deviation.

One way of establishing a connection between the two loss functions is by considering the SSE loss function as an approximation of the KL-divergence, where each interval is Poisson distributed. Concretely, let us define a Poisson distribution with true rate $\Xi(o_{i-1}, o_i]$ for an arbitrary event interval. Then suppose we have N many i.i.d. volume observations over this interval $\mathsf{C}_1(o_{i-1}, o_i], \ldots, \mathsf{C}_N(o_{i-1}, o_i]$. It follows that given the empirical mean of these observations $\hat{\mathsf{C}}(o_{i-1}, o_i]$, by the central limit theorem, in the limit we have that

$$\hat{\mathsf{C}}(o_{i-1}, o_i] \sim \mathsf{N}(\Xi(o_{i-1}, o_i], \Xi(o_{i-1}, o_i]/N)$$

as a Poisson distribution has variance equal to its rate. Thus, the averaging of event sequence provides an SSE loss function by the bijection between Bregman divergences and exponential families (Banerjee et al., 2005, Theorem 6). Consequently, we expect the KL loss and SSE loss functions to perform similarly when the number of samples is large.

Notably, optimal compensator parameters for our Bregman loss functions are sensitive to the generator function chosen. This contrasts the unconstrained minimization of the expectation of Bregman divergence, which was shown to have a unique minimum invariant of the generator function (Banerjee et al., 2005, Proposition 1). Thus the functional form of the compensator (the chosen triggering kernel) can be considered an additional constraint to the Bregman divergence minimization problem.

5.3 HIP as an MBPP approximation

We note that the SSE loss function given by Eq. (36) is similar to the HIP loss function Rizoiu et al. (2017b). This section shows that the SSE loss function is a generalization of the HIP loss function, which is not theoretically justified in the general case. In particular, it incorrectly attempts to match the point process intensity with interval-censored volumes. However, by using the framework of Bregman divergence losses presented in Section 5.1, we prove that the HIP loss function is equivalent to an SSE loss function with three additional assumptions: (1) that observation intervals are unit sized; (2) the MBPP intensity function is constant over the interval; and (3) a discrete convolution approximation is used on the intensity function.

One simple, practical method of using Hawkes processes with interval-censored data is to convert intervals into event times through uniformly sampling within each interval. The average intensity (over multiple samples of intervals) would amount to the MBPP intensity function (by Eq. (9)). In essence, this simplification assumes that events are equally likely in the intervals; thus, the simplification is equivalent to assuming that the underlying MBPP intensity function is constant over each interval.

In addition, by assuming that we have unit time observation intervals, we can show that the compensator function over that interval is equivalent to the intensity function of the MBPP process.

Lemma 9 *Suppose that the MBPP intensity function is constant over unit interval $(i-1, i]$, where $i \in \mathbb{Z}$. Then $\Xi(i-1, i] = \xi(i)$.*

Proof As the MBPP intensity function is constant over the interval, let c the value it obtains. Thus,

$$\Xi(i-1, i] = \int_{i-1}^{i} \xi(s) \, \mathrm{d}s = \int_{i-1}^{i} c \, \mathrm{d}s = c = \xi(i).$$

■

Lemma 9 shows that under the specific conditions we have specified, the MBPP intensity function can be considered instead of the MBPP compensator — which can be cumbersome to calculate/approximate for certain kernel functions chosen (as shown in Section 4.3). If, in addition, we use the SSE loss function, we recover the exact HIP loss function.

Theorem 10 *Suppose that observation intervals are unit length with constant MBPP intensity. Then the SSE loss function in Eq. (36) with the MBPP intensity approximated with discrete convolution recovers the HIP loss function (Rizoiu et al., 2017b, Eq. (6)), i.e.,*

$$\frac{1}{2} \sum_{i=1}^{t} (\xi(i) - \mathsf{C}(i-1, i])^2. \tag{37}$$

Proof Given the observation intervals are unit length and Lemma 9, the SSE loss function in Eq. (36) is given by

$$\mathrm{SSE}(\mathbf{c}, \mathbf{\Xi}) = \sum_{i=1}^{t} (\mathsf{C}(i-1, i] - \Xi(i-1, i])^2 = \sum_{i=1}^{t} (\mathsf{C}(i-1, i] - \xi(i))^2,$$

which when minimised, is equivalent to Eq. (37). ■

With the HIP loss function recovered, to recover the HIP model we consider an MBPP with power-law kernel function. Additionally, we also approximate the intensity function by discrete convolution. This approximation step changes the integral in Eq. (9) to a summation

$$s(i) + \int_{0}^{i} \xi(i)\phi(t - \tau) \, \mathrm{d}\tau \approx s[i] + \sum_{\tau=1}^{i} \xi[i - \tau]\phi[\tau], \tag{38}$$

where the square brackets indicate that the function is constant over the interval $(i-1, i]$.

Leaving the choice of kernel aside, the approximation via discrete convolution provides an MBPP intensity function which adheres to the assumption in Theorem 10. Thus the entire HIP model can be realised using the loss function framework proposed in this section, generalising Eq. (7).

Theorem 11 *Suppose that we have unit length interval-censored event volumes. Suppose that we have an MBPP with power-law kernel and that the MBPP intensity is approximated with a discrete convolution, Eq. (38). Then by using the SSE loss function, Eq. (36), we have the HIP model from Rizoiu et al. (2017b).*

Proof An MBPP with power-law kernel and discrete convolution approximation is equivalent to the MBPP introduced in (Rizoiu et al., 2017b, Eq. (4)).

By definition of the discrete convolution in Eq. (38), the intensity function is constant over each interval $(i-1, i]$. Thus by Theorem 10, the SSE loss function (which uses the MBPP compensator) is equivalent to the HIP loss function (Rizoiu et al., 2017b, Eq. (6)). ■

Theorem 11 shows that our framework can generalise the method proposed by Rizoiu et al. (2017b). In particular, we can consider different kernel functions or use the compensator approximation method in Section 4.3 when observation intervals are not uniformly unit length. Importantly, using this framework, we can provide a closed-form solution of HIP when the power-law kernel is replaced with the exponential kernel — as the discrete convolution and uniform unit length observations are no longer required to approximate the MBPP compensator.

6. Processes with observed exogenous stimuli

In Sections 3 to 5, we used the underlying assumption that the exogenous events of the point process – i.e., the events not generated through self-excitation – are solely determined by a corresponding exogenous intensity function $s(t)$. Furthermore, we did not differentiate between exogenous and endogenous events. However, in many real-world scenarios (including our real-world experiments in Section 8), the exogenous and endogenous are both separable and directly observed; while the underlying functional form of $s(t)$ is unknown. We denote the setting in which the exogenous events can be differentiated as a *separable scenario*. Furthermore, in the separable setting, the exogenous and endogenous events can be observed at different granularity levels (event time vs. interval-censored). For example, consider the problem of predicting flu trends using Twitter data (Achrekar et al., 2011). The tweets are observed as event times, while flu cases are recorded as aggregated volumes (for patient privacy reasons (Rizoiu et al., 2016)).

Table 2 lists the six possible scenarios in which a point process can be observed. Fig. 4 exemplifies how the same realization is presented in each of these scenarios. Scenarios A and D are not separable, and events are observed as event times (ET) or interval-censored (IC), respectively (shown in black color). Scenarios B, C, E, and F are separable and define all the possible combinations of ET and IC for exogenous and endogenous events. For these separable scenarios, Fig. 4 exemplifies how the exogenous events (upper red events) and endogenous events (lower blue events) can be observed with either one or both types of events as interval-censored.

This section introduces two exogenous functions to approximate $s(t)$, which allow us to characterize the exogenous contribution of a Hawkes process, or MBPP, in the separable scenarios. We introduce the *multi-impulse exogenous function* and the *latent homogeneous Poisson process exogenous function* which can be used when the exogenous data is observed

Table 2: Tabulation of possible data scenarios. ET = "event times" and IC = "interval-censored". The last column specifies if an endogenous loss function can be used.

Scenario	Setting		Point Process				
	Immigrant	Offspring	Exog. Func.	HP	MBPP	Loss Func.	Endo.?
A	-	ET	Any	✓	✓	PP-LL	×
B	ET	ET	Multi-Impulse	✓	✓	PP-LL	✓
C	IC	ET	LHPP	✓	✓	PP-LL	✓
D	-	IC	Any	×	✓	IC-LL/SSE	×
E	ET	IC	Multi-Impulse	×	✓	IC-LL/SSE	✓
F	IC	IC	LHPP	×	✓	IC-LL/SSE	✓

as event times or interval-censored, respectively. We also build endogenous loss functions that account for only endogenous events, allowing us to fit model parameters in all possible regimes of observed data. Table 2 summarises, for each scenario, the corresponding exogenous function, loss function, whether it is applicable for Hawkes, MBPP, or both, and whether an endogenous loss function should be used.

6.1 Exogenous Event Times – Multi-Impulse

Consider a separable scenario in which we observe exogenous events as event times (scenarios B and E in Table 2). That is, we have a set of exogenous event times $\mathfrak{s}_z \in \Psi_T$. We define the corresponding exogenous intensity as the sum of Dirac delta functions, one for each observed exogenous event, i.e., $s_{m\delta}(t)$ the multi-impulse function.

Definition 12 *The* multi-impulse *exogenous function for a set of immigrant event times* $\mathfrak{s}_z \in \Psi_T$ *is defined as*

$$s_{m\delta}(t) = \sum_{\mathfrak{s}_z \in \Psi_T} \delta(t - \mathfrak{s}_z). \tag{39}$$

Notably, the multi-impulse exogenous function is entirely deterministic for a set of immigrant event times and does not have any parameters to fit. Furthermore, $s_{m\delta}(t)$ is not a proper function, as the Dirac delta function is a generalized function not directly computable, i.e., it typically is computed under an integral. Thus, loss functions which require the multi-impulse exogenous to be directly evaluated are incomputable. For these cases, we introduce in Section 6.3 the endogenous loss functions. We use the multi-impulse exogenous function defined in Definition 12 in Section 7 for the scenarios where the immigrants are directly observed, and the intensity function generating them is unknown.

6.2 Exogenous Interval-Censored – Latent Homogeneous Poisson Process

Here we consider the separable scenarios in which the immigrant events are observed as interval-censored data, with no assumption on the endogenous response of the process (scenarios C and F in Table 2). Let $Q = \{q_0, \ldots, q_m\}$ be a pre-specified set of observations times for exogenous events. For each interval $(q_{i-1}, q_i]$ we only observe the total immigrant counts $\mathsf{S}(q_{i-1}, q_i]$, similarly to Section 4. Note that for scenario F (with interval-censored

endogenous response), the immigrant observation times $q_i \in Q$ can be distinct from the endogenous observation time $o_j \in O$, although they are often identical in practice and in our experiments in Sections 7 and 8. The rest of this section builds on the more general case where we do not assume that the immigrant observation times Q are the same as the endogenous events O.

To account for this scenario in which the immigrants are observed as interval-censored data, we assume that the volume of events for each observation interval $(q_{i-1}, q_i]$ is determined by a Poisson process. In particular, we define a *latent homogenous Poisson process (LHPP)* with constant intensity $s(t) = \lambda_i$ for each interval; which subsequently defines a piece-wise constant non-homogenous Poisson process as an exogenous function

$$\sum_{i=1}^{m} \lambda_i \cdot [\![q_{i-1} < t \le q_i]\!].$$
(40)

Notably, the rates λ_i are free parameters to be chosen. We consider the parameters of maximum likelihood.

Proposition 13 *Given a latent homogeneous Poisson process defined over observation intervals $(q_{i-1}, q_i]$ and corresponding immigrant event counts $\mathsf{S}(q_{i-1}, q_i]$, the most likely intensity values are*

$$\lambda_i = \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}}.$$
(41)

Proof Suppose we have a homogeneous Poisson point process $N(t)$ defined over arbitrary half-open interval $(x, y]$, $x \le y$ with a realization of C many events occurring on that interval. The number of events over a given interval in a homogeneous Poisson process follows a Poisson distribution of parameter $\lambda(y - x)$. Thus, the most likely intensity value is given by

$$\lambda^\star = \underset{\lambda}{\operatorname{argmax}} \, \mathbb{P}(N(x, y] = C]; \lambda)$$

$$= \underset{\lambda}{\operatorname{argmax}} \, \frac{(\lambda(y - x))^C}{C!} \exp(-\lambda(y - x))$$

$$= \underset{\lambda}{\operatorname{argmax}} \, C \log(\lambda(y - x)) - \log(C!) - \lambda(y - x)$$

$$= \underset{\lambda}{\operatorname{argmax}} \, C \log(\lambda(y - x)) - \lambda(y - x).$$

The function we are maximizing is concave, thus by setting the derivative to zero we have,

$$\lambda^\star = \frac{C}{y - x}.$$
(42)

Thus for each of the intervals $(q_{i-1}, q_i]$ and corresponding immigrant event counts $\mathsf{S}(q_{i-1}, q_i]$ we have Eq. (41). ∎

Thus using the optimal parameters for the latent homogeneous Poisson process, we have the LHPP exogenous function.

Definition 14 *Given the interval-censored exogenous volumes* $\mathsf{S}(q_{i-1}, q_i]$ *over observation times* $\{q_0, \ldots, q_m\} \in Q$, *then the* latent homogeneous Poisson process (LHPP) *exogenous function is*

$$s_{\blacksquare}(t) = \sum_{i=1}^{m} \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \cdot [\![q_{i-1} < t \leq q_i]\!]. \tag{43}$$

If more information is available about the process in which immigrant events are generated, an inductive bias can be added into the point process by changing the distribution of the exogenous function (for example, by considering a non-homogenous Poisson process in each interval). Furthermore, the LHPP exogenous function given in Eq. (43) can also be derived by first considering the volume distribution of exogenous events for each interval $(q_j, q_{j+1}]$ (see Appendix C). In particular, considering a uniform distribution in Appendix C results in the LHPP exogenous function.

6.3 Endogenous Loss Functions

When the exogenous and endogenous events are separable, both the intensity function and observed data can be separated in to endogenous and exogenous categories. As such, we propose the *endogenous loss function*, which is defined as only the endogenous contribution of the point process to any loss function. The motivation to do so is two fold. Firstly conceptually, as the exogenous events are already observed we only need to account for the endogenous response of the point process. Secondly computationally, the endogenous loss function does not require us to compute the exogenous function when evaluating the loss function. As a side effect, this allows us to circumvent the incomputability problems of the multi-impulse exogenous function, as per Section 6.1.

Suppose that we have a valid point process with corresponding intensity function $\lambda(t) \doteq \lambda(t; \theta)$, where θ denotes the parameterisation of the intensity function. Then for a loss function $\mathscr{L}(\theta; \lambda, \Psi \cup \Upsilon)$, the intensity function $\lambda(\cdot; \theta)$ is compared and evaluated against both the immigrant events Ψ and the offspring events Υ, whether event times or interval-censored. Suppose that the data is separable. Then similar to the separability of the data, we can decompose the intensity function into a sum of exogenous and endogenous function components:

$$\lambda(t; \theta) = \lambda^{\mathrm{exo}}(t; \theta^{\mathrm{exo}}) + \lambda^{\mathrm{endo}}(t; \theta^{\mathrm{endo}}), \tag{44}$$

where $\lambda^{\mathrm{exo}}(t) \doteq \lambda^{\mathrm{exo}}(t; \theta^{\mathrm{exo}})$ solely characterizes the immigrant events and $\lambda^{\mathrm{endo}}(t) \doteq \lambda^{\mathrm{endo}}(t; \theta^{\mathrm{endo}})$ solely characterizes the offspring events. The exogenous parameters θ^{exo} and endogenous parameters θ^{endo} corresponds to decomposition of the parameter θ as a result of Eq. (48).

Hawkes process. An example of a valid decomposition can be found in the definition of the standard Hawkes process intensity function, Eq. (3). Here, the Hawkes process intensity function is explicitly defined as the addition of an exogenous function $s(t)$ and endogenous intensity contribution $\sum_{t_i < t} \phi(t - t_i)$ determined by the past events and kernel function $\phi(\cdot)$.

MBPP. For MBPP, we can define the exogenous-endogenous function decomposition as

$$\xi(t) = \xi^{\mathrm{exo}}(t) + \xi^{\mathrm{endo}}(t) \tag{45}$$

29

$$\xi^{\text{exo}}(t) = s(t) \tag{46}$$

$$\xi^{\text{endo}}(t) = \int_0^t \xi(y)\phi(t-y)\mathrm{d}y. \tag{47}$$

Additionally, if a closed-formed solution of the intensity function $\xi(t)$ (as per Section 3.3) and exogenous function $\xi^{\text{exo}}(t)$ exists, we are able to find a closed-formed solution of the endogenous function

$$\xi^{\text{endo}}(t) = \xi(t) - \xi^{\text{exo}}(t). \tag{48}$$

Notably, the endogenous function $\lambda^{\text{endo}}(t)$ corresponds to the intensity function of a well defined point process. As such, instead of evaluating a loss function on the entire data $\Psi \cup \Upsilon$ and the intensity function $\lambda(t)$, we can instead evaluate the endogenous intensity function $\lambda^{\text{endo}}(t)$ and its corresponding events; which is exactly the endogenous events Υ we can identify from the separability of the data. Thus we define the *endogenous loss function* from the original loss function and point process intensity function:

$$\mathscr{L}^{\text{endo}}(\theta^{\text{endo}}) \doteq \mathscr{L}(\theta^{\text{endo}}; \lambda^{\text{endo}}, \Upsilon). \tag{49}$$

In particular, for the rest of this paper, we will assume that $\theta^{\text{exo}} = \emptyset$, as per the exogenous function we have introduced in this section, i.e., the multi-impulse (Eq. (39)) and LHPP (Eq. (43)) exogenous functions. Thus, we have that the endogenous parameters account for all model parameters of the original point process, $\theta^{\text{endo}} = \theta$.

MBPP + PP-LL. For example, the endogenous loss function of MBPP using the PP-LL loss function, as per Eq. (4), is defined as

$$\mathscr{L}^{\text{endo}}(\theta) = -\sum_{f_i \in \Upsilon_T} \log \xi^{\text{endo}}(f_i) + \int_0^T \xi^{\text{endo}}(s)\,\mathrm{d}s, \tag{50}$$

where Υ_T is the set of offspring events. Notably, this endogenous loss function is appropriate when the offspring data is in point format (i.e., scenarios B and C).

MBPP + IC-LL. In the scenarios when the offspring data is in an interval-censored format (i.e., scenarios E and F), we can no longer use Eq. (50). Instead, we can consider the endogenous version of an IC-LL loss function introduced in Sections 4 and 5. Consider the Bregman divergence interpretation of a loss function, with convex generator $\varphi(z)$. The endogenous version of the loss function is defined as,

$$\mathscr{L}^{\text{endo}}(\theta) = \sum_{i=1}^m B_\varphi(\hat{f}_i, \Xi^{\text{endo}}(o_{i-1}, o_i]), \tag{51}$$

$$\text{where } \Xi^{\text{endo}}(t) = \int_0^t \xi^{\text{endo}}(s)\mathrm{d}s$$
$$\text{and } \hat{f}_i = \mathsf{F}(o_{i-1}, o_i].$$

Here $\mathsf{F}(o_{i-1}, o_i]$ denotes offspring volumes on the interval $(o_{i-1}, o_i]$.

Table 2 outlines when using an endogenous loss function is suitable. In addition, experiments in Section 7 show that even across the different settings where the endogenous loss function is not required, the performances to the non-endogenous standard loss functions are similar.

7. Synthetic Experiments

In this section, we empirically validate our proposals by refitting a synthetically generated dataset with each of the six scenarios described in Table 2. In Section 7.1, we generate datasets of separable Hawkes realizations. We study in detail two parameter combinations – one clearly subcritical (branching factor $n^* \ll 1$) and another approaching criticality ($n^* \simeq 1$) – and two exogenous functions. We also perform a broad parameter sweep and fit over the parameter grid to analyze fitting stability. Next, in Section 7.2, we selectively interval-censor types of events in the datasets to render them compatible with the different scenarios in Table 2. Last, we present our results in Section 7.3.

7.1 Synthetic Datasets

We generate synthetic datasets usable in all scenarios (A-F in Table 2). We start from two exogenous functions and two sets of Hawkes parameters. We generate separable Hawkes realizations for each of the four combinations – i.e., immigrants are distinguishable from the offspring. This allows testing all defined scenarios (for example, to fit in scenario C we hide the exogenous generating function, and we interval-censor the immigrants, more details in Section 7.2).

Exogenous functions. We select two functions for $s(t)$ the exogenous events intensity. The first function is a piece-wise constant function, inspired from the latent homogeneous Poisson process defined in Section 6.2. We select this function to allow control over the number of exogenous events in each observation interval. We denote the synthetic datasets generated with this exogenous function as `HP-pc`, and the resulted sampling event intensity is defined as:

$$\lambda(t) = \sum_{j=1}^{n} \frac{\mathsf{S}(q_{j-1}, q_j]}{q_j - q_{j-1}} [\![q_{j-1} < t \leq q_j]\!] + \sum_{t_i < t} \kappa\theta e^{-\theta(t-t_i)}, \tag{52}$$

where $q_j \in Q_T = \{0, 5, 10, 15\}$, $n = 2$; and $\mathsf{S}(q_0, q_1] = 7$, $\mathsf{S}(q_1, q_2] = 6$, and $\mathsf{S}(q_2, q_3] = 8$.

The second immigrant function is a sinusoidal function, which emulates the seasonality often observed in real-world processes. We denote this dataset as `HP-sin`, and realizations are sampled from the event intensity:

$$\lambda(t) = \sin(t) + 2 + \sum_{t_i < t} \kappa\theta e^{-\theta(t-t_i)} \tag{53}$$

Hawkes kernel parametrization. For each of the above-defined exogenous function, we consider a Hawkes process with exponential memory kernel $\phi(t) = \kappa\theta e^{-\theta t}$. We consider two parameter combinations. The first combination defines a Hawkes process in a clearly subcritical regime (i.e. each event spawn significantly less than one offspring): $\kappa = 0.6$, $\theta = 0.8$, resulting in a branching factor $n^* = 0.6$. The second combination is in a nearly-critical regime, where each event spawns, on average, close to one offspring. This results in

realizations with more significant numbers of events and pronounced clustering effects. We use the parameter set: $\kappa = 0.95$, $\theta = 1.15$, resulting in a branching factor $n^* = 0.95$.

Simulation algorithm. To be usable with all scenarios in Table 2, we generate sequences in which the exogenous and endogenous event times are distinguishable. We use a simulation algorithm that leverages the cluster structure of Hawkes process (Hawkes and Oakes, 1974; Reinhart, 2018), by first sampling the exogenous events and then all their direct and indirect offspring. The steps for generating the synthetic event sequences are:

1. First, we sample immigrants from the fixed exogenous function.

2. Second, we obtain offspring by sampling a Hawkes process with no background intensity and with the immigrant as the first event. We use rejection sampling (Dassios and Zhao, 2013). We denote the collection of an immigrant together with its offspring as a cascade (Rizoiu et al., 2017b).

3. Finally, we combine all cascades to obtain a sequence, which is a realization of a Hawkes process with the benchmark parameters.

We combine the exogenous and endogenous time points for scenarios where we do not require the separation between immigrants and offspring (i.e., A and D). For scenarios that require one or both dimensions interval-censored, we compute the corresponding volumes by counting the number of events over pre-specified observation intervals.

7.2 Experimental Setup

Pre-process synthetic realizations. In our experiments in Section 7.3, the aim is to recover the parameters κ and θ of the *true* model used to generate the synthetic datasets. We test over all scenarios in Table 2, and for each scenario, we use the exact same set of realizations constructed in Section 7.1. Prior to fitting, for each scenario, we perform the following pre-processing on the realizations:

A. We remove the distinction between immigrants and offspring, and we provide the immigrant intensity $s(t)$;

B. We hide the immigrant intensity $s(t)$, and we observe immigrants and offspring separately;

C. We interval-censor the immigrants, and we observe offspring as event times;

D. We remove the distinction between immigrants and offspring, we interval-censor the resulting event stream, and we provide the immigrant intensity $s(t)$;

E. We hide the immigrant intensity $s(t)$, we interval-censor the offsprings, and we observe immigrants (event times) and offspring (interval-censored) separately;

F. We hide the immigrant intensity $s(t)$ and we individually interval-censor the immigrant stream and the offspring stream

Models and loss functions. For the scenarios where the offspring events are observed as event times (scenarios A-C), we fit both Hawkes and MBPP using the PP-LL loss (Eq. (4)). This allows estimating the performances of MBPP to estimate Hawkes parameters. For scenarios in which the offspring are interval-censored (D-F), we only fit MBPP using both IC-LL (Eq. (19)) and SSE (Eq. (36)) loss functions. This allows comparing the performances of the two loss functions. We prefer fitting using the endogenous version of the loss function for the separable settings (B, C, E, F). However, we compare the standard point process and the endogenous version of the loss function for scenario C. Finally, for the non-separable scenarios (A and D), we use the non-endogenous loss functions as there is no distinction between immigrants and offspring.

Closed-form vs. approximate loss functions. We test the fitting performance gap between the closed-form and approximate versions of our models, using the approximation method specified in Section 4.3. Scenario E is used to test our numerically approximated models when a high number of discretization points are given, additionally presenting both endogenous and non-endogenous versions of the IC-LL and SSE loss function. Additionally, discretization tests which do not involve parameter estimation can be found in Appendix E.1 and further non-endogenous loss functions experiments can be found in Appendix E.2.

Additional setup. For Scenarios A-C, we consider $1,000$ event sequences, which are further split into 50 groups to estimate parameters. For Scenarios D-F, we consider $10,000$ sequences, also split into 50 groups to be jointly fit. The additional volume of sequences is to account for the loss of information from interval censoring. For the scenarios where endogenous events are observed as event time (scenarios A-C), we consider both `HP-pc` and `HP-sin` datasets. For interval-censored endogenous events (scenarios D-F), only the more challenging `HP-sin` datasets are considered to limit computation time. In both cases, the sequences we generate are sampled until time-step $T = 30$. When we are testing interval-censored settings, we use uniform observation intervals with 5, 10, 15, 30, 60, and 100 observation periods — for example, with 10 observation intervals we calculate volumes over intervals $(0, 3], (3, 6], \ldots, (27, 30]$.

In each scenario and model tested we fit 50 of that model corresponding to the 50 groups of synthetic sequences. We then calculate the average of the estimated parameters and compare the aggregated statistic against the true parameters.

7.3 Results

Our investigations in this section have five goals. First, we test the performance of MBPP for retrieving the Hawkes parameters used to generate the datasets for scenarios A-C (the Hawkes process cannot be used for scenarios D-F). Second, we seek to quantify the performance loss due to increasingly larger data granularity when moving from scenario A to F. Third, we investigate the effects of using the endogenous loss introduced in Section 6.3 for the separable scenarios B, C, E, and F. Fourth, we study the effects of using the lower bound approximation in Section 4.3 for scenarios D, E, F, and the impact of the number of discretization points. Fifth and last, we study the MBPP fitting stability (the accuracy of retrieving generating parameters) over a wide parameter grid.

7.3.1 Fitting for scenarios A to F

Scenario A. Table 3 shows the results of fitting the Hawkes realizations with both the Hawkes process and MBPP for non-separable data and with knowledge of the parametrization of the exogenous function. Visibly, the parameters fitted with Hawkes are closest to the true parameters, which is expected as the data is generated from a Hawkes model—i.e., this is the upper bound of how well parameters can be retrieved on each dataset. Perhaps surprisingly, MBPP has very similar fitting performances, and we only observe a degradation on the `HP-sin` dataset. This suggests that MBPP can be used as a drop-in alternative for fitting the parameters of Hawkes processes.

Table 3: Synthetic experiments for Scenario A: non-separable data, event times.

True model	True Params.	Model Fit		
HP-pc	$\kappa = 0.6$ $\theta = 0.8$	HP	κ θ	0.58 ± 0.043 0.80 ± 0.087
		MBPP	κ θ	0.58 ± 0.043 0.78 ± 0.145
	$\kappa = 0.95$ $\theta = 1.15$	HP	κ θ	0.95 ± 0.015 1.17 ± 0.073
		MBPP	κ θ	0.95 ± 0.016 1.21 ± 0.246
HP-sin	$\kappa = 0.6$ $\theta = 0.8$	HP	κ θ	0.60 ± 0.019 0.80 ± 0.098
		MBPP	κ θ	0.60 ± 0.018 0.75 ± 0.192
	$\kappa = 0.95$ $\theta = 1.15$	HP	κ θ	0.95 ± 0.008 1.14 ± 0.062
		MBPP	κ θ	0.95 ± 0.010 1.22 ± 0.227

Table 4: Synthetic experiments for Scenario B: separable data, exogenous and endogenous event times.

True model	True Params.	Model		
HP-pc	$\kappa = 0.6$ $\theta = 0.8$	HP-Endo	κ θ	0.59 ± 0.056 0.81 ± 0.135
		MBPP-Endo	κ θ	0.59 ± 0.056 0.8 ± 0.093
	$\kappa = 0.95$ $\theta = 1.15$	HP-Endo	κ θ	0.94 ± 0.025 1.17 ± 0.067
		MBPP-Endo	κ θ	0.94 ± 0.025 1.37 ± 0.228
HP-sin	$\kappa = 0.6$ $\theta = 0.8$	HP-Endo	κ θ	0.59 ± 0.024 0.74 ± 0.11
		MBPP-Endo	κ θ	0.59 ± 0.024 0.75 ± 0.098
	$\kappa = 0.95$ $\theta = 1.15$	HP-Endo	κ θ	0.95 ± 0.021 1.07 ± 0.28
		MBPP-Endo	κ θ	0.95 ± 0.027 1.1 ± 0.124

Scenario B. We distinguish between immigrants and offspring in this scenario, but we do not know the immigrant intensity. We fit the two synthetic datasets using the endogenous versions of MBPP and Hawkes process with the multi-impulse exogenous function. Table 4 shows the obtained fitted parameters indicating very similar performances for both the Hawkes and MBPP. The MBPP even shows a slight advantage for all combinations of parameter and data sets, except for `HP-pc` dataset and parameters $\kappa = 0.95$ and $\theta = 1.15$. Furthermore, MBPP even shows a lower standard deviation indicating that it is more stable when predicting parameters. These results also show the effectiveness of using the multi-impulse exogenous function in the presence of separable data, together with the endogenous loss function.

Scenario C. Table 5 shows the results of fitting in Scenario C, in which immigrants and offspring are separable, immigrants are interval-censored, and offspring are event times. Again, we can conclude that the MBPP and the Hawkes process have very similar performances Similar to Scenario B, the only degradation in approximation quality is for `HP-pc` with the parameters $\kappa = 0.95$ and $\theta = 1.15$. Note that Table 5 also presents results for the non-endogenous version of the loss function, which we discuss in Section 7.3.2.

Table 5: Synthetic experiments for Scenario C: separable data, interval-censored exogenous events and endogenous event times.

True model	True Params.	Model			True model	True Params.	Model		
		HP	κ	0.57 ± 0.064			HP	κ	0.59 ± 0.025
			θ	0.74 ± 0.097				θ	0.52 ± 0.08
	$\kappa = 0.6$	HP-Endo	κ	0.59 ± 0.056		$\kappa = 0.6$	HP-Endo	κ	0.59 ± 0.024
	$\theta = 0.8$		θ	0.8 ± 0.093		$\theta = 0.8$		θ	0.75 ± 0.098
		MBPP	κ	0.59 ± 0.059			MBPP	κ	0.59 ± 0.024
			θ	0.83 ± 0.138				θ	0.72 ± 0.138
HP-pc		MBPP-Endo	κ	0.59 ± 0.056	HP-sin		MBPP-Endo	κ	0.59 ± 0.023
			θ	0.81 ± 0.139				θ	0.73 ± 0.107
		HP	κ	0.94 ± 0.025			HP	κ	0.95 ± 0.021
			θ	1.12 ± 0.065				θ	0.99 ± 0.117
	$\kappa = 0.95$	HP-Endo	κ	0.94 ± 0.025		$\kappa = 0.95$	HP-Endo	κ	0.95 ± 0.021
	$\theta = 1.15$		θ	1.17 ± 0.067		$\theta = 1.15$		θ	1.1 ± 0.124
		MBPP	κ	0.94 ± 0.025			MBPP	κ	0.95 ± 0.03
			θ	1.37 ± 0.237				θ	1.07 ± 0.296
		MBPP-Endo	κ	0.94 ± 0.025			MBPP-Endo	κ	0.95 ± 0.028
			θ	1.37 ± 0.229				θ	1.07 ± 0.282

Scenario D. In this scenario, we do not distinguish between immigrants and offspring, and we observe the process as interval-censored. We use for the first time the novel loss functions introduced in Section 4. I.e., we fit the synthetic data using the IC-LL and SSE loss functions together with the MBPP. Note that we cannot use the standard Hawkes process when the events are observed as interval-censored (see Section 4). Table 6 shows the obtained fitting results for an increasing number of observation intervals. We fit using both the closed-form solution and the approximation introduced in Section 4.3. For the approximation, we use as many approximation points as the number of intervals.

We observe that the closed-form MBPP performs exceptionally well, with no noticeable differences between IC-LL and SSE loss. In particular, we obtain highly accurate parameter fittings for closed-form models even for small numbers of observations. However, when we use the numeric approximation of compensators, we observe a decrease in the accuracy of the recovered parameters. The degradation is prevalent only for θ, which is consistently overestimated. Interestingly, an increase in observation intervals does not yield sizable improvements for the approximate models. Thus, for this scenario, knowing the functional form of the underlying interval-censored Hawkes process (the exogenous function and kernel function) appears more important than having finer-grained observation periods.

Scenario E. Here, we separately observe immigrants (as event times) and offspring (interval-censored). We fit using the multi-impulse exogenous function and the SSE and IC-LL loss functions, with both the closed-form solution and the numerical approximation. Table 7 shows that both loss functions achieve highly accurate parameter fittings for the closed-form models, even for low numbers of observation intervals. However, the approximate models require higher numbers of observation intervals to recover the parameters. For $\kappa = 0.6$ and $\theta = 0.8$, we require at least 30 observation intervals to converge close to the correct θ value. We need at least 60 intervals for the other parameter setting. Notably, the κ parameter is easier to fit and can be approximated using fewer intervals. We further explore the dependence of parameter estimation for the numerically approximated models in Section 7.3.2.

Table 6: Synthetic experiments for Scenario D: non-separable data, interval-censored events. Note that only MBPP can fit this kind of data.

True Params.	Loss Func.	Model		Number of Observation Intervals					
				5	10	15	30	60	100
$\kappa = 0.6$ $\theta = 0.8$	SSE	Closed	κ	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.006	0.6 ± 0.006
			θ	0.81 ± 0.092	0.81 ± 0.087	0.8 ± 0.07	0.8 ± 0.076	0.8 ± 0.072	0.8 ± 0.075
		Approx	κ	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.006	0.6 ± 0.006
			θ	0.85 ± 0.099	0.85 ± 0.096	0.84 ± 0.077	0.84 ± 0.082	0.84 ± 0.079	0.84 ± 0.081
	IC-LL	Closed	κ	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.006	0.6 ± 0.006
			θ	0.82 ± 0.091	0.81 ± 0.086	0.81 ± 0.072	0.8 ± 0.076	0.8 ± 0.073	0.8 ± 0.076
		Approx	κ	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.007	0.6 ± 0.006	0.6 ± 0.006
			θ	0.85 ± 0.098	0.85 ± 0.095	0.84 ± 0.08	0.84 ± 0.083	0.84 ± 0.08	0.84 ± 0.083
$\kappa = 0.95$ $\theta = 1.15$	SSE	Closed	κ	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.005
			θ	1.16 ± 0.082	1.16 ± 0.091	1.16 ± 0.065	1.16 ± 0.1	1.16 ± 0.079	1.16 ± 0.088
		Approx	κ	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.005
			θ	1.24 ± 0.093	1.24 ± 0.103	1.24 ± 0.074	1.24 ± 0.114	1.24 ± 0.09	1.24 ± 0.1
	IC-LL	Closed	κ	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.005
			θ	1.16 ± 0.074	1.16 ± 0.078	1.16 ± 0.06	1.16 ± 0.09	1.16 ± 0.07	1.16 ± 0.082
		Approx	κ	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.005
			θ	1.24 ± 0.083	1.24 ± 0.089	1.24 ± 0.068	1.24 ± 0.103	1.24 ± 0.079	1.24 ± 0.093

Table 7: Synthetic experiments for Scenario E: separable data, event time exogenous events and interval-censored endogenous events.

True Params.	Loss Func.	Model		Number of Observation Intervals					
				5	10	15	30	60	100
$\kappa = 0.6$ $\theta = 0.8$	SSE	Closed	κ	0.60 ± 0.006	0.60 ± 0.006	0.60 ± 0.005	0.60 ± 0.006	0.60 ± 0.005	0.60 ± 0.005
			θ	0.79 ± 0.055	0.79 ± 0.047	0.79 ± 0.042	0.79 ± 0.039	0.79 ± 0.036	0.79 ± 0.036
		Approx	κ	10 ± 0	0.62 ± 0.006	0.61 ± 0.006	0.60 ± 0.006	0.60 ± 0.006	0.60 ± 0.006
			θ	0.01 ± 0	8.82 ± 0.866	9.55 ± 1.589	1.09 ± 0.091	0.91 ± 0.052	0.86 ± 0.043
	IC-LL	Closed	κ	0.60 ± 0.006	0.60 ± 0.005	0.60 ± 0.005	0.60 ± 0.005	0.60 ± 0.005	0.60 ± 0.005
			θ	0.80 ± 0.053	0.80 ± 0.042	0.79 ± 0.037	0.79 ± 0.035	0.79 ± 0.032	0.79 ± 0.032
		Approx	κ	10 ± 0	0.63 ± 0.006	0.61 ± 0.006	0.60 ± 0.005	0.60 ± 0.005	0.60 ± 0.005
			θ	0.01 ± 0	8.22 ± 1.192	9.98 ± 0.103	1.17 ± 0.089	0.94 ± 0.045	0.87 ± 0.038
$\kappa = 0.95$ $\theta = 1.15$	SSE	Closed	κ	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004
			θ	1.15 ± 0.076	1.15 ± 0.075	1.15 ± 0.073	1.15 ± 0.073	1.15 ± 0.072	1.15 ± 0.072
		Approx	κ	10 ± 0	1.14 ± 0.007	1.03 ± 0.005	0.96 ± 0.003	0.95 ± 0.004	0.95 ± 0.004
			θ	0.01 ± 0	9.41 ± 0.764	10 ± 0	10 ± 0	1.77 ± 0.177	1.43 ± 0.111
	IC-LL	Closed	κ	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004
			θ	1.16 ± 0.068	1.16 ± 0.067	1.16 ± 0.065	1.16 ± 0.065	1.16 ± 0.065	1.16 ± 0.065
		Approx	κ	10 ± 0	1.17 ± 0.007	1.04 ± 0.005	0.96 ± 0.003	0.95 ± 0.004	0.95 ± 0.004
			θ	0.02 ± 0	9.95 ± 0.032	10 ± 0	10 ± 0	1.82 ± 0.164	1.46 ± 0.101

Table 8: Synthetic experiments for Scenario F: separable data, interval-censored exogenous and endogenous events.

True Params.	Loss Func.	Model		Number of Observation Intervals					
				5	10	15	30	60	100
$\kappa = 0.6$ $\theta = 0.8$	SSE	Closed	κ	0.726 ± 0.004	0.682 ± 0.005	0.668 ± 0.005	0.642 ± 0.005	0.62 ± 0.005	0.612 ± 0.005
			θ	0.647 ± 0.021	0.641 ± 0.024	0.801 ± 0.031	0.995 ± 0.052	0.915 ± 0.052	0.865 ± 0.045
		Approx	κ	10 ± 0	0.621 ± 0.006	0.605 ± 0.006	0.601 ± 0.006	0.601 ± 0.006	0.601 ± 0.006
			θ	0.005 ± 0	8.823 ± 0.866	9.546 ± 1.589	1.09 ± 0.091	0.909 ± 0.052	0.856 ± 0.043
	IC-LL	Closed	κ	0.729 ± 0.004	0.69 ± 0.005	0.672 ± 0.005	0.642 ± 0.005	0.621 ± 0.005	0.613 ± 0.005
			θ	0.642 ± 0.019	0.675 ± 0.023	0.828 ± 0.029	0.997 ± 0.045	0.929 ± 0.044	0.878 ± 0.039
		Approx	κ	10 ± 0	0.628 ± 0.006	0.611 ± 0.006	0.604 ± 0.005	0.603 ± 0.005	0.602 ± 0.005
			θ	0.006 ± 0	8.223 ± 1.192	9.982 ± 0.103	1.174 ± 0.089	0.938 ± 0.045	0.872 ± 0.038
$\kappa = 0.95$ $\theta = 1.15$	SSE	Closed	κ	0.958 ± 0.002	1.496 ± 2.17	1.494 ± 2.171	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004
			θ	1.765 ± 0.098	1.362 ± 0.358	1.301 ± 0.344	1.283 ± 0.089	1.216 ± 0.08	1.19 ± 0.076
		Approx	κ	10 ± 0	1.139 ± 0.007	1.033 ± 0.005	0.958 ± 0.003	0.95 ± 0.004	0.95 ± 0.004
			θ	0.014 ± 0	9.409 ± 0.764	10 ± 0	10 ± 0	1.765 ± 0.177	1.433 ± 0.111
	IC-LL	Closed	κ	0.959 ± 0.002	0.953 ± 0.003	0.95 ± 0.003	0.949 ± 0.003	0.95 ± 0.004	0.95 ± 0.004
			θ	1.687 ± 0.085	1.501 ± 0.087	1.448 ± 0.092	1.323 ± 0.081	1.239 ± 0.073	1.205 ± 0.069
		Approx	κ	10 ± 0	1.165 ± 0.007	1.043 ± 0.005	0.959 ± 0.003	0.949 ± 0.004	0.95 ± 0.004
			θ	0.019 ± 0	9.953 ± 0.032	10 ± 0	10 ± 0	1.823 ± 0.164	1.456 ± 0.101

Scenario F. This is a separable scenario in which both immigrants and offspring are interval-censored. Table 8 shows the fitting results, and we observe that the quality of parameter estimation improves as we increase the number of observation intervals. Interestingly, the performance improvement is more constant for the IC-LL than SSE, as the number of intervals grows. When using the SSE, the parameters are poorly estimated for 10 and 15 intervals. We also note that the numerical approximation requires between 60 and 100 intervals to approach the true parameters for this scenario. Knowing that these results are directly comparable with those for Scenario E (Table 7), we conclude that fitting in Scenario F is the most challenging as most information gets lost when interval-censoring.

7.3.2 ENDOGENOUS LOSS AND NUMBER OF APPROXIMATION POINTS

Endogenous vs. non-endogenous loss. Table 5 includes results for both the endogenous and non-endogenous loss function for Scenario C. We see that the endogenous versions consistently outperform the non-endogenous version. This is most prominent in the `HP-sin` dataset with parameters $\kappa = 0.6$ and $\theta = 0.8$, where the two Hawkes process models have drastic differences in the fitted θ parameter. Table 9 further shows comparisons between endogenous and non-endogenous numeric approximation models for Scenario E, with many approximation points. Here we see little or no improvement between endogenous and non-endogenous when using the SSE loss function. For the IC-LL loss function, we see that the endogenous version is more stable. Together with Scenario F's conclusions (that IC-LL is more stable than SSE), we conjecture that using the endogenous IC-LL is preferable.

Number of approximation points. All results previously reported in this section use the same number of approximation points for the numerical approximation as the number of observation intervals. In Table 9, we report on the numeric approximation for Scenario E when 300 approximation points are used irrespective of the number of observation intervals. Intuitively, a larger number of approximation points should yield better estimates of the compensator and better parameter fits, with the cost of higher computation time. We see only a slight improvement for the normal IC-LL loss function as the number of

Table 9: Large number of discretization points experiment. Scenario E: separable data, interval-censored exogenous and endogenous events.

Loss Func.	Model		Number of Observation Intervals					
			5	10	15	30	60	100
SSE	Normal	κ	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003
		θ	1.23 ± 0.072	1.23 ± 0.071	1.23 ± 0.071	1.23 ± 0.07	1.23 ± 0.069	1.23 ± 0.069
	Endogenous	κ	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003
		θ	1.23 ± 0.072	1.23 ± 0.071	1.23 ± 0.071	1.23 ± 0.07	1.23 ± 0.069	1.23 ± 0.069
IC-LL	Normal	κ	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003
		θ	0.95 ± 0.054	0.96 ± 0.071	0.97 ± 0.071	0.98 ± 0.06	1 ± 0.055	1.03 ± 0.055
	Endogenous	κ	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003	0.95 ± 0.003
		θ	1.24 ± 0.065	1.24 ± 0.064	1.24 ± 0.064	1.24 ± 0.063	1.24 ± 0.063	1.24 ± 0.063

observation intervals increases. For all other settings, endogenous or not, the estimated parameters are surprisingly stable. This suggests that the performances of the approximation method are dependent on the number of approximation points rather than intervals – as the number of approximation points increases, the error quickly decreases and converges (see Appendix E.1).
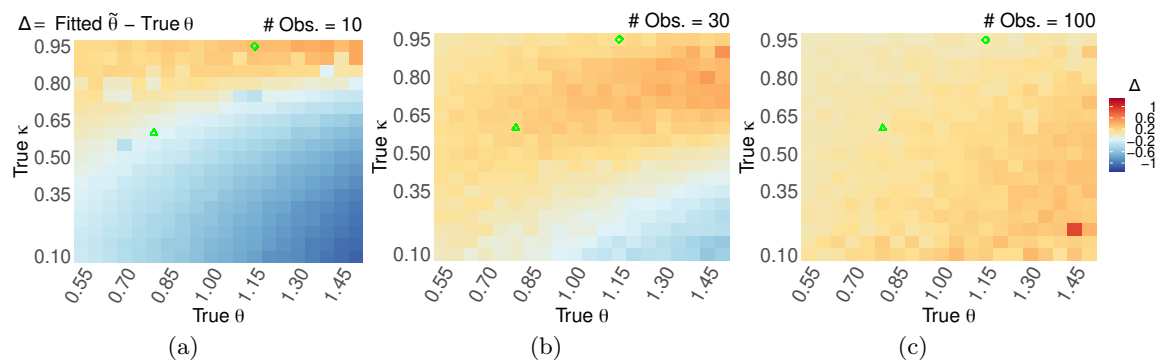
Figure 5: Mean parameter θ fitting error (i.e., fitted $\tilde{\theta}$ − true θ) over a parameter grid $\kappa \in [0.1, 0.95]$ and $\theta \in [0.55, 1.5]$. Realizations are sampled using the Hawkes process with the sinusoidal exogenous function and fitted using MBPP in Scenario F with 10 **(a)**, 30 **(b)** and 100 **(c)** observational intervals. The green shapes highlight the specific parameters described in Section 7.1: $\{\kappa = 0.6, \theta = 0.8\}$ (green triangle) and $\{\kappa = 0.95, \theta = 1.15\}$ (green diamond). For graphics for parameter κ and more observational invervals, see Appendix E.3.

7.3.3 Fitting MBPP with a parameter grid

The synthetic experiments here above only use two parameter combinations defined in Section 7.1. Here, we verify the stability of the MBPP fitting over a wide parameter grid – $\kappa \in [0.1, 0.95]$ (step 0.05) and $\theta \in [0.5, 1.55]$ (step 0.05). We sample realizations from a Hawkes process with a sinusoidal exogenous function (see Eq. (53)), and we fit MBPP in Scenario F (fully interval-censored).

Fig. 5 shows the color map of the fitting error of pameter θ over the grid. In every panel, we highlight the two specific parameter combinations used in the synthetic experiments sections above: $\{\kappa = 0.6, \theta = 0.8\}$ (green triangle) and $\{\kappa = 0.95, \theta = 1.15\}$ (green diamond). Parameter κ is consistently retrieved with only minor errors (see Appendix E.3). We notice that, for high values of θ the fitting under-estimates the true θ for low values of κ (bottom-right corner) and over-estimates it for high values of κ (top-right corner). Note that high θ corresponds to a faster decay rate for the Hawkes kernel $\phi(t)$, which typically translates to offspring events being tightly clustered close to their parent event. As θ decreases (slower kernel decay), the parameter is increasingly well recovered. As expected, the fitting error decreases as the data granularity (i.e., the number of observation intervals) is increased from 10 (Fig. 5a), to 30 (Fig. 5b) and 100 (Fig. 5c). For more data granularities, see Appendix E.3. Finally, for high data granularity (100 intervals in Fig. 5c), we observe a slight over-estimation for θ due to the model mismatch between the generating model (Hawkes process) and the fitting model (MBPP).

8. Real World Experiments

We evaluate our proposed MBPP and its variants on `ACTIVE` (Rizoiu et al., 2017b), a real-world social media dataset containing views to YouTube videos and the tweets that mention them. The task is to predict the future popularity of YouTube videos (their future views) by accounting for the exogenous influence of tweets and the endogenous amplification within YouTube (for example, due to word-of-mouth processes). Both the exogenous (immigrants) and the endogenous (offspring) are in interval-censored format – counts per day. Therefore, we fit using Scenario F (also denoted here below as IC-IC) on a subset of days, and we forecast the future. We compare our MBPP to the HIP process presented in Rizoiu et al. (2017b), the current state of the art in predicting popularity under exogenous influence. The predicted number of views produced by each model is evaluated using average percentile error (APE) and symmetric mean absolute percentage error (sMAPE).

8.1 Dataset and prediction setting

Dataset. We use the `ACTIVE` dataset introduced by Rizoiu et al. (2017b) and further studied in (Wu et al., 2019, 2020). The dataset contains information about more than 14,000 YouTube videos published between 2014-05-29 and 2014-12-26, and the corresponding tweets that share the video in the same time range. For each video, we consider the first 120 days of views and tweets. We fit model parameters on the first 90 days and we use the later 30 days for evaluation. We denote the number of tweets at day i for a video as $[\#\texttt{Tweets on day } i]$ and the number of YouTube views for a video at day i as $[\#\texttt{Views on day } i]$.

Prediction setting. We replicate the evaluation setup introduced in (Rizoiu et al., 2017b). To evaluate our models, we predict the number of times a YouTube video is viewed from day 91 to 120. We consider that the number of tweets (the exogenous influence) is known in the prediction period.

Compared approaches. We evaluate the approximate compensator MBPP, closed-form MBPP, and HIP process, which can be considered an approximation of the MBPP, as per Section 5.3. We also fit and evaluate these models for both the IC-LL and SSE loss func-

tions; and the exponential and power-law trigger kernel when applicable (the closed-form MBPP only exists for the exponential kernel). Furthermore, we use the LHPP exogenous function as the ACTIVE dataset is in the IC-IC setting. It follows that the LHPP exogenous function simplifies to the number of tweets per day as the ACTIVE dataset presents unit-sized observation intervals (in days). Notably, in (Rizoiu et al., 2017b), the HIP process uses only the power-law triggering kernel with the SSE loss function. Thus, the HIP process with exponential kernel and IC-LL loss function combinations are new to our MBPP framework.

Three fitting procedure augmentations. Following Rizoiu et al. (2017b), we introduce three augmentations to improve fitting and forecasting performances. The first two augmentations enhance the exogenous intensity function, while the third is a workaround for the local optima in the loss function.

The first two augmentations are required because the models we discuss in this work make the implicit assumption that the exogenous intensity $s(t)$ completely describes all the immigrants that enter the process. For scenarios B, C, E, F, this implies that the observed exogenous events are the only possible parents for the offspring. However, this assumption is unrealistic for real data. For example, the views of YouTube videos can also be driven by other sources of influences outside tweets, such as Facebook or the radio. We account for unobserved exogenous influence by augmenting the exogenous intensity $s(t)$. First, we scale $s(t)$ with a fitted constant μ. Second, we add two learnable parameters (γ and ν) to account for unobserved external intensity. The augmented exogenous intensity for each day i is:

$$\hat{s}[i] = \gamma \cdot [\![t = 0]\!] + \nu \cdot [\![t > 0]\!] + \mu \cdot [\#\texttt{Tweets on day } i], \tag{54}$$

where μ corresponds to the scaling constant of the observed exogenous influence; and γ and ν correspond to parameters accounting for unobserved external influence.

The third augmentation deals with the non-convexity of the loss surface (see Appendix D for more details), which can cause the gradient descent minimizer to get stuck into local optima. We address this by refitting each video 10 times, starting with randomly selected initial parameters. We chose as the final parameters the combination with the lowest loss.

Forecasting future views using MBPP. We forecast the expected views in the days 91 to 120 using the numeric approximation scheme for the compensators, found in Proposition 7, and leveraging the observed views in days 1 to 90. Specifically, the forecasted number of views on day i is:

$$\begin{aligned}
\texttt{Predicted}[i] = \hat{s}[90 + i] &+ \sum_{j=1}^{90} [\#\texttt{Views on day } j] \int_{i-j-1}^{i-j} \phi(y + 90) \mathrm{d}y \\
&+ \sum_{k=91}^{90+i-1} \Xi[k-1, k] \int_{i-k-1}^{i-k} \phi(y + 90) \mathrm{d}y, \quad \text{for } i \geq 91
\end{aligned} \tag{55}$$

where $\hat{s}[j]$ is defined as per Eq. (54).

We obtain this prediction scheme by replacing the compensator values evaluated between days 1 and 90 with the observed views — as the expectation of the compensator is equivalent to the expected number of events for a non-homogeneous point process. Thus, the fitted compensator only needs to be used when extrapolating past days 90.
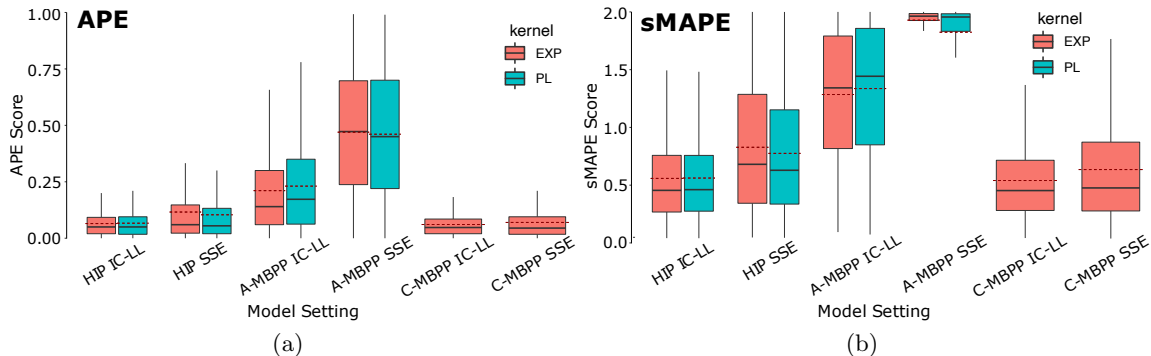
Figure 6: **Evaluation of future popularity prediction using APE (a) and sMAPE (b).** For each model, the boxplots aggregate the APE for all videos in the dataset. A-MBPP denotes the approximate compensator MBPP variant, and C-MBPP denotes the closed-form compensator MBPP variant. For each boxplot, the solid line denotes the median, and the dashed line denotes the mean. The combination HIP SSE PL is the baseline reported in (Rizoiu et al., 2017b)

Evaluation measures. We evaluate the prediction for days 91 to 120 using two evaluation metrics: (1) average percentile error (APE) and (2) symmetric mean absolute percentage error (sMAPE). APE was introduced in Rizoiu et al. (2017b) and is defined as the absolute difference between the recorded YouTube views percentiles and the predicted YouTube views percentile. The measure accounts for the long-tail distribution of online popularity. Intuitively, the impact of errors of the same size should be more significant for unpopular videos than for popular videos. sMAPE is defined as a normalized error of the true YouTube views and the predicted YouTube views over the entire sequence. Noteworthy, sMAPE is a local error measure (only depends on one video). In contrast, APE is a global measure that depends on both the error on the current video and the errors made on all other videos.

8.2 Results

We measure the APE (Fig. 6a) and sMAPE (Fig. 6b) for all HIP and MBPP configurations, summarized as boxplots. We measure the impact of the choices of three components: the kernel (power-law vs. exponential), the loss functions (IC-LL vs. SSE), and the model (closed-form MBPP, approximate MBPP, and HIP). Note that the combination HIP, SSE, and PL kernel is the baseline reported in (Rizoiu et al., 2017b), for which we match the performances claimed by its authors.

We can see that the choice of the kernel does not change the APE value by much, with power-law slightly over-performing exponential (as previously reported in literature (Mishra et al., 2016)). Visibly, the choice of loss function has a more significant impact on performances, with IC-LL constantly outperforming SSE. This is particularly visible for the approximate MBPP (A-MBPP) and also present for both HIP and the closed-form MBPP to a lesser extent. The result is not surprising as the IC-LL loss function comes from the minimization of the corresponding Poisson distribution defined by the MBPP compensator,

whereas the SSE loss function comes from changing the convex generator of a Bregman divergence to obtain an alternative loss function.

The choice of model is also significant, with the closed-form MBPP performing best for each of the loss function choices. This was expected, as HIP and the approximate MBPP are both approximations to the closed-form MBPP (for power-law, MBPP not having a closed-form solution). Perhaps surprisingly, the HIP has lower APE values than approximate MBPP with the same loss function and triggering kernel. This is because the HIP approximation is designed explicitly for unit time intervals — as presented in Section 5.3 — and would perform better than the more general approximate MBPP.

We draw very similar conclusions from studying sMAPE performance measure (Fig. 6b). The only main difference is that the choice of kernel engenders a larger performance difference; for example, the 75% quantile for HIP SSE is much lower for power-law than exponential; similarly, the approximate MBPP (A-MBPP) SSE with power-law has a lower 25% quantile than the exponential kernel. A-MBPP IC-LL performs slightly worse with a power-law triggering kernel than the exponential triggering kernel, similar to the APE boxplot.

Summary. Overall, we summarize the three main findings shown by both the local and global error metrics. First, the IC-LL loss function outperforms SSE. Second, the closed-form MBPP outperforms numerical approximations (A-MBPP and HIP). This follows naturally, given both of those processes are approximations of the closed-form MBPP. Third, the HIP approximation (Rizoiu et al., 2017b) – which is specialized for unit time intervals – outperforms the general approximation presented in the compensator approximation MBPP.

9. Related Work

We structure the related work into three main themes. First, we focus on prior work dedicated to estimating the parameters of Hawkes processes in both event-time and interval-censored settings. Second, we consider alternative theoretic modeling paradigms and augmented Hawkes processes, and we investigate their link to the Hawkes processes. Finally, we consider the applications of Hawkes processes and other models in various domains.

9.1 Hawkes model parameter estimation

Several techniques have been proposed for estimating the parameters of a Hawkes model from data. Here, we concentrate on techniques using event-time data (other than the typical maximum likelihood estimates detailed in Section 2.3) and interval-censored data, both parametric and non-parametric.

Event-time models. We start with works that estimate Hawkes model parameters starting from event-time data in several settings. For examples, Piggott and Solo (2018) propose an online estimator for the parameters of a Hawkes process by using a Laguerre basis expansion; their work targets an online learning setting. The RedQueen algorithm (Zarezade et al., 2017) provides a framework for expressing a Hawkes process's counting process and intensity function as a jump stochastic differential equation, providing an online algorithm with provable guarantees (Zarezade et al., 2018). Several studies have linked the Hawkes process to the infinite-server queues when the Hawkes process characterizes the arrivals into

the queue. Koops et al. (2018) obtain a system of differential equations that characterize the joint distribution of arrival intensity and the number of customers. Similarly, Daw and Pender (2018b) find differential equations to characterize the moments in several different settings. Finally, Gao and Zhu (2018) show that the queue length process is limited to a non-Markovian Gaussian process in the asymptotic regime of a stationary Hawkes process. More related to the current work, Xu et al. (2017) address the problem of fitting a Hawkes process observed during a finite interval of time – denoted as a double-censored event sequence. Note that their definition of *censored* is different than our interval-censoring in Section 4.1, i.e., they observe all event times during the given interval and nothing before and after. To avoid edge effects when the history is missing, Xu et al. (2017) simulate many possible prefixes for the given observed window.

Most of the approaches mentioned above define the Hawkes process using a parametric family of kernels and estimate their parameters. In contrast, other works estimate the kernel in a non-parametric way. For example, Dion and Lemler (2020) use trigonometric basis vectors to provide a non-parametric estimator of one-dimensional diffusion processes characterized by Hawkes processes. Similarly, Li et al. (2017) use a Fourier transform to provide a non-parametric estimator for the multidimensional Hawkes process. Zhang et al. (2019, 2020a,b) propose Bayesian estimation procedures for the parameters of the Hawkes process kernel, which leverage the Hawkes process's cluster structure.

The above approaches are typically limited to event time data and cannot estimate model parameters in the presence of interval-censored or mixed data. Our approach can estimate parameters seamlessly from both event time and interval-censored data and operate on mixtures of the two.

Interval-censored models. A series of studies show that the Hawkes process can be estimated in the interval-censored setup using an integer-valued auto-regressive (INAR) model. Kirchner (2016), and in their follow-up work in (Kirchner, 2017), show that the distribution of the resulting 'bin-count sequences' (i.e., number of events in each interval) can be approximated by the (multivariate) INAR(p) model. They establish a least-squares estimation of the INAR(p) model, which yields estimates for the underlying multivariate Hawkes process after appropriate scaling. The quality of the estimates is compared against maximum likelihood estimates (on event-times) in (Kirchner and Bercher, 2018). At first glance, these studies are similar to our work, as the link to the INAR model provides a discrete-time estimator of the Hawkes process. However, they do not expose the connection between the Hawkes process's event-time and interval-censored formulations as shown in this paper in Sections 4.1 and 4.2. Therefore, they do not establish a correspondence between the discrete and continuous regimes.

Additionally, the INAR model and its estimation have been developed under the assumption of a constant exogenous intensity $s(t) = s$. This is incompatible with the synthetic and real-world experiments in Section 7 and Section 8, respectively, since these experiments involve a non-constant $s(t)$ driving the system.

For other interval-censored settings, Lu (2019) presents several Taylor's expansion approximation algorithms for several thinning-based count processes, including the previously mentioned INAR model. Similarly, Manolakis et al. (2019) provide a survey of signal processing approaches for interval-censored processes, detailing many different auto-regressive models. Like the INAR model, there is no theoretical correspondence between the various

discrete models and the underlying true continuous-time process. In the context of survival analysis, Hudgens et al. (2014) propose a scheme for estimating the parameters of the cumulative incidence function in interval-censored and partly interval-censored settings.

Closest to our work, the Hawkes intensity process (HIP) (Rizoiu et al., 2017b) defines an estimator as the Hawkes process intensity function's expectation. It fits the new process in an interval-censored setting by using a sum of squared errors loss function. This study has a series of shortcomings already outlined in Section 2.4. In comparison, this work introduces a new point process (the non-homogeneous Mean Behavior Poisson process (MBPP)) whose intensity is the expected intensity of a Hawkes process of equivalent parameters. The MBPP allows fitting both event-time and interval-censored data. We further show the specific assumption needed to recover the HIP process (see Section 5.3).

Among the non-parametric approaches, Shlomovich et al. (2020) propose the estimation of aggregated Hawkes process using a Monte Carlo Expectation-Maximization algorithm. In aggregated data, event times are rounded to a given precision for data capturing or storage reasons (e.g., only the hour and minute of arrival is recorded, the seconde and below are omitted). The aggregated process at each time point is the number of events with that rounded time-stamp – in effect a version of the interval-censored setting defined in Section 4 and Table 2. A spectral method of approximating stationary Hawkes process, based on Whittle's approach, is established by considering a mixing condition with polynomial decay rates from their Poisson clustering structure (Cheysson and Lang, 2020). There are prior works that propose solutions using panel count data analysis – an inherently non-parametric approach. Ding et al. (2018) propose a Gaussian process-modulated Poisson process that permits panel data observations (i.e., interval-censored). Later, Moreno et al. (2020) wrap several popular panel count inference methods to deal with incomplete data and the Poisson process assumption's mis-specification. Note that these approaches estimate Poisson processes, not Hawkes processes, and are not directly linked to the current work apart from the presentation of the data. Because of the non-parametric estimation of the Poisson intensity, this type of approaches cannot be applied to MBPP either out of the box, as MBPP requires a parametric form to connect back to the Hawkes process as shown in Section 3.1.

Our study differs from the above-mentioned by connecting the Hawkes process with a non-homogeneous Poisson process and providing a generalized treatment of the HIP models. Notably, our proposed mean behavior Poisson process relates to the Hawkes process by a direct parameter equivalence, which provides a bijection between the models which the previous literature did not have. Furthermore, we differentiate from prior studies by providing a singular model to fit both interval-censored and event-time data; only the log-likelihood function changes to accommodate the type of data. Finally, our proposed exogenous functions allow a mixed dataset of interval-censored and event-time data.

9.2 Other modeling paradigms and their links to the Hawkes model

The Hawkes processes are not the only option for modeling events in continuous time and various other models have also been used. Here we present some of these alternatives and how they related to Hawkes.

Linking Hawkes to other models. In their earlier work, Lewis and Shedler (1979) propose a foundational connection between homogeneous and non-homogeneous Poisson processes. They designed a thinning procedure – a simple and efficient algorithm for simulating non-homogeneous Poisson processes from homogeneous Poisson processes through thinning. This method was also adapted for Hawkes processes and is the de-facto method of simulating Hawkes process event sequences (Ogata, 1981). More recently, the Hawkes process has been linked to epidemic models (Rizoiu et al., 2018b). More specifically, the expectation of the intensity of the new infection in a stochastic Susceptible-Infectious-Recovery (SIR) epidemic model over all possible recovery processes is equal to the event intensity in a finite population Hawkes process (HawkesN). HawkesN is a Hawkes process in which the total number of events is limited and for which the remaining number of events modulates the intensity. Kong et al. (2020a) presented a generalized connection between Hawkes processes and epidemic models for arbitrary recovery time distributions and established relationships between the (latent and arbitrary) recovery time distribution, recovery hazard function, and the infection kernel of self-exciting processes.

Hawkes properties and augmented models. Other studies focus on analyzing other aspects of the Hawkes process that are not directly used to connect different models and estimation procedures. Although these studies are only tangentially related to our work, we include some of these theoretical studies for completeness. Cordi et al. (2018) show that univariate symmetric Hawkes processes are only weakly causal. They find that reversed event sequences have similar point process log-likelihood scores and sometimes even yield better fits. Zino et al. (2018) investigate the epidemic threshold of Susceptible-Infected-Susceptible models in the presence and absence of self-excitation in time-varying networks. The paper additionally proposes a model where self-excitement govern the temporal evolution of an activity-driven network. Embrechts and Kirchner (2018) propose several graph-theoretic objects which summarize the branching structure of a multivariate Hawkes process. Daw and Pender (2018a) propose an ephemerally self-excited process which adds an additional stochastic component to the time in which excitement affects the process. Intuitively, this is motivated by how past events in epidemics only influence the future as long as the event stays active (say, as long as a person is infectious). Finally, Dion et al. (2019) introduce a new class of diffusion processes where jumps are driven by multivariate nonlinear Hawkes processes.

Other point process modeling paradigms. Other models have been used in event time modeling as an alternative to Hawkes processes. Before being used to estimate Hawkes processes (as described above), the integer auto-regressive (INAR) model was proposed as a discrete alternative to the continuous-valued auto-regressive process (Al-Osh and Alzaid, 1987). Renewal processes are also a common alternative to the Hawkes process. For example, Mishra et al. (2020) propose a generalized renewal process that arises from the expectation of an age-dependent branching process. The renewal process generalization accounts for exogenous events and the time varying-reproduction number for modeling epidemics; they tested their approach on COVID-19 case data in South Korea. Other point process modeling paradigms make use of Markov processes. For example, Soumaré and Tafolong (2017) propose a two-state Markov switching model to model risk-based pricing models. Stochastic differential equations are another method of characterizing the dynamics of event time processes, with links to the Hawkes process as per the prior mentioned

RedQueen algorithm Zarezade et al. (2018). A stochastic differential equation framework has been proposed for modeling information diffusion over networks (Wang et al., 2016). Similarly, Zarezade et al. (2017) uses an optimal control problem for jump stochastic differential equations to predict when to post in a social network for visibility.

Our study presents a novel variant of the non-homogeneous Poisson process, the mean behavior Poisson. Although we mainly use MBPP to approximate the Hawkes process in interval-censored settings, it can be used on its own as an alternative to the Hawkes process with nicer properties (see the independent increments discussion in Section 2). Furthermore, the exogenous functions that we introduce can be used for other point process intensity functions.

9.3 Applications of Hawkes processes

Earthquakes, finance, and epidemics. Hawkes processes have long been used for modeling events in various domains, from modeling the aftershocks of earthquakes to predicting social network interactions. One of the first domains to adapt Hawkes processes was earthquake modeling. The epidemic-type aftershock (ETAS) (Ogata, 1988) model assumes that each earthquake generates aftershocks at a rate governed by a Hawkes model with a power-law decay rate. Later works show the connection to the Richter intensity scale and provide an analytical solution for Hawkes process variants (Ogata, 1999; Helmstetter and Sornette, 2002a,b). The Hawkes processes have also been used in financial markets. One such work uses a multivariate Hawkes process to account for dynamics of market prices by tailoring four Hawkes kernels for their domain (Bacry and Muzy, 2014). Similarly, Hawkes processes have also been used to study epidemic spread. Zino et al. (2020) use Hawkes processes to study the bursty activity patterns in social, temporal networks to help determine where targeted behavioral changes could accelerate the erratic of epidemics.

Information spread. With the rise of social media platforms, many studies use Hawkes processes to characterize how information develops and moves around social networks. These studies concentrate on either the users of the network or the individual pieces of content (like a post). Similarly, Kobayashi and Lambiotte (2016) predict the time evolution of retweet activity by taking into account the circadian nature of users and the aging of information. Hawkes has been effectively applied to social media platforms other than Twitter. The Hawkes process was found to outperform reinforced Poisson processes for individual micro-blogs on Sina Weibo, a Chinese micro-blogging network. Zarezade et al. (2016) propose a spatial-temporal Hawkes process with a periodic decaying kernel, which considers location-based data of users checking in to leverage the observation that users are influenced by locations in which their close friends have recently visited. Wang and Resnick (2019) incorporate the periodic nature of users on social media by switching from a Hawkes process in the day into a non-homogeneous Poisson process at night.

Cascade size. Cha et al. (2009) analyze the Flickr social network, finding that popular photos do not necessarily spread widely and popular photos spread slowly. The study additionally found that information exchanged between friends likely accounted for over 50% of content sharing, with significant delays between any additional hops. Similarly, Goel et al. (2012) find that in a variety of social network domains, most of the cascade sizes in network diffusions are small, terminating within one degree of an initial adapting "seed".

For large cascades, a study found that in addition to temporal and structural features being key predictors, breadth rather than depth is a better indicator of a large cascade. Yu et al. (2017) describe a stochastic model based on survival analysis to unify the task of predicting cascade sizes at a given time and the time when a cascade reaches a specific size. The authors propose an estimation of the cascading process, and the behavioral dynamics are aggregated. Zhao et al. (2015) use a doubly stochastic extension of the Hawkes process to examine the Twitter social network and make predictions on a tweet's final number of reshares by looking at its current reshare history. Mishra et al. (2016) predict the final popularity of Twitter cascades by combining a marked Hawkes process with a power-law kernel and feature-driven methods. Finally, Mishra et al. (2018) emulate a Hawkes process unfolding using Recurrent Neural Networks that leverage the Hawkes process intensity function.

User influence. In a Flickr dataset, Goyal et al. (2010) investigate the influence probabilities between users in a structured social network graph. The study presents three classes of models for the influence probabilities: the first which assumes that the influence does not change with time; the second assumes influence is a function of continuous-time; the third proposes a discrete approximation of the continuous-time models. Matsubara et al. (2012) provide an economical alternative to the Hawkes process to model the rise and fall pattern of influence propagation used to investigate the fast propagation of news rumors in social networks. Kempe et al. (2003) propose a greedy strategy to find the most influential node in a network to determine which nodes can be targeted to trigger a large cascade of idea adaptation. Rizoiu et al. (2018a) propose a method – implemented as the software package `birdspotter` (Ram et al., 2021) – to estimate user influence in retweet data sets when the direct retweeting relation (who retweets whom directly) is latent. They use a Hawkes-inspired retweeting probability and compute user influence as the expected number of retweets over all possible unfolding of retweet cascades. Ram and Rizoiu (2021) further extend this Hawkes process-based method to include social science grounded phenomena, such as network conductance and influence-capital distribution.

Content popularity. Making predictions about the popularity of online content has been another focus of study in social networks. Szabo and Huberman (2010) present a method for accurately predicting the long-time popularity from early measurements of user access, with empirical results on YouTube and Digg datasets. Another study found that group-level popularity is a necessary and more practical measurement to predict the popularity of online content (Hoang et al., 2017). To predict online video popularity, Ding et al. (2015) propose a dual sentimental Hawkes process that incorporates sentimental features in addition to event counts. Rizoiu et al. (2017b) propose the HIP model already discussed in Section 2.4 and apply it to characterize and predict the popularity of online YouTube videos. They also use the same model to build optimal promotion schedules to boost online popularity (Rizoiu and Xie, 2017). Kong et al. (2018) build a dashboard web application that leverages HIP to analyze online videos.

Fake users and news. There have also been studies on detecting differences in online content, such as differentiating between fake news and regular users. (Tschiatschek et al., 2018) is one example of such a study, where a Bayesian inference algorithm is used for detecting fake news and jointly learning about the accuracy of users who flag such content over time. Kong et al. (2020b) propose a dual mixture model of Hawkes processes to jointly fit a group of cascades (large and small) relating to content from the same source. They

use the resulted parameters to build diffusion embeddings for news sources – i.e., a vector that describes a news source based on how the retweet cascades concerning produced articles propagate through Twitter. The authors show that the diffusion embeddings induce a space in which controversial and reputable news sources are separable.

Software tools that implement Hawkes processes. There exist several tools that implement the Hawkes process. These are usually programming language-specific (see `THAP` (Xu and Zha, 2017) in Matlab, `PoPPy` (Xu, 2018) in PyTorch) and some allow extensions to the Hawkes process (`pyhawkes` (Linderman and Adams, 2014) implements network Hawkes models). The most popular package, `tick` (Bacry et al., 2017), has the most active community and supplies a comprehensive set of models and helper functions for general time-dependent modeling, including Hawkes processes. Recently, `evently` (Kong et al., 2021) was developed as a Hawkes process toolbox and designed with an emphasis on online information diffusion modeling.

In this work, we address the application of popularity prediction — therefore, we classify the present work under the heading content popularity here above. In our experiments in Section 8.2, we show that we outperform HIP (the current state of the art in content popularity prediction).

10. Conclusion and future work

This work studies the self-exciting Hawkes point process in an interval-censored setting – i.e. when we do not observe individual event times but instead counts of events during pre-defined observation intervals. As Hawkes processes lack the independent increment property, computing the volumes of events is intractable as it depends on all the past observed volumes. This research overcomes the above difficulties by making several contributions. First, we establish the Mean Behavior Poisson process (MBPP), a novel Poisson process with a direct parameter correspondence to the popular self-exciting Hawkes process. The event intensity function of the MBPP is the expected intensity over all possible Hawkes realizations with the same parameter set. We fit MBPP in the interval-censored setting using an interval-censored Poisson log-likelihood (IC-LL). Second, we introduce two novel exogenous functions to distinguish the exogenous from the endogenous events. We propose the multi-impulse exogenous function when the exogenous events are observed as event time and the latent homogeneous Poisson process exogenous function when the exogenous events are presented as interval-censored volumes. Third, we provide several approximation methods to estimate the intensity and compensator function of MBPP when no analytical solution exists. Fourth and finally, we connect the interval-censored loss of MBPP to a broader class of Bregman divergence-based functions. Using the connection, we show that the current state of the art in popularity estimation (Hawkes Intensity Process (HIP) (Rizoiu et al., 2017b)) is a particular case of the MBPP. We verify our models through empirical testing on synthetic data and real-world data. We find that on real-world datasets that our MBPP outperforms HIP for the task of popularity prediction.

Future work. This work only addresses the uni-variate Hawkes process. However, drawing from the experiments in Section 8, one could argue that tweets are not only ex-ogenous sources, and they are influenced by the views also. In other words, tweets and views should be jointly modeled using a bi-variate Hawkes, where one dimension is interval-

censored. Future work will address this gap by investigating partially censored multivariate MBPP and marked variants of the Hawkes process. A second direction for future work is finding theoretical bounds for our approximation method for the MBPP compensator (Eq. (26)).

Acknowledgments and Disclosure of Funding

References

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707, 2011. doi: 10.1109/INFCOMW.2011.5928903.

M. A. Al-Osh and A. A. Alzaid. First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8(3):261–275, 1987. ISSN 14679892. doi: 10.1111/j.1467-9892.1987.tb00438.x.

Ifigeneia Apostolopoulou, Scott Linderman, Kyle Miller, and Artur Dubrawski. Mutually regressive point processes. In *NeurIPS*, 2019.

G Arfken. *Mathematical Methods for Physicists 3rd edn (San Diego, CA: Academic)*. 1985.

Emmanuel Bacry and Jean François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014. ISSN 14697696. doi: 10.1080/14697688.2014.897000. URL https://www.tandfonline.com/doi/abs/10.1080/14697688.2014.897000.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 01(01), 2015.

Emmanuel Bacry, Martin Bompaire, Philip Deegan, Stéphane Gaïffas, and Søren V Poulsen. Tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *JMLR*, 2017.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Viktor Benes and Jan Rataj. *Stochastic geometry: selected topics*. Springer Science & Business Media, 2007.

Daniel Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pages 1–45, 1760.

E Borel. Sur l'emploi du théoreme de bernoulli pour faciliter le calcul d'une infinité de coefficients. application au probleme de l'attentea un guichet. *CR Acad. Sci. Paris*, 214: 452–456, 1942.

Clive Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876 – 912, 2007.

D.R. Brillinger, P.M. Guttorp, and F.P. Schoenberg. *Encyclopedia of environmetrics*, volume 3. Wiley, 2002.

Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. DeepHawkes: Bridging the gap between prediction and understanding of information cascades. In *International Conference on Information and Knowledge Management, Proceedings*, volume Part F131841, pages 1149–1158, New York, NY, USA, nov 2017. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3132973. URL `https://dl.acm.org/doi/10.1145/3132847.3132973`.

Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the Flickr social network. In *WWW'09 - Proceedings of the 18th International World Wide Web Conference*, pages 721–730, 2009. ISBN 9781605584874. doi: 10.1145/1526709.1526806.

Felix Cheysson and Gabriel Lang. Strong mixing condition for Hawkes processes and application to Whittle estimation from count data. mar 2020. URL `http://arxiv.org/abs/2003.04314`.

Michael Collins, Robert E Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Marcus Cordi, Damien Challet, and Ioane Muni Toke. Testing the causality of Hawkes processes with time reversal. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(3):033408, mar 2018. ISSN 17425468. doi: 10.1088/1742-5468/aaac3f.

D. R. Cox and V. Isham. *Point Processes*. Chapman & Hall, 1980.

D. R. Cox and P. A. W. Lewis. Multivariate point processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Probability Theory*, pages 401–448, Berkeley, Calif., 1972. University of California Press.

D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003.

Angelos Dassios and Hongbiao Zhao. Exact simulation of hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18:13 pp., 2013.

Andrew Daw and Jamol Pender. An Ephemerally Self-Exciting Point Process. nov 2018a. URL http://arxiv.org/abs/1811.04282.

Andrew Daw and Jamol Pender. Queues Driven by Hawkes Processes. *Stochastic Systems*, 8(3):192–229, sep 2018b. ISSN 1946-5238. doi: 10.1287/stsy.2018.0014.

Hongyi Ding, Young Lee, Issei Sato, and Masashi Sugiyama. Variational inference for Gaussian processes with panel count data. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, volume 1, pages 290–299. Association For Uncertainty in Artificial Intelligence (AUAI), mar 2018. ISBN 9781510871601. URL http://arxiv.org/abs/1803.04232.

Wanying Ding, Yue Shang, Lifan Guo, Xiaohua Hu, Rui Yan, and Tingting He. Video popularity prediction by sentiment propagation via implicit network. In *International Conference on Information and Knowledge Management, Proceedings*, volume 19-23-Oct-2015, pages 1621–1630. Association for Computing Machinery, oct 2015. ISBN 9781450337946. doi: 10.1145/2806416.2806505.

Charlotte Dion and Sarah Lemler. Nonparametric drift estimation for diffusions with jumps driven by a Hawkes process. *Statistical Inference for Stochastic Processes*, 23(3):489–515, oct 2020. ISSN 15729311. doi: 10.1007/s11203-020-09213-5. URL https://link.springer.com/article/10.1007/s11203-020-09213-5.

Charlotte Dion, Sarah Lemler, and Eva Löcherbach. Exponential ergodicity for diffusions with jumps driven by a Hawkes process. apr 2019. URL http://arxiv.org/abs/1904.06051.

P. Embrechts and M. Kirchner. Hawkes graphs. *Theory of Probability and its Applications*, 62(1):132–155, may 2018. ISSN 10957219. doi: 10.1137/S0040585X97T988538.

Xuefeng Gao and Lingjiong Zhu. Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems*, 90(1-2):161–206, oct 2018. ISSN 15729443. doi: 10.1007/s11134-018-9570-5.

Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. The structure of online diffusion networks. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 623–638, 2012. ISBN 9781450314152. doi: 10.1145/2229012.2229058.

Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 241–250, 2010. ISBN 9781605588896. doi: 10.1145/1718487.1718518.

Jan Grandell. Point processes and random measures. *Advances in Applied Probability*, pages 502–526, 1977.

Stephen J. Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. Critical reflexivity in financial markets: a Hawkes process analysis. *European Physics Journal B*, 86(10):442, 2013.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83, 1971.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Agnès Helmstetter and Didier Sornette. Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research: Solid Earth*, 107 (B10):ESE 10–1–ESE 10–21, oct 2002a. ISSN 2169-9356. doi: 10.1029/2001jb001580.

Agnès Helmstetter and Didier Sornette. Diffusion of epicenters of earthquake aftershocks, Omori's law, and generalized continuous-time random walk models. *Phys. Rev. E*, 66(6): 061104.1–061104.24, 2002b. URL `http://link.aps.org/doi/10.1103/PhysRevE.66.061104`.

Agnes Helmstetter and Didier Sornette. Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research: Solid Earth*, 107 (B10), 2002c.

Minh X. Hoang, Xuan Hong Dang, Xiang Wu, Zhenyu Yan, and Ambuj K. Singh. GPOP: Scalable group-level popularity prediction for online content in social networks. In *26th International World Wide Web Conference, WWW 2017*, pages 725–733. International World Wide Web Conferences Steering Committee, 2017. ISBN 9781450349130. doi: 10.1145/3038912.3052626.

Michael G Hudgens, Chenxi Li, and Jason P Fine. Parametric likelihood inference for interval censored competing risks data. *Biometrics*, 70(1):1–9, 2014.

Martin Jacobsen. Compensators and Martingales. In *Point Process Theory and Applications*, pages 33–102. Birkhäuser-Verlag, Boston, 2006. doi: 10.1007/0-8176-4463-6_4. URL `http://link.springer.com/10.1007/0-8176-4463-6_4`.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003. doi: 10.1145/956750.956769.

M. Kirchner and A. Bercher. A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation*, 88(6):1106–1116, apr 2018. ISSN 15635163. doi: 10.1080/00949655.2017.1422126. URL `https://www.tandfonline.com/doi/abs/10.1080/00949655.2017.1422126`.

Matthias Kirchner. Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, aug 2016. ISSN 03044149. doi: 10.1016/j.spa.2016.02.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S0304414916000399`.

Matthias Kirchner. An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4):571–595, apr 2017. ISSN 1469-7688. doi: 10.1080/14697688.2016.1211312. URL `https://www.tandfonline.com/doi/full/10.1080/14697688.2016.1211312`.

Fima C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, 3 edition, 2012.

Donald E. Knuth. Two notes on notation. *The American Mathematical Monthly*, 99(5): 403–422, 1992. doi: 10.1080/00029890.1992.11995869. URL https://doi.org/10.1080/00029890.1992.11995869.

Ryota Kobayashi and Renaud Lambiotte. TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 191–200. AAAI Press, 2016. ISBN 9781577357582.

Quyu Kong, Marian Andrei Rizoiu, Siqi Wu, and Lexing Xie. Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pages 175–178, Lyon, France, 2018. ACM Press. ISBN 9781450356404. doi: 10.1145/3184558.3186972. URL https://arxiv.org/abs/1801.04117http://dl.acm.org/citation.cfm?doid=3184558.3186972.

Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. Modeling Information Cascades with Self-exciting Processes via Generalized Epidemic Models. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 286–294, New York, NY, USA, jan 2020a. ACM. ISBN 9781450368223. doi: 10.1145/3336191.3371821.

Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 645–654, New York, NY, USA, oct 2020b. ACM. ISBN 9781450368599. doi: 10.1145/3340531.3411861.

Quyu Kong, Rohit Ram, and Marian-Andrei Rizoiu. Evently. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1097–1100, New York, NY, USA, mar 2021. ACM. ISBN 9781450382977. doi: 10.1145/3437963.3441708. URL http://arxiv.org/abs/2006.06167https://dl.acm.org/doi/10.1145/3437963.3441708.

D. T. Koops, M. Saxena, O. J. Boxma, and M. Mandjes. Infinite-server queues with Hawkes input. *Journal of Applied Probability*, 55(3):920–943, sep 2018. ISSN 00219002. doi: 10.1017/jpr.2018.58. URL /core/journals/journal-of-applied-probability/article/infiniteserver-queues-with-hawkes-input/29A985198A710AB1F7636DCEAEA9C0E6.

P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes Processes. *ArXiv e-prints*, July 2015.

P. A.W. Lewis and G. S. Shedler. SIMULATION OF NONHOMOGENEOUS POISSON PROCESSES BY THINNING. *Naval research logistics quarterly*, 26(3):403–413, 1979. ISSN 00281441. doi: 10.1002/nav.3800260304.

Sha Li, Xiaofeng Gao, Weiming Bao, and Guihai Chen. FM-Hawkes: A Hawkes process based approach for modeling online activity correlations. In *International Conference*

on Information and Knowledge Management, Proceedings, volume Part F131841, pages 1119–1128, New York, NY, USA, nov 2017. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3132883. URL `https://dl.acm.org/doi/10.1145/3132847.3132883`.

Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *ICML*, 2014.

Yang Lu. The predictive distributions of thinning-based count processes. *Scandinavian Journal of Statistics*, page sjos.12438, dec 2019. ISSN 0303-6898. doi: 10.1111/sjos.12438. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12438`.

Dimitris Manolakis, Nicholas Bosowski, and Vinay K. Ingle. Count Time-Series Analysis: A Signal Processing Perspective. *IEEE Signal Processing Magazine*, 36(3):64–81, may 2019. ISSN 15580792. doi: 10.1109/MSP.2018.2885853.

Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 6–14, 2012. ISBN 9781450314626. doi: 10.1145/2339530.2339537.

Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Feature Driven and Point Process Approaches for Popularity Prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, pages 1069–1078, Indianapolis, IN, USA, 2016. ACM Press. ISBN 9781450340731. doi: 10.1145/2983323.2983812. URL `http://dl.acm.org/citation.cfm?doid=2983323.2983812http://arxiv.org/pdf/1608.04862.pdf`.

Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, pages 1–10, Stanford, CA, USA, 2018. URL `https://arxiv.org/pdf/1804.02101.pdf`.

Swapnil Mishra, Tresnia Berah, Thomas A. Mellan, H. Juliette T. Unwin, Michaela A Vollmer, Kris V Parag, Axel Gandy, Seth Flaxman, and Samir Bhatt. On the derivation of the renewal equation from an age-dependent branching process: an epidemic modelling perspective. jun 2020. URL `http://arxiv.org/abs/2006.16487`.

Alexander Moreno, Zhenke Wu, Jamie Yap, David Wetter, Cho Lam, Inbal Nahum-Shani, Walter Dempsey, and James M Rehg. A Robust Functional EM Algorithm for Incomplete Panel Count Data. mar 2020.

Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. Bregman voronoi diagrams: Properties, algorithms and applications. *arXiv preprint arXiv:0709.2196*, 2007.

Joseph D. O'Brien, Alberto Aleta, Yamir Moreno, and James P. Gleeson. Quantifying Uncertainty in a Predictive Model for Popularity Dynamics. jan 2020. URL `https://arxiv.org/abs/2001.09490`.

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

Yosihiko Ogata. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association*, 83(401):9–27, mar 1988. ISSN 0162-1459. doi: 10.1080/01621459.1988.10478560.

Yosihiko Ogata. Seismicity Analysis through Point-process Modeling: A Review. In *Seismicity Patterns, their Statistical Significance and Physical Meaning*, volume 155, pages 471–507. Birkhäuser Basel, Basel, aug 1999. doi: 10.1007/978-3-0348-8677-2_14.

Krunal Parmar, Samuel Bushi, Sourangshu Bhattacharya, and Surender Kumar. Forecasting ad-impressions on online retail websites using non-homogeneous hawkes processes. In *CIKM*, 2017.

Charles L Phillips, John M Parr, Eve Ann Riskin, and Telagarapu Prabhakar. *Signals, systems, and transforms*. Prentice Hall Upper Saddle River, 2003.

Marc Piggott and Victor Solo. Non-Negative Online Estimation for Hawkes Process Networks. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2018-April, pages 6333–6337. Institute of Electrical and Electronics Engineers Inc., sep 2018. ISBN 9781538646588. doi: 10.1109/ICASSP.2018. 8462138.

Rohit Ram and Marian-Andrei Rizoiu. Conductance and Social Capital: Modeling and Empirically Measuring Online Social Influence. oct 2021. URL `http://arxiv.org/abs/2110.12569`.

Rohit Ram, Quyu Kong, and Marian-Andrei Rizoiu. Birdspotter: A Tool for Analyzing and Labeling Twitter Users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 918–921, New York, NY, USA, mar 2021. ACM. ISBN 9781450382977. doi: 10.1145/3437963.3441695. URL `https://dl.acm.org/doi/10.1145/3437963.3441695`.

Alex Reinhart. A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications. *Statistical Science*, 33(3):299–318, aug 2018. ISSN 0883-4237. doi: 10.1214/17-STS629. URL `https://projecteuclid.org/euclid.ss/1534147221`.

Marian-Andrei Rizoiu and Lexing Xie. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *International AAAI Conference on Web and Social Media (ICWSM '17)*, pages 182–191, Montréal, Québec, Canada, 2017. URL `https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553https://arxiv.org/pdf/1703.01012.pdf`.

Marian Andrei Rizoiu, Lexing Xie, Tiberio Caetano, and Manuel Cebrian. Evolution of Privacy Loss in Wikipedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 215–224, New York, New York, USA, dec 2016. ACM, ACM Press. ISBN 9781450337168. doi: 10.1145/2835776.

2835798. URL `http://dl.acm.org/citation.cfm?doid=2835776.2835798http://arxiv.org/abs/1512.03523http://dx.doi.org/10.1145/2835776.2835798`.

Marian-Andrei Rizoiu, Young Lee, and Swapnil Mishra. Hawkes processes for events in social media. *Frontiers of Multimedia Research*, pages 191–218, aug 2017a. doi: 10.1145/3122865.3122874. URL `http://arxiv.org/abs/1708.06401`.

Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be hip: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 735–744, 2017b.

Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, pages 1–10, Stanford, CA, USA, 2018a.

Marian Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. SIR-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations. In *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, pages 419–428, New York, New York, USA, apr 2018b. Association for Computing Machinery, Inc. ISBN 9781450356398. doi: 10.1145/3178876.3186108. URL `http://dl.acm.org/citation.cfm?doid=3178876.3186108`.

Sheldon M. Ross. *Stochastic Processes*. Wiley, 2nd edition, 1995. ISBN 0471120626.

Leigh Shlomovich, Edward Cohen, Niall Adams, and Lekha Patel. A Monte Carlo EM Algorithm for the Parameter Estimation of Aggregated Hawkes Processes. jan 2020. URL `http://arxiv.org/abs/2001.07160`.

Donald L. Snyder and Michael I. Miller. *Random point processes in time and space*. Springer-Verlag, 2nd edition, 1991.

Issouf Soumaré and Ernest Tafolong. Risk-based capital for credit insurers with business cycles and dynamic leverage. *Quantitative Finance*, 17(4):597–612, apr 2017. ISSN 14697696. doi: 10.1080/14697688.2016.1206960.

Jianguo Sun and Xingqiu Zhao. *Statistical Analysis of Panel Count Data*, volume 80 of *Statistics for Biology and Health*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-8714-2. doi: 10.1007/978-1-4614-8715-9.

Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, aug 2010. ISSN 0001-0782. doi: 10.1145/1787234.1787254. URL `https://dl.acm.org/doi/10.1145/1787234.1787254`.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake News Detection in Social Networks via Crowd Signals. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW*

2018, pages 517–524, New York, New York, USA, apr 2018. Association for Computing Machinery, Inc. ISBN 9781450356404. doi: 10.1145/3184558.3188722. URL `http://dl.acm.org/citation.cfm?doid=3184558.3188722`.

George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

H Juliette T Unwin, Isobel Routledge, Seth Flaxman, Marian-Andrei Rizoiu, Shengjie Lai, Justin Cohen, Daniel J Weiss, Swapnil Mishra, and Samir Bhatt. Using Hawkes Processes to model imported and local malaria cases in near-elimination settings. *PLOS Computational Biology*, 17(4):e1008830, apr 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008830. URL `http://medrxiv.org/content/early/2020/07/17/2020.07.17.20156174.abstracthttps://dx.plos.org/10.1371/journal.pcbi.1008830`.

Gerrit van Dijk. *Distribution Theory Convolution, Fourier Transform, and Laplace Transform*. De Gruyter Graduate Lectures, De Gruyter,, Berlin, 2013.

Tiandong Wang and Sidney I. Resnick. Common Growth Patterns for Regional Social Networks: a Point Process Approach. nov 2019. URL `http://arxiv.org/abs/1911.07902`.

Yichen Wang, Evangelos Theodorou, Apurv Verma, and Le Song. A Stochastic Differential Equation Framework for Guiding Information Diffusion. *arXiv preprint*, pages 1–22, 2016. URL `http://arxiv.org/abs/1603.09021`.

Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–25, nov 2019. ISSN 25730142. doi: 10.1145/3359285. URL `http://dl.acm.org/citation.cfm?doid=3371885.3359285`.

Siqi Wu, Marian Andrei Rizoiu, and Lexing Xie. Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 715–725, mar 2020. ISBN 9781577357889. URL `https://arxiv.org/abs/2003.09557`.

Hongteng Xu. Poppy: A point process toolbox based on pytorch. *arXiv*, 2018.

Hongteng Xu and Hongyuan Zha. Thap: A matlab toolkit for learning with hawkes processes. *arXiv*, 2017.

Hongteng Xu, Dixin Luo, and Hongyuan Zha. Learning Hawkes Processes from Short Doubly-Censored Event Sequences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3831–3840. JMLR.org, 2017.

Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. Uncovering and predicting the dynamic process of information cascades with survival model. *Knowledge and Information Systems*, 50(2):633–659, feb 2017. ISSN 02193116. doi: 10.1007/s10115-016-0955-7.

Ali Zarezade, Sina Jafarzadeh, and Hamid R. Rabiee. Spatio-Temporal Modeling of Users' Check-ins in Location-Based Social Networks. nov 2016. URL http://arxiv.org/abs/1611.07710.

Ali Zarezade, Utkarsh Upadhyay, Hamid R. Rabiee, and Manuel Gomez-Rodriguez. RedQueen: An online algorithm for smart broadcasting in social networks. In *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pages 51–60. Association for Computing Machinery, Inc, feb 2017. ISBN 9781450346757. doi: 10.1145/3018661.3018684.

Ali Zarezade, Abir De, Utkarsh Upadhyay, Hamid R. Rabiee, and Manuel Gomez-Rodriguez. Steering Social Activity: A Stochastic Optimal Control Point of View. *Journal of Machine Learning Research*, 18(205):1–35, 2018. URL http://jmlr.org/papers/v18/17-416.html.

Rui Zhang, Christian Walder, Marian-Andrei Rizoiu, and Lexing Xie. Efficient Non-parametric Bayesian Hawkes Processes. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4299–4305, California, aug 2019. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/597.

Rui Zhang, Christian Walder, Edwin V. Bonilla, Marian-Andrei Rizoiu, and Lexing Xie. Quantile Propagation for Wasserstein-Approximate Gaussian Processes. In H. Larochelle Lin, M. Ranzato, R. Hadsell, M. F. Balcan, and H., editors, *Conference on Neural Information Processing Systems (NeurIPS'20)*, pages 21566—-21578. Curran Associates, Inc., 2020a. URL https://arxiv.org/abs/1912.10200.

Rui Zhang, Christian Walder, and Marian-Andrei Rizoiu. Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6803–6810, apr 2020b. ISSN 2374-3468. doi: 10.1609/aaai.v34i04.6160.

Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2015-August, pages 1513–1522. Association for Computing Machinery, aug 2015. ISBN 9781450336642. doi: 10.1145/2783258.2783401.

Liang Zhu, Xinwei Tong, Jianguo Sun, Manhua Chen, Deo Kumar Srivastava, Wendy Leisenring, and Leslie L Robison. Regression analysis of mixed recurrent-event and panel-count data. *Biostatistics*, 15(3):555–568, 2014.

Lorenzo Zino, Alessandro Rizzo, and Maurizio Porfiri. Modeling memory effects in activity-driven networks. *SIAM Journal on Applied Dynamical Systems*, 17(4):2830–2854, 2018. ISSN 15360040. doi: 10.1137/18M1171485.

Lorenzo Zino, Alessandro Rizzo, and Maurizio Porfiri. Analysis and control of epidemics in temporal networks with self-excitement and behavioral changes. *European Journal of Control*, 54:1–11, jul 2020. ISSN 09473580. doi: 10.1016/j.ejcon.2019.12.007.

Contents (Appendix)

A. Proofs and Derivations

Here, we present the proofs and derivations that we omitted in the main text for readability reasons.

Proof [of Eq. (6) — which derives $\xi(t)$, the expected intensity of the Hawkes process and the event intensity of MBPP.] We shall show that for a Hawkes process,

$$\xi(t) = \mu(t) + \int_0^t \phi(s) \cdot \xi(t - s)\, \mathrm{d}s.$$

Similar identities also exist (Laub et al., 2015, pg. 8), (Helmstetter and Sornette, 2002c, Equation 9). Invoking an idea from stochastic calculus (see for instance, Lipster and Shiryaev or Klebaner), let us define the martingale

$$M(\cdot) \doteq N(\cdot) - \Lambda(\cdot),$$

where $\Lambda(\cdot) = \int_0^t \lambda(s)\, \mathrm{d}s$ is a compensator. We have:

$$
\begin{aligned}
\xi(t) &= \mathbb{E}\left[\lambda(t)\right] \\
&= \mathbb{E}\left[\mu(t) + \int_0^{t^-} \phi(t - s)\, \mathrm{d}N(s)\right] \\
&= \mathbb{E}\left[\mu(t) + \int_0^{t^-} \phi(t - s)\left[\mathrm{d}N(s) - \lambda(s)ds\right] + \int_0^{t^-} \phi(t - s)\lambda(s)\, \mathrm{d}s\right] \\
&= \mu(t) + \mathbb{E}\left[\int_0^{t^-} \phi(t - s)\left(\mathrm{d}N(s) - \lambda(s)\mathrm{d}s\right) + \int_0^{t^-} \phi(t - s)\lambda(s)\, \mathrm{d}s\right]
\end{aligned}
$$

$$= \mu(t) + \mathbb{E}\left[\int_0^{t^-} \phi(t-s)\left(dN(s) - \lambda(s)ds\right) + \int_0^{t^-} \phi(t-s)\lambda(s)\,ds\right]$$

$$= \mu(t) + \mathbb{E}\left[\int_0^{t^-} \phi(t-s)\lambda(s)\,ds\right].$$

Since the term is $\int_s^t \phi(t-s)(dN(s) - \lambda(s))ds$ is a martingale and has expectation 0,

$$= \mu(t) + \int_0^{t^-} \phi(t-s)\mathbb{E}\left[\lambda(s)\right]ds$$

$$= \mu(t) + \int_0^{t^-} \phi(t-s)\xi(s)\,ds$$

$$= \mu(t) + \int_0^{t^-} \phi(s)\xi(t-s)\,ds.$$

$\blacksquare$

Proof [of Theorem 1 — which derives the solution for $\xi(t)$ using the Laplace transform.] The closed form MBPP equation is given by

$$\xi(t) = s(t) + \int_0^t \phi(t-\tau)\cdot\xi(\tau)\,d\tau = s(t) + (\phi * \xi)(t).$$

Applying the Laplace transform to both sides, we have,

$$\mathcal{L}\{\xi(t)\}(\omega) = \mathcal{L}\{s(t) + (\phi * \xi)(t)\}(\omega)$$
$$= \mathcal{L}\{s(t)\}(\omega) + \mathcal{L}\{(\phi * \xi)(t)\}(\omega)$$
$$= \mathcal{L}\{s(t)\}(\omega) + \mathcal{L}\{\phi(t)\}(\omega)\cdot\mathcal{L}\{\xi(t)\}(\omega).$$

Solving for $\mathcal{L}\{\xi(t)\}(\omega)$, we have

$$\mathcal{L}\{\xi(t)\}(\omega) = \frac{\mathcal{L}\{s(t)\}(\omega)}{1 - \mathcal{L}\{\phi(t)\}(\omega)}.$$

Thus,

$$\xi(t) = \mathcal{L}^{-1}\left\{\frac{\mathcal{L}\{s(t)\}(\omega)}{1 - \mathcal{L}\{\phi(t)\}(\omega)}\right\}(t).$$

$\blacksquare$

Proof [of Proposition 8 — which links the IC-LL loss function and the KL divergence.] Let $\varphi(z) \doteq z\log z$ define a convex function. For clarity, further define $\mathsf{C}_i \doteq \mathsf{C}(o_{i-1}, o_i]$ and $\Xi_i \doteq \Xi(o_{i-1}, o_i]$. We may compute the loss composed of the sum of Bregman divergences with convex generator φ:

$$\mathscr{L}(\theta) = \sum_{i=1}^m B_\varphi(\mathsf{C}_i, \Xi_i)$$

60

$$= \sum_{i=1}^{m} [\varphi(\mathsf{C}_i) - \varphi(\Xi_i) - \varphi\prime(\Xi_i) \cdot (\mathsf{C}_i - \Xi_i)]$$

$$= \sum_{i=1}^{m} [\mathsf{C}_i \log \mathsf{C}_i - \Xi_i \log \Xi_i - (\log \Xi_i + 1) \cdot (\mathsf{C}_i - \Xi_i)]$$

$$= \sum_{i=1}^{m} [\mathsf{C}_i \log \mathsf{C}_i - \mathsf{C}_i + \Xi_i - \mathsf{C}_i \log \Xi_i]$$

$$= \sum_{i=1}^{m} \Xi_i - \sum_{i=1}^{m} \mathsf{C}_i \log \Xi_i + \sum_{i=1}^{m} \mathsf{C}_i (\log \mathsf{C}_i - 1).$$

As $\sum_{i=1}^{m} \mathsf{C}_i (\log \mathsf{C}_i - 1)$ is just a constant with respect to θ, the minimisation of $\mathscr{L}(\theta)$ with respect to θ, is equivalent to

$$\mathscr{L}_{IC-LL}(\theta) = \sum_{i=1}^{m} \Xi_i - \sum_{i=1}^{m} \mathsf{C}_i \log \Xi_i.$$

∎

Proof [of Eq. (36) — which shows the equivalence between IC-LL and SSE.]

Let $\hat{\varphi}(z) \doteq z^2$ define a convex function. For clarity, further define $\mathsf{C}_i \doteq \mathsf{C}(o_{i-1}, o_i]$ and $\Xi_i \doteq \Xi(o_{i-1}, o_i]$.

$$\mathscr{L}(\theta) = \sum_{i=1}^{m} B_{\hat{\varphi}}(\mathsf{C}_i, \Xi_i) = \sum_{i=1}^{m} [\mathsf{C}_i^2 - \Xi_i^2 - 2\Xi_i(\mathsf{C}_i - \Xi_i)] = \sum_{i=1}^{m} [\mathsf{C}_i - \Xi_i]^2$$

which is exactly the SSE between C_i and Ξ_i.

∎

Proof [of Proposition 6 — both the lower-bound counting process and upper-bound counting process converge to the original counting process as the discretization points increase.]

To consider the convergence of our counting process approximation, we further assume that the approximation points are equidistant apart

$$d_j^D = \frac{T}{D} j.$$

We explicitly add the superscript D is to denote the number of approximation points in the discretization sequence. Further we let $\mathcal{D}^D = \left\{ d_j^D \right\}_{j=0}^{D}$ as the set of approximation points.

We consider the convergence of the counting functions in L^1-space, as counting processes are simple functions by definition. In particular, for the lower-bound counting process we aim to show that

$$\lim_{D \to \infty} \int_0^T |M(t) - M_D^-(t)| \, \mathrm{d}t = 0.$$

61

However, as $M_D^-(t)$ is a lower-bound to $M(t)$ and by linearity of integration we can instead consider

$$\int_0^T |M(t) - M_D^-(t)|\mathrm{d}t = \int_0^T M(t)\mathrm{d}t - \int_0^T M_D^-(t)\mathrm{d}t.$$

For the original counting process $M(t)$ we assume that we have an arbitrary realisation $\mathcal{T} = \{t_i\}_{i=1}$ over a time period $t_i \in (0, T]$. Thus we consider the integral of the counting process $M(t)$ as the following,

$$
\begin{aligned}
\int_0^T M(t)\mathrm{d}t &= \int_0^T \left[\sum_{s \in [0,T]} M(\{s\}) \cdot [\![s \le t]\!] \right] \mathrm{d}t \\
&= \sum_{s \in [0,T]} M(\{s\}) \cdot \int_s^T \mathrm{d}t \\
&= \sum_{s \in [0,T]} M(\{s\}) \cdot (T - s) \\
&= \sum_{t_i \in \mathcal{T}} \delta(t_i) \cdot (T - t_i).
\end{aligned}
$$

For the lower-bound counting process, we first introduce indexing sets $I_j^D = \{i : t_i \in (d_{j-1}^D, d_j^D]\}$ which are the indices of the event times which occur in the half-open interval $(d_{j-1}^D, d_j^D]$. Thus we have,

$$
\begin{aligned}
\int_0^T M_D^-(t)\mathrm{d}t &= \int_0^T \left[\sum_{j=1}^{2^D} M(d_{j-1}^D, d_j^D] \cdot [\![d_j^D < t \le d_D]\!] \right] \mathrm{d}t \\
&= \int_0^T \left[\sum_{j=1}^{2^D} M(d_{j-1}^D, d_j^D] \cdot [\![d_j^D < t \le T]\!] \right] \mathrm{d}t \\
&= \sum_{j=1}^{2^D} M(d_{j-1}^D, d_j^D] \cdot \int_{d_j^D}^T \mathrm{d}t \\
&= \sum_{j=1}^{2^D} M(d_{j-1}^D, d_j^D] \cdot (T - d_j^D) \\
&= \sum_{j=1}^{2^D} \left(\sum_{i \in I_j^D} \delta(t_i) \right) \cdot (T - d_j^D).
\end{aligned}
$$

However, we know that $(d_{j-1}^D, d_j^D]$ will become a singleton set as $D \to \infty$ from the uniformity of discretization points. Further we know that for any D, the union $\bigcup_{j=1}^{2^D}(d_{j-1}^D, d_j^D] = (0, T]$. Thus in the limit of $D \to \infty$, we have $\{(d_{j-1}^D, d_j^D] : i \in \{1, 2, \ldots, 2^D\}\} \to \{\{t\} : t \in$

$(0, T]$}. Thus, we compute the limit as the following,

$$\lim_{D\to\infty} \int_0^T M_D^-(t)\mathrm{d}t = \lim_{D\to\infty} \left(\sum_{j=1}^{2^D} \sum_{i\in I_j^D} \delta(t_i) \cdot \left(T - d_j^D\right) \right)$$

$$= \sum_{t\in(0,T]} \delta(t) \cdot [\![t \in \mathcal{T}]\!] \cdot (T - t)$$

$$= \sum_{t_i\in\mathcal{T}} \delta(t_i) \cdot (T - t_i)$$

$$= \int_0^T M(t)\mathrm{d}t.$$

Thus,

$$\lim_{D\to\infty} \int_0^T |M(t) - M_D^-(t)|\mathrm{d}t = \lim_{D\to\infty} \left(\int_0^T M(t)\mathrm{d}t - \int_0^T M_D^-(t)\mathrm{d}t \right)$$

$$= \lim_{D\to\infty} \int_0^T M(t)\mathrm{d}t - \lim_{D\to\infty} \int_0^T M_D^-(t)\mathrm{d}t$$

$$= \int_0^T M(t)\mathrm{d}t - \lim_{D\to\infty} \int_0^T M_D^-(t)\mathrm{d}t$$

$$= \int_0^T M(t)\mathrm{d}t - \int_0^T M(t)\mathrm{d}t$$

$$= 0.$$

The proof is identical when the upper-bound approximation $M_D^+(t)$ is chosen instead. ∎

Proof [of Eq. (25) – derivation of the upper bound approximation of the MBPP compensator $\Xi(t)$.] We consider the following compensator generated by the upper-bound counting process $M_D^+(t)$:

$$\Xi^+(t) \doteq S(t) + \int_0^t \mathbb{E}[M_D^+(y)]\phi(t - y)\mathrm{d}y$$

$$= S(t) + \int_0^t \mathbb{E}\left[\sum_{j=1}^D M(d_{j-1}, d_j] \cdot [\![d_{j-1} < y \le d_D]\!] \right] \phi(t - y)\mathrm{d}y$$

$$= S(t) + \sum_{j=1}^D \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_0^t [\![d_{j-1} < y \le d_D]\!]\phi(t - y)\mathrm{d}y$$

$$= S(t) + \sum_{j=1}^{\bar{D}(t)+1} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{d_{j-1}}^{\min(t,d_D)} \phi(t - y)\mathrm{d}y,$$

where $\bar{D}(t) = \max\{i \in \{1, \ldots, D\} : d_i < t\}$. This is a similar equation to Eq. (24). ∎

Proof [of Proposition 7 — derivation of $\Xi_{\bar{D}}^{\bar{=}}(o_{i-1}, o_i]$ the approximation of the MBPP compensator over an interval] We assume that the anti-derivative of the kernel function $\phi = \Phi'$ exists. Thus we have

$$\Xi_{\bar{D}}^{\bar{=}}(o_{i-1}, o_i] = \Xi_{\bar{D}}^{-}(o_i) - \Xi_{\bar{D}}^{-}(o_{i-1})$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{d_j}^{\min(o_i, d_D)} \phi(o_i - y) \mathrm{d}y$$

$$- \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{d_j}^{\min(o_{i-1}, d_D)} \phi(o_{i-1} - y) \mathrm{d}y$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{d_j}^{o_i} \phi(o_i - y) \mathrm{d}y$$

$$- \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{d_j}^{o_{i-1}} \phi(o_{i-1} - y) \mathrm{d}y$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \left[\int_{d_j}^{o_i} \phi(o_i - y) \mathrm{d}y - \int_{d_j}^{o_{i-1}} \phi(o_{i-1} - y) \mathrm{d}y \right]$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{d_j}^{o_i} \phi(o_i - y) \mathrm{d}y$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \left[-(\Phi(0) - \Phi(o_i - d_j)) + (\Phi(0) + \Phi(o_{i-1} - d_j)) \right]$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \left[-\Phi(0) + \Phi(o_i - d_j) \right]$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \left[\Phi(o_i - d_j) - \Phi(o_{i-1} - d_j) \right]$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \left[\Phi(o_i - d_j) - \Phi(0) \right]$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{o_{i-1}-d_j}^{o_i-d_j} \phi(y) \mathrm{d}y$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j]\right] \int_{0}^{o_i-d_j} \phi(y) \mathrm{d}y$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_{i-1})} \mathbb{E}\left[M(d_{j-1}, d_j)\right] \int_{o_{i-1}-d_j}^{o_i-d_j} \phi(y)\mathrm{d}y$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j)\right] \int_{o_{i-1}-d_j}^{o_i-d_j} \phi(y)\mathrm{d}y$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j)\right] \int_{0}^{o_{i-1}-d_j} \phi(y)\mathrm{d}y$$

$$= S(o_i) - S(o_{i-1}) + \sum_{j=1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j)\right] \int_{o_{i-1}-d_j}^{o_i-d_j} \phi(y)\mathrm{d}y$$

$$+ \sum_{j=\bar{D}(o_{i-1})+1}^{\bar{D}(o_i)} \mathbb{E}\left[M(d_{j-1}, d_j)\right] \int_{0}^{o_{i-1}-d_j} \phi(y)\mathrm{d}y.$$

$\blacksquare$

Table 10: Solutions for different immigrant intensities $s(t)$, and the exponentially time-decaying kernel in Eq. (13) of the main text. Ψ_T acts as a set of immigrant event times.

	$s(t)$	MBPP Intensity Function $\xi(t)$
I	$\delta(t-a)$	$\delta(t-a) + \kappa\theta\exp((k-1)\theta(t-a))\cdot[\![t>a]\!]$
II	$\sum_{\mathfrak{s}_z\in\Psi_T}\delta(t-\mathfrak{s}_z)$	$\sum_{\mathfrak{s}_z\in\Psi_T}\left[\delta(t-\mathfrak{s}_z)+\kappa\theta\exp((k-1)\theta(t-\mathfrak{s}_z))\cdot[\![t>\mathfrak{s}_z]\!]\right]$
III	$[\![a\le t<b]\!]$	$\frac{\kappa}{1-\kappa}\left[\frac{1}{k}-\exp((k-1)\theta(t-a))\right]\cdot[\![a<t\le b]\!]+\frac{\kappa}{1-\kappa}\left[\exp((k-1)\theta(t-b))-\exp((k-1)\theta(t-a))\right]\cdot[\![t>b]\!]$
IV	$\sum_{i=1}^m \frac{S(q_{i-1},q_i)}{q_i-q_{i-1}}\cdot[\![q_{i-1}<t\le q_i]\!]$	$\sum_{i=1}^m \frac{S(q_{i-1},q_i)}{q_i-q_{i-1}}\left[\frac{\kappa}{1-\kappa}\left[\frac{1}{k}-\exp((k-1)\theta(t-q_{i-1}))\right]\cdot[\![q_{i-1}<t\le q_i]\!]+\frac{\kappa}{1-\kappa}\left[\exp((k-1)\theta(t-q_i))-\exp((k-1)\theta(t-q_{i-1}))\right]\cdot[\![t>q_i]\!]\right]$
V	$\kappa\theta+(u_0-\kappa\theta)\exp(-\theta t)$	$\frac{\kappa\theta}{1-\kappa}\left(1-\exp(-(1-\kappa)\theta t)\right)+u_0\exp(-(1-\kappa)\theta t)$
VI	$\sin(t)+\alpha$	$-\frac{\alpha}{\kappa-1}+\frac{\kappa}{\kappa-1}\frac{\alpha+\alpha\theta^2-2\alpha\kappa\theta^2+\alpha\kappa^2\theta^2+\theta\kappa-\theta}{1+\theta^2-2\kappa\theta^2+\kappa^2\theta^2}\exp((k-1)\theta t)+\frac{\sin(t)+\theta^2\sin(t)-\kappa\theta^2\sin(t)-\kappa\theta\cos(t)}{1+\theta^2-2\kappa\theta^2+\kappa^2\theta^2}$

B. MBPP Calculations

B.1 MBPP Intensity Functions

We consider the closed form solutions of an MBPP intensity function $\xi(t)$ with exponential kernel for various immigrant intensity functions. We use Corollary 3 where we have

$$h(t) = k\theta \exp((k-1)\theta t) \cdot [\![t > 0]\!].$$

and

$$E(t) = \delta(t) + h(t)$$
$$\xi(t) = (E * s)(t) = s(t) + (h * s)(t).$$

Specifically, we calculate each row of Table 10.

Row I. Consider $s(t) = \delta(t - a)$ for $a \geq 0$. Thus,

$$(h * s)(t) = (h(t) * \delta(t - a))$$
$$= h(t - a) \cdot [\![t > a]\!].$$

Together we have,

$$\xi(t) = \delta(t - a) + \kappa\theta \exp((k-1)\theta(t-a)) \cdot [\![t > a]\!].$$

Row II. Consider $s(t) = \sum_{\mathfrak{s}_z \in \Psi_T} \delta(t - \mathfrak{s}_z)$ for some set of events Ψ_T. As the MBPP intensity is a LTI (Rizoiu et al. (2017b)[Corollary 2.2]), we can simply express the MBPP intensity as the sum of Row I.

Together we have,

$$\xi(t) = \sum_{\mathfrak{s}_z \in \Psi_T} \left[\delta(t - \mathfrak{s}_z) + \kappa\theta \exp((k-1)\theta(t - \mathfrak{s}_z)) \cdot [\![t > \mathfrak{s}_z]\!] \right].$$

Row III. Consider $s(t) = [\![a < t \leq b]\!]$ an interval. Then,

$$(h * s)(t) = (h(t) * [\![a < t \leq b]\!]).$$

We break the calculation into three cases: (1) $t \leq a$; (2) $a < t \leq b$; and (3) $t > b$.
Case (1): Suppose that $t \leq a$, then simply $(h(t) * [\![a < t \leq b]\!]) = 0$.
Case (2): Suppose that $a < t \leq b$. Then we have,

$$(h(t) * [\![a < t \leq b]\!]) = \int_0^\infty [\![a < \tau \leq b]\!] \cdot h(t - \tau) \, d\tau$$
$$= \int_a^t h(t - \tau) \, d\tau$$
$$= \frac{\kappa\theta}{(1 - \kappa)\theta} \exp((k-1)\theta(t - \tau)) \Big]_{\tau=a}^{\tau=t}$$
$$= \frac{\kappa}{1 - \kappa} [1 - \exp((k-1)\theta(t - a))].$$

<u>Case (3)</u>: Suppose that $t > b$. Then we have,

$$
\begin{aligned}
(h(t) * [\![a < t \le b]\!]) &= \int_0^\infty [\![a < \tau \le b]\!] \cdot h(t - \tau) \, d\tau \\
&= \int_a^b h(t - \tau) \, d\tau \\
&= \frac{\kappa \theta}{(1 - \kappa)\theta} \cdot \exp((k - 1)\theta(t - \tau)) \Big]_{\tau = a}^{\tau = b} \\
&= \frac{\kappa}{1 - \kappa} \left[\exp((k - 1)\theta(t - b)) - \exp((k - 1)\theta(t - a)) \right].
\end{aligned}
$$

Thus together we have that

$$
(h(t) * [\![a < t \le b]\!]) = \begin{cases} 0 & t \le a \\ \frac{\kappa}{1-\kappa} [1 - \exp((k - 1)\theta(t - a))] & a < t \le b \,. \\ \frac{\kappa}{1-\kappa} [\exp((k - 1)\theta(t - b)) - \exp((k - 1)\theta(t - a))] & t > b \end{cases}
$$

Giving,

$$
\begin{aligned}
\xi(t) = {}& [\![a < t \le b]\!] + \frac{\kappa}{1 - \kappa} [1 - \exp((k - 1)\theta(t - a))] \cdot [\![a < t \le b]\!] \\
&+ \frac{\kappa}{1 - \kappa} [\exp((k - 1)\theta(t - b)) - \exp((k - 1)\theta(t - a))] \cdot [\![t > b]\!] \\
= {}& \frac{\kappa}{1 - \kappa} \left[\frac{1}{k} - \exp((k - 1)\theta(t - a)) \right] \cdot [\![a < t \le b]\!] \\
&+ \frac{\kappa}{1 - \kappa} [\exp((k - 1)\theta(t - b)) - \exp((k - 1)\theta(t - a))] \cdot [\![t > b]\!].
\end{aligned}
$$

Row IV. Consider $s(t) = \sum_{i=1}^m \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \cdot [\![q_{i-1} < t \le q_i]\!]$ for a set of intervals with non-negative immigrant volumes $\mathsf{S}(q_{i-1}, q_i]$. As the MBPP intensity is an LTI (Rizoiu et al. (2017b)[Corollary 2.2]), we can simply express the MBPP intensity as the weighted sum of Row III.

$$
\begin{aligned}
\xi(t) = \sum_{i=1}^m \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \Bigg[&\frac{\kappa}{1 - \kappa} \left[\frac{1}{k} - \exp((k - 1)\theta(t - q_{i-1})) \right] \cdot [\![q_{i-1} < t \le q_i]\!] \\
&+ \frac{\kappa}{1 - \kappa} [\exp((k - 1)\theta(t - q_i)) - \exp((k - 1)\theta(t - q_{i-1}))] \cdot [\![t > q_i]\!] \Bigg].
\end{aligned}
$$

Row V. Proof shown in the main text.

Row VI. Consider $s(t) = \sin(t) + \alpha$. Then

$$
(h * s)(t) = (h * \alpha)(t) + (h * \sin)(t).
$$

For $(h * \alpha)(t)$ we have,

$$
(h * \alpha)(t) = \alpha \int_0^t h(\tau) \, d\tau
$$

$$= \frac{\alpha\kappa}{\kappa - 1}\left(\exp((k-1)\theta t) - 1\right)$$

$$= \frac{\alpha\kappa}{\kappa - 1}\exp((k-1)\theta t) - \frac{\alpha\kappa}{\kappa - 1}.$$

For $(h * \sin)(t)$ first consider the indefinite integral

$$\triangle = \int \sin(\tau)\exp(a(t-\tau))\mathrm{d}\tau.$$

By integration by parts,

$$-\frac{1}{a}\sin(\tau)\exp(a(t-\tau)) + \frac{1}{a}\int\cos(\tau)\exp(a(t-\tau))\mathrm{d}\tau.$$

Then by integration by parts again,

$$-\frac{1}{a}\sin(\tau)\exp(a(t-\tau)) + \frac{1}{a}\left[-\frac{1}{a}\cos(\tau)\exp(a(t-\tau)) - \frac{1}{a}\int\sin(\tau)\exp(a(t-\tau))\mathrm{d}\tau\right]$$

$$= -\frac{1}{a}\sin(\tau)\exp(a(t-\tau)) - \frac{1}{a^2}\cos(\tau)\exp(a(t-\tau)) - \frac{1}{a^2}\int\sin(\tau)\exp(a(t-\tau))\mathrm{d}\tau$$

$$= -\exp(a(t-\tau))\left[\frac{1}{a}\sin(\tau) + \frac{1}{a^2}\cos(\tau)\right] - \frac{1}{a^2}\int\sin(\tau)\exp(a(t-\tau))\mathrm{d}\tau$$

$$= -\exp(a(t-\tau))\left[\frac{1}{a}\sin(\tau) + \frac{1}{a^2}\cos(\tau)\right] - \frac{1}{a^2}\triangle.$$

Thus,

$$\triangle = -\exp(a(t-\tau))\left[\frac{1}{a}\sin(\tau) + \frac{1}{a^2}\cos(\tau)\right] - \frac{1}{a^2}\triangle$$

$$(\iff) \qquad \triangle = \frac{-\exp(a(t-\tau))\left[a\sin(\tau) + \cos(\tau)\right]}{a^2 + 1}.$$

Therefore for $(h * \sin)(t)$ we have

$$(h * \sin)(t) = \kappa\theta\frac{-\exp((k-1)\theta(t-\tau))\left[(k-1)\theta\sin(\tau) + \cos(\tau)\right]}{(k-1)^2\theta^2 + 1}\Bigg]_{\tau=0}^{\tau=t}$$

$$= \kappa\theta\frac{-\exp((k-1)\theta(t-\tau))\left[(k-1)\theta\sin(\tau) + \cos(\tau)\right]}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}\Bigg]_{\tau=0}^{\tau=t}$$

$$= -\kappa\theta\frac{\left[(k-1)\theta\sin(t) + \cos(t)\right]}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} + \kappa\theta\frac{\exp((k-1)\theta t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}.$$

Thus for the MBPP intensity,

$$\xi(t) = \sin(t) + \alpha + (h * \sin)(t)$$

$$= \sin(t) + \alpha + \frac{\alpha\kappa}{\kappa - 1}\exp((k-1)\theta t) - \frac{\alpha\kappa}{\kappa - 1}$$

$$- \kappa\theta\frac{\left[(k-1)\theta\sin(t) + \cos(t)\right]}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} + \kappa\theta\frac{\exp((k-1)\theta t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}.$$

Now we collect the trigonometric terms, the exponential terms, and the other constants. The trigonometric terms:

$$
\begin{aligned}
&\sin(t) - \kappa\theta\frac{[(k-1)\theta\sin(t) + \cos(t)]}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \\
&= \sin(t) - \frac{\kappa^2\theta^2\sin(t) - \kappa\theta^2\sin(t) + \kappa\theta\cos(t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \\
&= \frac{\sin(t) + \theta^2\sin(t) - 2\kappa\theta^2\sin(t) + \kappa^2\theta^2\sin(t) - \kappa^2\theta^2\sin(t) + \kappa\theta^2\sin(t) - \kappa\theta\cos(t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \\
&= \frac{\sin(t) + \theta^2\sin(t) - \kappa\theta^2\sin(t) - \kappa\theta\cos(t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}.
\end{aligned}
$$

The exponential terms:

$$
\begin{aligned}
&\frac{\alpha\kappa}{\kappa - 1}\exp((k-1)\theta t) + \kappa\theta\frac{\exp((k-1)\theta t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \\
&= \frac{\kappa}{\kappa - 1}\frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}\exp((k-1)\theta t).
\end{aligned}
$$

The constant terms:

$$
\alpha - \frac{\alpha\kappa}{\kappa - 1} = -\frac{\alpha}{\kappa - 1}.
$$

Thus together,

$$
\begin{aligned}
\xi(t) = &-\frac{\alpha}{\kappa - 1} + \frac{\kappa}{\kappa - 1}\frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}\exp((k-1)\theta t) \\
&+ \frac{\sin(t) + \theta^2\sin(t) - \kappa\theta^2\sin(t) - \kappa\theta\cos(t)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}.
\end{aligned}
$$

Table 11: Solutions for different immigrant intensities $s(t)$, and the exponentially time-decaying kernel in Eq. (13) of the main text. Ψ_T acts as a set of immigrant event times.

	$s(t)$	MBPP Compensator Function $\Xi(t)$
I	$\delta(t-a)$	$[\![t > a]\!]\left[1 + \frac{\kappa}{\kappa-1}\left(\exp((\kappa-1)\theta(t-a))-1\right)\right]$
II	$\sum_{\mathfrak{s}_z\in\Psi_T}\delta(t-\mathfrak{s}_z)$	$\sum_{\mathfrak{s}_z\in\Psi_T}[\![t > \mathfrak{s}_z]\!]\cdot\left[1 + \frac{\kappa}{\kappa-1}\left(\exp(\kappa-1)\theta(t-\mathfrak{s}_z))-1\right)\right]$
III	$[\![a \le t < b]\!]$	$(t-a)\cdot\frac{\kappa}{1-\kappa}\left[\frac{1}{k}-\exp((k-1)\theta(t-a))\right]\cdot[\![a < t \le b]\!] + \frac{\kappa t}{1-\kappa}\left[\exp((k-1)\theta(t-b)) - \exp((k-1)\theta(t-a))\right]\cdot[\![t > b]\!] + \frac{\kappa b}{1-\kappa}\left[\frac{1}{k}-\exp((k-1)\theta(t-b))\right]\cdot[\![t > b]\!] + \frac{\kappa a}{1-\kappa}\left[\exp((k-1)\theta(t-a)) - \frac{1}{k}\right]\cdot[\![t > b]\!]$
IV	$\sum_{i=1}^m \frac{S(q_{i-1},q_i)}{q_i-q_{i-1}}\cdot[\![q_{i-1} < t \le q_i]\!]$	$\frac{(t-q_{m^\star(t)})S(q_{m^\star(t)+1},q_{m^\star(t)})}{q_{m^\star(t)+1}-q_{m^\star(t)}} + \sum_{i=1}^{m^\star(t)}S(q_{i-1},q_i) + \sum_{i=1}^{m^\star(t)}\frac{S(q_{i-1},q_i)}{q_i-q_{i-1}}\frac{\kappa}{1-\kappa}\left[\exp((\kappa-1)\theta(t-q_i)) - 1 - \exp((\kappa-1)\theta(t-q_{i-1})) + \exp((\kappa-1)\theta(q_i-q_{i-1}))\right] + \frac{\kappa}{1-\kappa}\left[t-q_{m^\star(t)}\right. - \frac{1}{\theta(\kappa-1)}\exp((k-1)\theta(t-q_{m^\star(t)}))\,d\tau\right]$, where $m^\star(t)=1+\max\{i \mid q_i > t\}$
V	$\kappa\theta + (u_0 - \kappa\theta)\exp(-\theta t)$	$\frac{\kappa\theta t}{(1-\kappa)} + \frac{u_0-1}{(1-\kappa)\theta}(1-\exp(-(1-\kappa)\theta t))$
VI	$\sin(t) + \alpha$	$-\frac{\alpha t}{\kappa-1} + \frac{\kappa}{(\kappa-1)^2\theta} + \frac{\alpha+\alpha\theta^2-2\alpha\kappa\theta^2+\alpha\kappa^2\theta^2+\theta\kappa-\theta}{1+\theta^2-2\kappa\theta^2+\kappa^2\theta^2}(\exp((k - 1)\theta\tau) - 1) + \frac{-\cos(t)-\theta^2\cos(t)+\kappa\theta^2\cos(t)-\kappa\theta\sin(t)+1+\theta^2-\kappa\theta^2}{1+\theta^2-2\kappa\theta^2+\kappa^2\theta^2}$

71

B.2 MBPP Compensator Functions

We similarly calculate the corresponding MBPP compensator function of the MBPP intensity function we looked at above. In particular we similarly calculate each row of Table 11.

Row I. Consider $s(t) = \delta(t - a)$ for $a \geq 0$. We simply integrate the result of Row I in Table 10,

$$
\begin{aligned}
\Xi(t) &= \int_0^t \xi(\tau)\mathrm{d}\tau \\
&= \int_0^t \left(\delta(\tau - a) + \kappa\theta \exp((\kappa - 1)\theta(\tau - a)) \cdot [\![\tau > a]\!] \right) \mathrm{d}\tau \\
&= \int_0^t \delta(\tau - a)\mathrm{d}\tau + \int_0^t \kappa\theta \exp((\kappa - 1)\theta(\tau - a)) \cdot [\![\tau > a]\!] \mathrm{d}\tau \\
&= [\![t > a]\!] + [\![t > a]\!] \cdot \int_a^t \kappa\theta \exp((\kappa - 1)\theta(\tau - a))\mathrm{d}\tau \\
&= [\![t > a]\!] + [\![t > a]\!] \cdot \left. \frac{\kappa}{\kappa - 1} \exp((\kappa - 1)\theta(\tau - a)) \right]_{\tau = a}^{\tau = t} \\
&= [\![t > a]\!] + [\![t > a]\!] \cdot \frac{\kappa}{\kappa - 1} \left(\exp((\kappa - 1)\theta(t - a)) - 1 \right) \\
&= [\![t > a]\!] \cdot \left[1 + \frac{\kappa}{\kappa - 1} \left(\exp((\kappa - 1)\theta(t - a)) - 1 \right) \right].
\end{aligned}
$$

Row II. Consider $s(t) = \sum_{\mathfrak{s}_z \in \Psi_T} \delta(t - \mathfrak{s}_z)$ for some set of events Ψ_T. As MBPP compensator is LTI, we compute the weighted sum of Row II.

$$
\Xi(t) = \sum_{\mathfrak{s}_z \in \Psi_T} [\![t > \mathfrak{s}_z]\!] \cdot \left[1 + \frac{\kappa}{\kappa - 1} \left(\exp((\kappa - 1)\theta(t - \mathfrak{s}_z)) - 1 \right) \right].
$$

Row III. Consider $s(t) = [\![a < t \leq b]\!]$ an interval. First we consider the integral of $s(t)$.

$$
S(t) = \int_0^t s(\tau)\mathrm{d}\tau = (t - a) \cdot [\![a < t \leq b]\!] + (b - a) \cdot [\![t > b]\!],
$$

which can be simply computed by considering different values of t.

As the MBPP compensator is a LTI, the result can be calculated as the weighted sum of of Row III in Table 10.

The first term,

$$
\begin{aligned}
(t - a) \cdot \frac{\kappa}{1 - \kappa} & \left[\frac{1}{k} - \exp((k - 1)\theta(t - a)) \right] \cdot [\![a < t \leq b]\!] \\
& + (t - a) \cdot \frac{\kappa}{1 - \kappa} \left[\exp((k - 1)\theta(t - b)) - \exp((k - 1)\theta(t - a)) \right] \cdot [\![t > b]\!].
\end{aligned}
$$

The second term,

$$
(b - a) \cdot \frac{\kappa}{1 - \kappa} \left[\frac{1}{k} - \exp((k - 1)\theta(t - b)) \right] \cdot [\![t > b]\!].
$$

Together we have

$$
\begin{aligned}
\Xi(t) &= (t-a) \cdot \frac{\kappa}{1-\kappa}\left[\frac{1}{k} - \exp((k-1)\theta(t-a))\right] \cdot [\![a < t \leq b]\!] \\
&\quad + (t-a) \cdot \frac{\kappa}{1-\kappa}\left[\exp((k-1)\theta(t-b)) - \exp((k-1)\theta(t-a))\right] \cdot [\![t > b]\!] \\
&\quad + (b-a) \cdot \frac{\kappa}{1-\kappa}\left[\frac{1}{k} - \exp((k-1)\theta(t-b))\right] \cdot [\![t > b]\!] \\
&= (t-a) \cdot \frac{\kappa}{1-\kappa}\left[\frac{1}{k} - \exp((k-1)\theta(t-a))\right] \cdot [\![a < t \leq b]\!] \\
&\quad + \frac{\kappa t}{1-\kappa}\left[\exp((k-1)\theta(t-b)) - \exp((k-1)\theta(t-a))\right] \cdot [\![t > b]\!] \\
&\quad + \frac{\kappa b}{1-\kappa}\left[\frac{1}{k} - \exp((k-1)\theta(t-b))\right] \cdot [\![t > b]\!] \\
&\quad + \frac{\kappa a}{1-\kappa}\left[\exp((k-1)\theta(t-a)) - \frac{1}{k}\right] \cdot [\![t > b]\!]
\end{aligned}
$$

Row IV. To calculate this row, one could utilise the LTI with Row III. However, we derive this differently, which provides an equation which is easier to implement.

First, let us calculate the exogenous component.

$$
\begin{aligned}
S(t) &= \int_0^t s(\tau)\, d\tau \\
&= \int_0^t \left(\sum_{i=1}^m \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \cdot [\![q_{i-1} < \tau \leq q_i]\!]\right) d\tau \\
&= \sum_{i=1}^m \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \cdot \int_0^t [\![q_{i-1} < \tau \leq q_i]\!]\, d\tau
\end{aligned}
$$

The integral within the summation is what we need to calculate now $\int_0^t [\![q_{i-1} < \tau \leq q_i]\!]\, d\tau$. We split up the calculation into three cases: (1) $t \leq q_{i-1}$; (2) $q_{i-1} < t \leq q_i$; and (3) $t > q_i$.

Case (1): Suppose that $t \leq q_{i-1}$, then simply $\int_0^t [\![q_{i-1} < \tau \leq q_i]\!]\, d\tau = 0$.

Case (2): Suppose that $q_{i-1} < t \leq q_i$,

$$
\int_0^t [\![q_{i-1} < \tau \leq q_i]\!]\, d\tau = \int_{q_{i-1}}^t 1\, d\tau = t - q_{i-1}.
$$

Case (3): Suppose that $t > q_i$,

$$
\int_0^t [\![q_{i-1} < \tau \leq q_i]\!]\, d\tau = \int_{q_{i-1}}^{q_i} 1\, d\tau = q_i - q_{i-1}.
$$

For arbitrary fixed t, let $m^\star(t) = 1 + \max\{i \mid q_i > t\}$. Then, by combining the three cases we have that,

$$
S(t) = \frac{(t - q_{m^\star(t)})\mathsf{S}(q_{m^\star(t)+1} - q_{m^\star(t)})}{q_{m^\star(t)+1} - q_{m^\star(t)}} + \sum_{i=1}^{m^\star(t)} \mathsf{S}(q_{i-1}, q_i].
$$

73

Similarly to the exogenous function, we consider the endogenous response in the same three cases. We directly use the piecewise derivations found in Row III of Table 10. Thus, we only need to compute $\int_0^t (h * s)\,\mathrm{d}\tau$. First we consider $\int_0^t (h(\tau) * [\![q_{i-1} < \tau \le q_i]\!])\,\mathrm{d}\tau$

Case (1): Suppose that $t \le q_{i-1}$, then simply $\int_0^t (h(\tau) * [\![q_{i-1} < \tau \le q_i]\!]) [\![t \le q_{i-1}]\!]\,\mathrm{d}\tau = 0$.

Case (2): Suppose that $q_{i-1} < t \le q_i$,

$$\int_0^t (h(\tau) * [\![q_{i-1} < \tau \le q_i]\!]) [\![q_{i-1} < \tau \le q_i]\!]\,\mathrm{d}\tau$$

$$= \int_{q_{i-1}}^t (h * s)(\tau)\,\mathrm{d}\tau$$

$$= \int_{q_{i-1}}^t \frac{\kappa}{1 - \kappa} \left[1 - \exp((k - 1)\theta(\tau - q_{i-1})) \right]\,\mathrm{d}\tau$$

$$= \frac{\kappa}{1 - \kappa} \left[t - q_{i-1} - \int_{q_{i-1}}^t \exp((k - 1)\theta(\tau - q_{i-1}))\,\mathrm{d}\tau \right]$$

$$= \frac{\kappa}{1 - \kappa} \left[t - q_{i-1} + \frac{1}{\theta(\kappa - 1)} - \frac{1}{\theta(\kappa - 1)} \exp((k - 1)\theta(t - q_{i-1}))\,\mathrm{d}\tau \right].$$

Case (3): Suppose that $t > q_i$,

$$\int_0^t (h(\tau) * [\![q_{i-1} < \tau \le q_i]\!]) [\![\tau > q_i]\!]\,\mathrm{d}\tau$$

$$= \int_{q_i}^t (h * s)(\tau)\,\mathrm{d}\tau$$

$$= \int_{q_i}^t \frac{\kappa}{1 - \kappa} \left[\exp((\kappa - 1)\theta(\tau - q_i)) - \exp((\kappa - 1)\theta(\tau - q_{i-1})) \right]\,\mathrm{d}\tau$$

$$= \frac{\kappa}{1 - \kappa} \left[\exp((\kappa - 1)\theta(t - q_i)) - 1 - \exp((\kappa - 1)\theta(t - q_{i-1})) \right.$$

$$\left. + \exp((\kappa - 1)\theta(q_i - q_{i-1})) \right].$$

Thus, combining the three cases, we have that,

$$\int_0^t (h * s)(\tau)\,\mathrm{d}\tau$$

$$= \sum_{i=1}^m \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \int_0^t (h(t) * [\![q_{i-1} < t \le q_i]\!])\,\mathrm{d}\tau$$

$$= \sum_{i=1}^{m^\star(t)} \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \frac{\kappa}{1 - \kappa} \left[\exp((\kappa - 1)\theta(t - q_i)) - 1 - \exp((\kappa - 1)\theta(t - q_{i-1})) \right.$$

$$\left. + \exp((\kappa - 1)\theta(q_i - q_{i-1})) \right] +$$

$$\frac{\kappa}{1 - \kappa} \left[t - q_{m^\star(t)} + \frac{1}{\theta(\kappa - 1)} - \frac{1}{\theta(\kappa - 1)} \exp((k - 1)\theta(t - q_{m^\star(t)}))\,\mathrm{d}\tau \right].$$

Together, we can now compute the compensator function.

$$\Xi(t) = \frac{(t - q_{m^\star(t)})\mathsf{S}(q_{m^\star(t)+1} - q_{m^\star(t)})}{q_{m^\star(t)+1} - q_{m^\star(t)}} + \sum_{i=1}^{m^\star(t)} \mathsf{S}(q_{i-1}, q_i]$$

$$+ \sum_{i=1}^{m^\star(t)} \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \frac{\kappa}{1 - \kappa} \left[\exp((\kappa - 1)\theta(t - q_i)) - 1 - \exp((\kappa - 1)\theta(t - q_{i-1})) \right.$$

$$+ \exp((\kappa - 1)\theta(q_i - q_{i-1}))]$$

$$+ \frac{\kappa}{1 - \kappa} \left[t - q_{m^\star(t)} + \frac{1}{\theta(\kappa - 1)} - \frac{1}{\theta(\kappa - 1)} \exp((k - 1)\theta(t - q_{m^\star(t)})) \, \mathrm{d}\tau \right].$$

Row V. To calculate the MBPP compensator, we integrate Row V of Table 10.

$$\Xi(t) = \int_0^t \xi(\tau)\mathrm{d}\tau$$

$$= \int_0^t \frac{\kappa\theta}{1 - \kappa} (1 - \exp(-(1 - \kappa)\theta\tau)) + u_0 \exp(-(1 - \kappa)\theta\tau)\mathrm{d}\tau$$

$$= \frac{\kappa\theta}{(1 - \kappa)} \left[\tau + \frac{1}{(1 - \kappa)\theta} \exp(-(1 - \kappa)\theta\tau) \right] - \frac{u_0}{(1 - \kappa)\theta} \exp(-(1 - \kappa)\theta\tau) \Big]_{\tau=0}^{\tau=t}$$

$$= \frac{\kappa\theta t}{(1 - \kappa)} + \frac{u_0 - 1}{(1 - \kappa)\theta}(1 - \exp(-(1 - \kappa)\theta t)).$$

Row VI. To calculate the MBPP compensator, we integrate Row VI of Table 10.

$$\Xi(t) = \int_0^t \xi(\tau)\mathrm{d}\tau = \int_0^t \Big[-\frac{\alpha}{\kappa - 1} + \frac{\kappa}{\kappa - 1} \frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \exp((k - 1)\theta\tau)$$

$$+ \frac{\sin(\tau) + \theta^2 \sin(\tau) - \kappa\theta^2 \sin(\tau) - \kappa\theta \cos(\tau)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \Big] \mathrm{d}\tau.$$

We calculate the integral term by term. For the first term,

$$\int_0^t -\frac{\alpha}{\kappa - 1}\mathrm{d}\tau = -\frac{\alpha t}{\kappa - 1}.$$

For the second term,

$$\int_0^t \frac{\kappa}{\kappa - 1} \frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \exp((k - 1)\theta\tau)\mathrm{d}\tau$$

$$= \frac{\kappa}{\kappa - 1} \frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \int_0^t \exp((k - 1)\theta\tau)\mathrm{d}\tau$$

$$= \frac{\kappa}{(\kappa - 1)^2\theta} \frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} (\exp((k - 1)\theta\tau) - 1).$$

For the third term,

$$\int_0^t \frac{\sin(\tau) + \theta^2 \sin(\tau) - \kappa\theta^2 \sin(\tau) - \kappa\theta \cos(\tau)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}\mathrm{d}\tau$$

$$= \frac{-\cos(\tau) - \theta^2 \cos(\tau) + \kappa\theta^2 \cos(\tau) - \kappa\theta \sin(\tau)}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} \Bigg]_{\tau=0}^{\tau=t}$$

$$= \frac{-\cos(t) - \theta^2 \cos(t) + \kappa\theta^2 \cos(t) - \kappa\theta \sin(t) + 1 + \theta^2 - \kappa\theta^2}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}.$$

Thus together,

$$\Xi(t) = -\frac{\alpha t}{\kappa - 1} + \frac{\kappa}{(\kappa - 1)^2 \theta} \frac{\alpha + \alpha\theta^2 - 2\alpha\kappa\theta^2 + \alpha\kappa^2\theta^2 + \theta\kappa - \theta}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2} (\exp((k-1)\theta\tau) - 1)$$

$$+ \frac{-\cos(t) - \theta^2 \cos(t) + \kappa\theta^2 \cos(t) - \kappa\theta \sin(t) + 1 + \theta^2 - \kappa\theta^2}{1 + \theta^2 - 2\kappa\theta^2 + \kappa^2\theta^2}.$$

C. Latent Homogeneous Poisson Process Exogenous Function - Probabilistic Approach

The Latent Homogenous Poisson Process (LHPP) exogenous function presented in Section 6.2 can be considered in a probabilistic approach. Instead of assigning a point process to dictate the interval-censored exogenous events, we consider a probability distribtion $p(\mathfrak{s})$ corresponding to the probability an exogenous event occurs from $(0, T]$. Furthermore, we have the corresponding immigrant event times which are sampled from the probability distributions $\mathfrak{s} \in \Psi_T$.

We first consider the impulse response MBPP with the position of the impulse as a parameter,

$$\xi_\delta(t; \mathfrak{s}) \doteq \delta(t - \mathfrak{s}) + \int_0^t \phi(s) \cdot \xi_\delta(t - s; \mathfrak{s}) \mathrm{d}s. \tag{56}$$

Then, we can consider the MBPP corresponding to the expectation of the intensity function with respect to the position of the impulse:

$$\mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} [\xi_\delta(t; \mathfrak{s})] = \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} \left[\delta(t - \mathfrak{s}) + \int_0^t \phi(s) \cdot \xi_\delta(t - s; \mathfrak{s}) \mathrm{d}s \right]$$

$$= \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} [\delta(t - \mathfrak{s})] + \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} \left[\int_0^t \phi(s) \cdot \xi_\delta(t - s; \mathfrak{s}) \mathrm{d}s \right]$$

$$= \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} [\delta(t - \mathfrak{s})] + \int_0^t \phi(s) \cdot \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} [\xi_\delta(t - s; \mathfrak{s})] \mathrm{d}s$$

$$= \int_0^T \delta(t - \mathfrak{s}) \cdot p(\mathfrak{s}) \mathrm{d}\mathfrak{s} + \int_0^t \phi(s) \cdot \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} [\xi_\delta(t - s; \mathfrak{s})] \mathrm{d}s$$

$$= p(t) + \int_0^t \phi(s) \cdot \mathbb{E}_{\mathfrak{s} \sim p(\mathfrak{s})} [\xi_\delta(t - s; \mathfrak{s})] \mathrm{d}s.$$

We now consider probability distributions $p_i(\mathfrak{s})$ defined as a uniform distribution over the exogenous event observation period $(q_{i-1}, q_i]$, for $i \in \{q_1, \ldots, q_m\}$:

$$p_i(\mathfrak{s}) = \begin{cases} \frac{1}{q_i - q_{i-1}} & \mathfrak{s} \in (q_{i-1}, q_i] \\ 0 & \text{otherwise} \end{cases}$$

76

$$= \frac{1}{q_i - q_{i-1}} \cdot [\![q_{i-1} < \mathfrak{s} \le q_i]\!].$$

The probabilistic interpretation of the LHPP exogenous function can now be expressed as the weighted sum of MBPP impulse expectations, where the weights are given by exogenous event volumes $\mathsf{S}(q_{i-1}, q_i]$ and the summation is over the probability distribution $p_i(\mathfrak{s})$. That is, the intensity function is given by:

$$
\begin{aligned}
\xi_\blacksquare(t) &\doteq \sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot \mathbb{E}_{\mathfrak{s} \sim p_i(\mathfrak{s})} \left[\xi_\delta(t; \mathfrak{s}) \right] \\
&= \sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot \left[p_i(t) + \int_0^t \phi(s) \cdot \mathbb{E}_{\mathfrak{s} \sim p_i(\mathfrak{s})} \left[\xi_\delta(t - s; \mathfrak{s}) \right] \mathrm{d}s \right] \\
&= \sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot p_i(t) + \sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot \left[\int_0^t \phi(s) \cdot \mathbb{E}_{\mathfrak{s} \sim p_i(\mathfrak{s})} \left[\xi_\delta(t - s; \mathfrak{s}) \right] \mathrm{d}s \right] \\
&= \sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot p_i(t) + \int_0^t \phi(s) \cdot \left[\sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot \mathbb{E}_{\mathfrak{s} \sim p_i(\mathfrak{s})} \left[\xi_\delta(t - s; \mathfrak{s}) \right] \right] \mathrm{d}s \\
&= \sum_{i=1}^{m} \mathsf{S}(q_{i-1}, q_i] \cdot p_i(t) + \int_0^t \phi(s) \cdot \xi_\blacksquare(t - s) \mathrm{d}s \\
&= \sum_{i=1}^{m} \frac{\mathsf{S}(q_{i-1}, q_i]}{q_i - q_{i-1}} \cdot [\![q_{i-1} < t \le q_i]\!] + \int_0^t \phi(s) \cdot \xi_\blacksquare(t - s) \mathrm{d}s,
\end{aligned}
$$

where the exogenous term of the intensity function is the LHPP exogenous function $s_\blacksquare(t)$, Eq. (43).

D. Convexity analysis

This section analyzes the loss functions of the Hawkes process and MBPP using the standard point process log-likelihood. We find that with the exponential triggering kernel and multi-impulse response in Equation (39), both the Hawkes process and MBPP are convex in κ for fixed θ.

To account for exogenous functions, the exogenous events Ψ_T and the endogenous events Υ_T are considered to be separable.

D.1 Hawkes process

It is trivial to verify that the negative log-likelihood functions corresponding to the exogenous functions below are convex in κ, as the terms consist of a negative-log function and linear functions which are convex functions of κ, but not θ.

- Multiple impulse response: Equation (39)

$$
\mathcal{L}(\theta; T) \doteq - \sum_{t_n \in \Omega_T} \log \left(\sum_{\mathfrak{s}_j \in \Psi_T} \delta(t_n - \mathfrak{s}_j) + \sum_{t_i < t_n} \kappa \theta e^{-\theta(t_n - t_i)} \right)
$$

$$+ |\Psi_T| + \kappa \sum_{t_i \in \Omega_T} (1 - e^{-\theta(T - t_i)}). \tag{57}$$

- Step response: $s(t) = [\![t \geq 0]\!]$

$$\mathcal{L}(\theta; T) \doteq - \sum_{t_n \in \Omega_T} \log \left([\![t_n \geq 0]\!] + \sum_{t_i < t_n} \kappa \theta e^{-\theta(t_n - t_i)} \right)$$
$$+ T + \kappa \sum_{t_i \in \Omega_T} (1 - e^{-\theta(T - t_i)}). \tag{58}$$

- Dassios response: Equation (12)

$$\mathcal{L}(\theta; T) \doteq - \sum_{t_n \in \Omega_T} \log \left(\kappa \theta + (u_0 - \kappa \theta) e^{-\theta t_n} + \sum_{t_i < t_n} \kappa \theta e^{-\theta(t_n - t_i)} \right)$$
$$+ \kappa \theta T + \frac{u_0 - \kappa \theta}{\theta} (1 - e^{-\theta T}) + \kappa \sum_{t_i \in \Omega_T} (1 - e^{-\theta(T - t_i)}) \tag{59}$$

D.2 Mean Behaviour Poisson

The negative log likelihood for the multiple impulse response is given by

$$\mathcal{L}(\theta; T) \doteq - \sum_{t_n \in \Omega_T} \log \left(\sum_{\mathfrak{s}_j \in \Psi_T} \delta(t_n - \mathfrak{s}_j) + \kappa \theta e^{(\kappa - 1)\theta(t_n - \mathfrak{s}_j)} \cdot [\![t_n > \mathfrak{s}_j]\!] \right)$$
$$+ |\Psi_T| + \sum_{\mathfrak{s}_j \in \Psi_T} \left[\frac{\kappa}{\kappa - 1} \left(e^{(\kappa - 1)\theta(T - \mathfrak{s})} - 1 \right) \right] \tag{60}$$

We will prove that for fixed θ, $\mathcal{L}(\theta, T)$ is convex with respect to κ over $\kappa \in [0, 1)$.

Our first step is to prove convexity in κ of the NLL for the single impulse response. Suppose that $s(t) = \delta(t - \mathfrak{s})$ for $\mathfrak{s} \geq 0$. The corresponding NLL is given by

$$\mathcal{L}(\theta; T) \doteq \overbrace{- \sum_{t_n \in \Omega_T} \log \left(\delta(t_n - \mathfrak{s}) + \kappa \theta e^{(\kappa - 1)\theta(t_n - \mathfrak{s})} \cdot [\![t_n > \mathfrak{s}]\!] \right)}^{l(\xi(\cdot; \kappa))}$$
$$+ \underbrace{1 + \frac{\kappa}{\kappa - 1} \left(e^{(\kappa - 1)\theta(T - \mathfrak{s})} - 1 \right)}_{\Xi(T; \kappa)}.$$

As we can see, the single impulse NLL can be written as a sum of two terms, a function of the conditional intensity $l(\xi(\cdot; \kappa))$, where $l : x \mapsto - \sum \log(x)$, and the compensator at the final time $\Xi(T; \kappa)$. To prove convexity of $\mathcal{L}(\theta, T)$, we show that both components are convex in κ.

Given that the function l is non-decreasing and convex (since a sum of convex functions is still convex), it suffices to show that

$$\xi(t_n; \kappa) = \delta(t_n - \mathfrak{s}) + \kappa\theta e^{(\kappa-1)\theta(t_n-\mathfrak{s})} \cdot [\![t_n > \mathfrak{s}]\!]$$

is convex in κ to prove that $l(\xi(\cdot; \kappa))$ is convex in κ. This follows from the property that if f and g are convex and f is non-decreasing, then $f \circ g$ is convex.

It can be shown that

$$\frac{\partial^2 \xi(t_n; \kappa)}{\partial \kappa^2} = \theta^2(t_n - \mathfrak{s})e^{\theta(t_n-\mathfrak{s})(\kappa-1)}(\kappa\theta(t_n - \mathfrak{s}) + 2) \cdot [\![t_n > \mathfrak{s}]\!] \geq 0,$$

which proves convexity of $\xi(t_n; \kappa)$ for κ. Thus, $l(\xi(\cdot; \kappa))$ is a convex function.

We now show convexity of $\Xi(T; \kappa)$ in κ. It can be shown that

$$\frac{\partial^2 \Xi(T; \kappa)}{\partial \kappa^2} = \frac{2 - e^{-\theta(T-\mathfrak{s})(1-\kappa)}(2 + 2\theta(T - \mathfrak{s})(1 - \kappa) + (\theta(T - \mathfrak{s}))^2\kappa(1 - \kappa)^2)}{(1 - \kappa)^3}$$

So that $\frac{\partial^2 \Xi(T;\kappa)}{\partial \kappa^2} \geq 0$, we must show that

$$p(k) = e^{-\theta(T-\mathfrak{s})(1-\kappa)}(2 + 2\theta(T - \mathfrak{s})(1 - \kappa) + (\theta(T - \mathfrak{s}))^2\kappa(1 - \kappa)^2)(1 - \kappa)^3 \leq 2$$

whenever $0 \leq \kappa < 1$. Since

$$\frac{\mathrm{d}p}{\mathrm{d}k} = \theta^2(T - \mathfrak{s})^2(1 - \kappa)^2 e^{-\theta(T-\mathfrak{s})(1-(\theta(T-\mathfrak{s})))}(\theta(T - \mathfrak{s})\kappa + 3) > 0,$$

for $0 \leq \kappa < 1$, which means that $p(k)$ is non-decreasing over this interval, and $p(1) = 2$, we have

$$\frac{\partial^2 \Xi(T; \kappa)}{\partial \kappa^2} \geq 0,$$

for $0 \leq \kappa < 1$. Thus, $\Xi(T; \kappa)$ is convex in κ over this range.

Since $l(\xi(\cdot; \kappa))$ and $\Xi(T; \kappa)$ are both convex, their sum $\mathcal{L}(\theta, T)$, the NLL for the single impulse response, is convex.

Convexity of the NLL for the multi impulse response follows from the following observation: Observe that the conditional intensity ξ and the compensator Ξ for the multi impulse response can be expressed, respectively, as sums of ξ and Ξ of the single impulse response. Given that a nonnegative-weighted sum of convex functions is still convex, it follows that the NLL for the multi impulse response is convex as well.

For θ, observe that

$$\theta \to -\log\left(\delta(t_n - \mathfrak{s}) + \kappa\theta e^{(\kappa-1)\theta(t_n-\mathfrak{s})} \cdot [\![t_n > \mathfrak{s}]\!]\right)$$

is convex in θ but

$$\frac{\partial^2}{\partial \theta^2}\left(\frac{\kappa}{\kappa - 1}e^{(\kappa-1)\theta(T-\mathfrak{s})}\right) = \kappa(\kappa - 1)(T - \mathfrak{s})^2 e^{(\kappa-1)\theta(T-\mathfrak{s})} < 0$$

for $0 < \kappa < 1$. Thus $\frac{\partial^2 \mathcal{L}}{\partial \theta^2} \geq 0$ does not hold, implying that the NLL is not convex in θ for fixed κ.

E. Experiments

E.1 Number of Approximation Points

Fig. 7 shows the RSS of the compensator values of the discrete compensator function versus the closed-form compensator function.
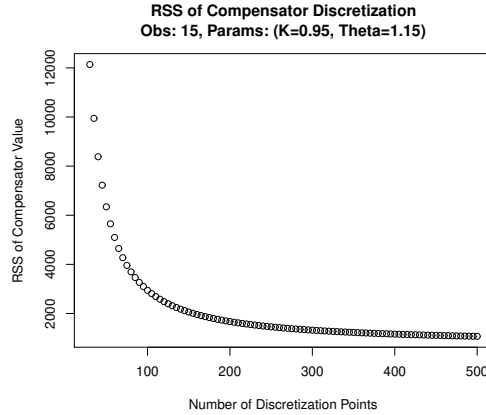

Figure 7: Compensator approximation error with respect to the number of discretization points.

E.2 Other Non-endogenous Experiments

Fig. 8 shows the parameter fittings resulting from a different number of discretization points.

Table 12: Non-endogenous synthetic experiments for Scenario F.

True Params.	Loss Func.	Model		Number of Observation Intervals					
				5	10	15	30	60	100
$\kappa = 0.6$ $\theta = 0.8$	SSE	Closed	κ	0.726 ± 0.004	0.682 ± 0.005	0.668 ± 0.005	0.642 ± 0.005	0.62 ± 0.005	0.612 ± 0.005
			θ	0.647 ± 0.021	0.641 ± 0.024	0.801 ± 0.031	0.995 ± 0.052	0.915 ± 0.052	0.865 ± 0.045
		Approx	κ	10 ± 0	0.621 ± 0.006	0.605 ± 0.006	0.601 ± 0.006	0.601 ± 0.006	0.601 ± 0.006
			θ	0.005 ± 0	8.823 ± 0.866	9.546 ± 1.589	1.09 ± 0.091	0.909 ± 0.052	0.856 ± 0.043
	IC-LL	Closed	κ	0.728 ± 0.004	0.69 ± 0.005	0.673 ± 0.005	0.644 ± 0.005	0.622 ± 0.005	0.613 ± 0.005
			θ	0.644 ± 0.019	0.67 ± 0.023	0.828 ± 0.031	1.014 ± 0.054	0.931 ± 0.05	0.876 ± 0.045
		Approx	κ	10 ± 0	0.618 ± 0.007	0.599 ± 0.006	0.594 ± 0.006	0.595 ± 0.006	0.596 ± 0.006
			θ	0.005 ± 0	8.355 ± 1.02	9.607 ± 1.603	1.091 ± 0.088	0.905 ± 0.051	0.85 ± 0.043
$\kappa = 0.95$ $\theta = 1.15$	SSE	Closed	κ	0.958 ± 0.002	1.496 ± 2.17	1.494 ± 2.171	0.95 ± 0.004	0.95 ± 0.004	0.95 ± 0.004
			θ	1.765 ± 0.098	1.362 ± 0.358	1.301 ± 0.344	1.283 ± 0.089	1.216 ± 0.08	1.19 ± 0.076
		Approx	κ	10 ± 0	1.139 ± 0.007	1.033 ± 0.005	0.958 ± 0.003	0.95 ± 0.004	0.95 ± 0.004
			θ	0.014 ± 0	9.409 ± 0.764	10 ± 0	10 ± 0	1.765 ± 0.177	1.433 ± 0.111
	IC-LL	Closed	κ	0.959 ± 0.002	0.953 ± 0.003	0.95 ± 0.003	0.949 ± 0.004	0.95 ± 0.004	0.95 ± 0.004
			θ	1.703 ± 0.086	1.52 ± 0.091	1.451 ± 0.096	1.314 ± 0.083	1.233 ± 0.075	1.202 ± 0.071
		Approx	κ	10 ± 0	1.161 ± 0.007	1.041 ± 0.005	0.958 ± 0.003	0.95 ± 0.004	0.95 ± 0.004
			θ	0.017 ± 0	9.767 ± 0.403	10 ± 0	10 ± 0	1.776 ± 0.163	1.433 ± 0.102

E.3 Examine the MBPP fitting bias

Setup and protocol. In this section, we examine the parameter fitting bias for the MBPP process (shown in Figs. 9 and 10). We achieve this by sampling the Hawkes process
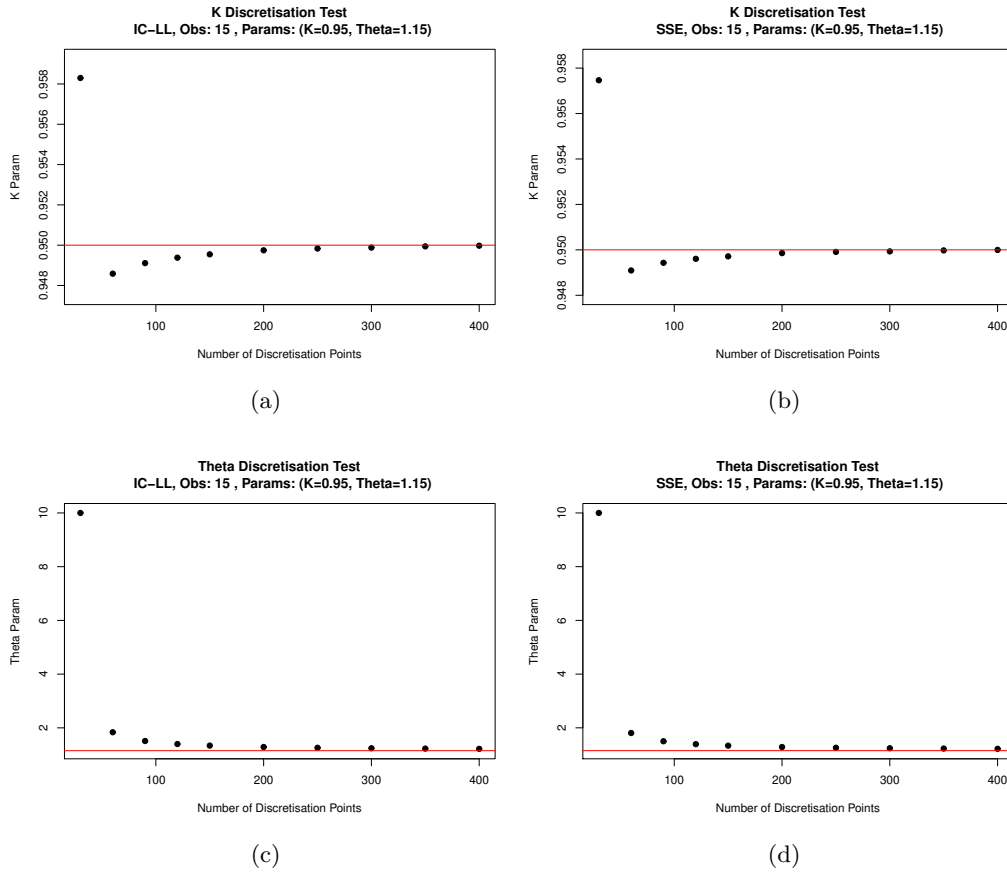
80

Figure 8: Plots detailing the parameter fits of synthetic data in setting D with varying numbers of discretization points. The plots are taken over 15 observation intervals, using $\kappa = 0.95$ and $\theta = 1.15$ parameters. The red line indicates these true parameters of the synthetic data.

81

over a wide range of parameters, fitting the MBPP process for each parameter combination in Scenario F—as per Table 2—, and examining the difference between the true parameter (used for sampling) and the fitted parameter. We sample synthetic Hawkes realizations using `HP-sin` (see Eq. (53)) with a parameter grid. We vary κ between 0.1 and 0.95 (step 0.05) and θ between 0.5 to 1.55 (step 0.05). For each parameter combination, we jointly fit 50 groups of realizations, and we obtain 50 MBPP parameter estimates (see the main-text experiments section). We use six values for the number of observation intervals, controlling the width of the observation interval: 5 intervals (largest width, Fig. 9a), 10 intervals (Fig. 9b), 15 intervals (Fig. 9c), 30 intervals (Fig. 10a), 60 intervals (Fig. 10b), and 100 intervals (smallest width, Fig. 10c). We report the mean difference between fitted and true values for κ (left panel of each figure) and θ (right panel). Furthermore we highlight the specific parameters used in the main-text by the green shapes, with the triangle signifying $\kappa = 0.6$ and $\theta = 0.8$; and diamond signifying $\kappa = 0.95$ and $\theta = 1.15$.

Results and interpretation There are several conclusions that emerge from Figs. 9 and 10. First, we observe that we retrieve parameter κ with only minor errors over every parameter configuration. The occasional more consistent over-estimation is likely due to convergence failures of the optimization algorithm. Second, we observe that fitting the θ parameter is significantly more challenging. For high values of θ the fitting under-estimates the true θ for low values of κ (bottom-right corner) and over-estimates it for high values of κ (top-right corner). Note that high θ corresponds to a faster decay rate for the Hawkes kernel $\phi(t)$, which typically translates to offspring events being tightly clustered close to their parent event. As θ decreases (slower kernel decay), the parameter is increasingly well recovered. Third observation, there seems to be a breakpoint between regions where parameters are underestimated (blue) and overestimated (red) for the θ plots. Forth and as expected, for higher data granularity — i.e., the higher the number of intervals — the θ are recovered closer to their true values. Fifth and finally, for high data granularity (100 intervals in Fig. 10c) we observe a slight over-estimation for both κ and θ. We posit that this consistent over-estimation is due to the model mismatch between the generating model (Hawkes process) and the fitting model (MBPP).
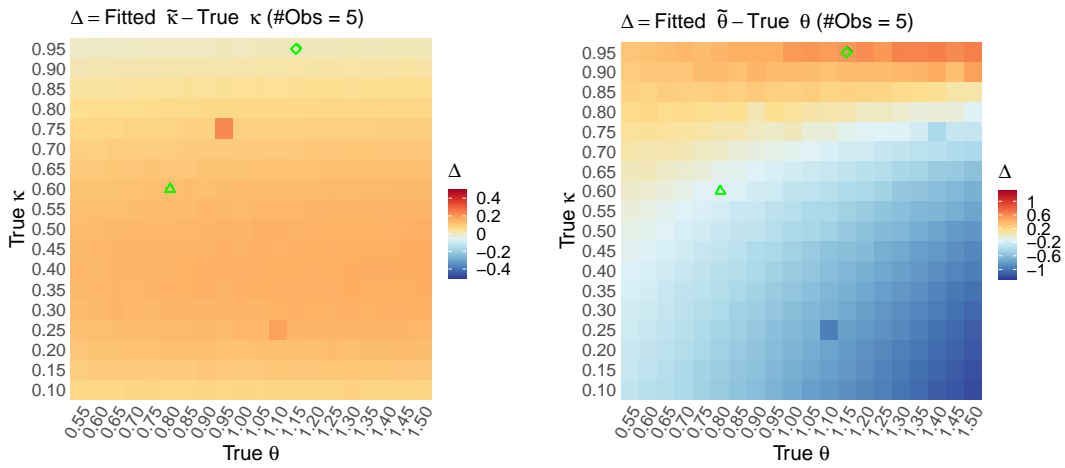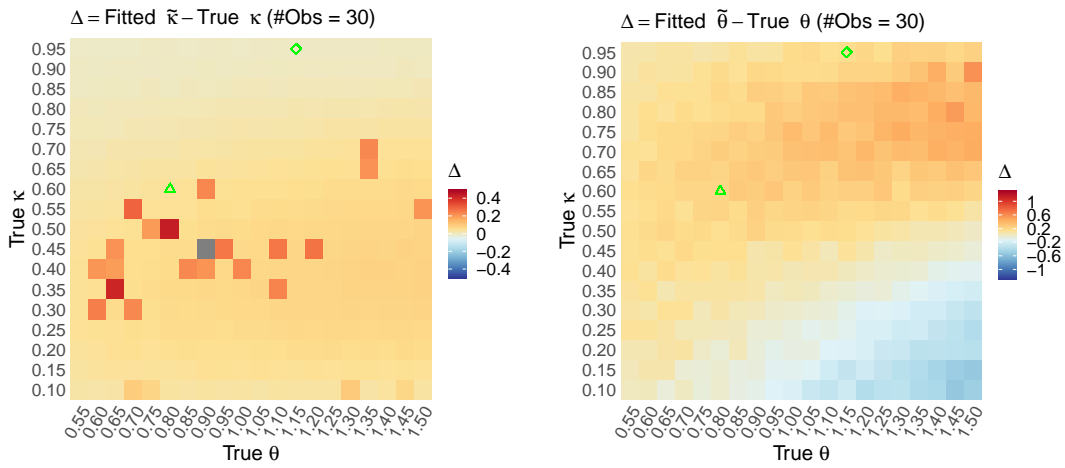
(a) 5 observational intervals
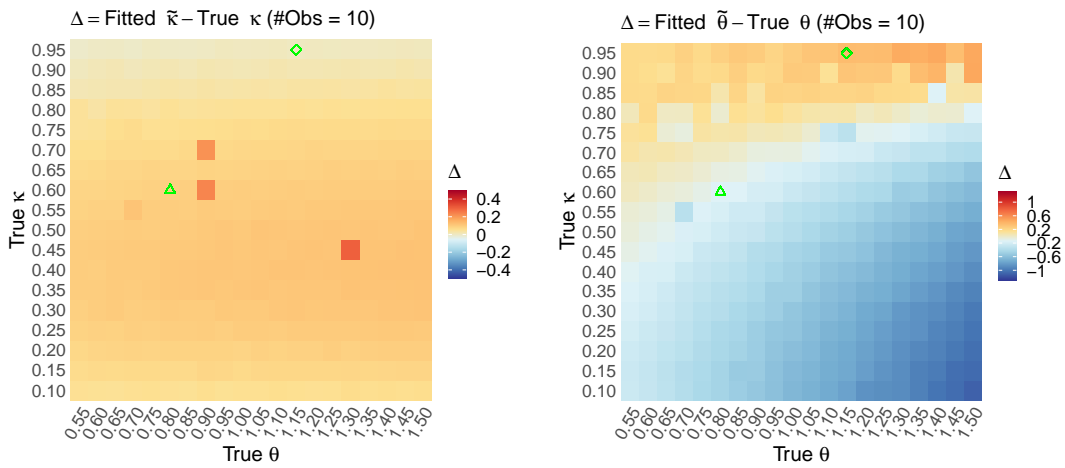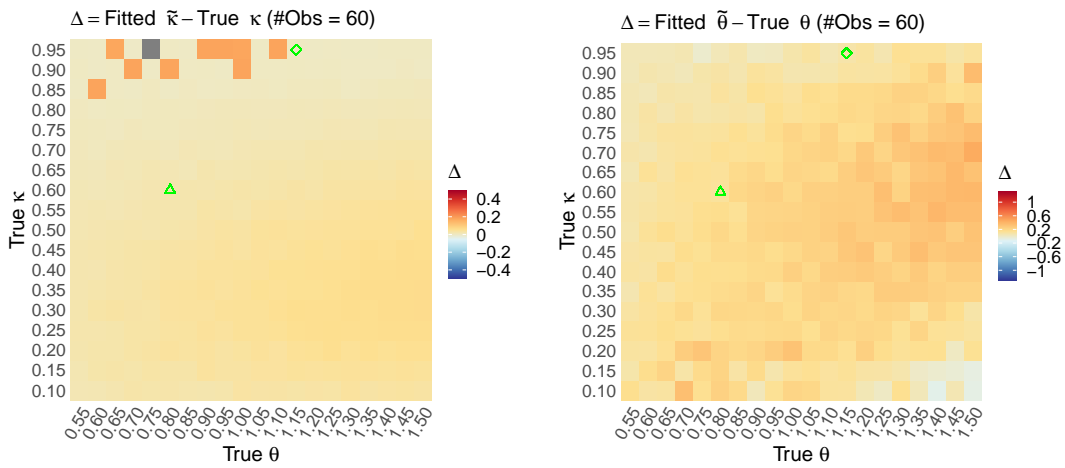

(b) 10 observational intervals
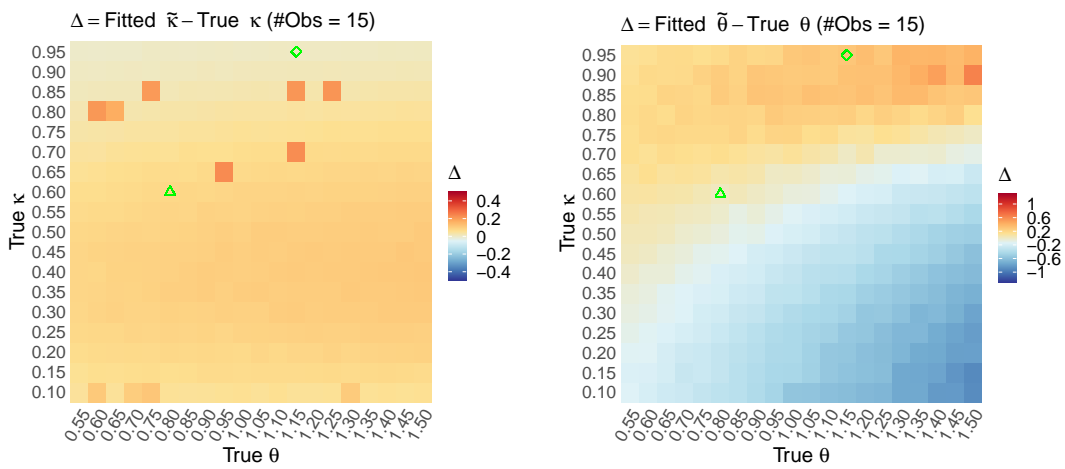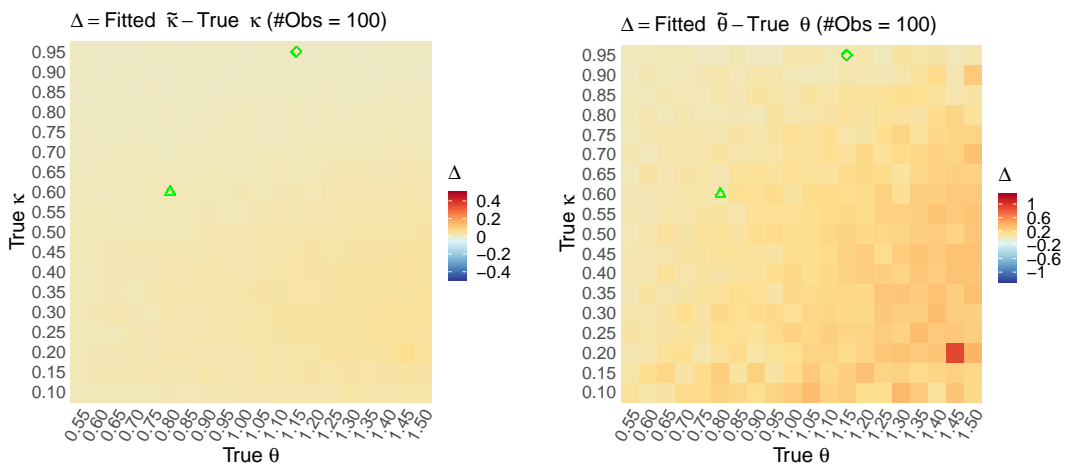

(c) 15 observational intervals

Figure 9: MBPP parameter error heatmap for κ **(left)** and θ **(right)**.

(a) 30 observational intervals

(b) 60 observational intervals

(c) 100 observational intervals

Figure 10: MBPP parameter error heatmap for κ **(left)** and θ **(right)**.