DELAYED HAWKES BIRTH-DEATH PROCESSES

JUSTIN BAARS, ROGER J. A. LAEVEN, AND MICHEL MANDJES

ABSTRACT. We introduce, and formally establish, a variant of the Hawkes-fed birth-death process — the *delayed Hawkes birth-death process* — in which the conditional intensity does not increase at arrivals but at departures from the system. In a scaling limit where sojourn times are stretched out by a factor $\sqrt{T}$, after which time gets contracted by a factor T, the delayed Hawkes process behaves markedly differently from its classical counterpart. We design a family of models admitting a cluster representation and containing the Hawkes and delayed Hawkes processes as special cases. The cluster representation allows for transform characterizations by a fixed-point equation and for analysis of heavy-tailed asymptotics. We compare the delayed Hawkes process to the classical Hawkes process using stochastic ordering, which enables us to describe stationary distributions and heavy-traffic behavior. In the Markovian network case, a recursive procedure is presented to calculate the dth-order moments analytically.

KEYWORDS. Self-exciting processes ∘ Hawkes processes ∘ Birth-death processes ∘ Scaling limits ∘ Branching processes ∘ Transform analysis ∘ Stochastic ordering.

MSC 2020 CLASSIFICATIONS. Primary: 60G55; Secondary: 60E10, 60E15, 62E20.

AFFILIATIONS. JB and RL are with the Dept. of Quantitative Economics, University of Amsterdam. RL is also with EURANDOM, Eindhoven University of Technology, and with CENTER, Tilburg University. MM is with the Mathematical Institute, Leiden University, and is also affiliated with the Korteweg-de Vries Institute for Mathematics, University of Amsterdam; EURANDOM, Eindhoven University of Technology, Eindhoven; Amsterdam Business School, University of Amsterdam. The research of JB and RL was funded in part by the Netherlands Organization for Scientific Research under an NWO VICI grant (2020–2027). The research of MM was funded in part by the NWO Gravitation project NETWORKS, grant number 024.002.003.

EMAIL ADDRESSES. j.r.baars@uva.nl, r.j.a.laeven@uva.nl, and m.r.h.mandjes@math.leidenuniv.nl.

Date: October 21, 2026.

1. INTRODUCTION

Since their introduction in 1971 [22, 23], Hawkes processes have gained significant attention in the academic literature. One notable application is in finance [1, 2, 6, 7], where they have been used to capture the clustering behavior of financial returns and transactions, such as stock trades or order arrivals in electronic markets. Hawkes processes have also been applied in social network analysis [35], to represent the contagious nature of information diffusion or the spread of online content in social media platforms. Additionally, in the field of seismology [26, 39, 47] they are used to model earthquake aftershock sequences. Other applications include the analysis of disease outbreaks [12], the prediction of online user activity [45], crime modeling [42], and the assessment of neuronal spike trains [44]. The versatility of Hawkes processes makes them a valuable tool in various domains, providing insights into the underlying mechanisms driving the observed events.

The Hawkes process is a self-exciting càdlàg point process $(N(t))_{t\geqslant 0}$, which can be defined through its conditional intensity process $(\Lambda(t))_{t\geqslant 0}$ [15]. In the simplest linear, unmarked, univariate case, the (left-continuous, predictable) conditional intensity process is given by

$$\Lambda(t) = \lambda_0 + \sum_{t_i<t} h(t-t_i) = \lambda_0 + \int_{(-\infty,t)} h(t-s)\,\mathrm{d}N(s), \tag{1}$$

where $\lambda_0 > 0$ is the *baseline intensity*, or *immigration intensity*, $(t_i)_{i\in\mathbb{N}}$ is an increasing sequence of *arrival times*, $h : [0,\infty) \to [0,\infty)$ is the *excitation kernel*, which is assumed to be integrable, and $N((-\infty,0))$ is some initial condition, typically either a random initial condition resulting in a stationary version of the process, or an empty history.

Besides being a suitable process to model real-world phenomena, the Hawkes process owes much of its popularity to its high tractability. In particular, recursive procedures have been developed to determine corresponding moments [14, 18, 20, 34]; a procedure has been devised by which, in the context of Hawkes-fed population processes, transforms can be approximated by iterates of a certain operator [30]; heavy-tailed and heavy-traffic asymptotics have been identified [30, 34]; techniques for nonparametric estimation of the model parameters, with provable performance guarantees, have been set up [32]; a broad range of scaling and large deviation limits have been studied [5, 25, 28, 29, 31, 48]; existence, uniqueness and stability results have been established that apply under great generality [10, 36, 43]; and recently results on the distribution of the Hawkes process' underlying cluster duration have become available [16]. Evidently, this list is by no means exhaustive, but it provides an illustration of the process' amenability for analysis, focusing on contributions of direct relevance to this paper.

Since its inception, various generalized versions of the basic variant of the Hawkes process have been examined, all of them being point or population processes in which the occurrence of events affect the conditional intensity process. For example, Massoulié [36] considers a highly flexible family of models involving a (possibly) nonlinear intensity function λ:

$$\Lambda(t) = \lambda\left(\int_{(-\infty,t)} h(t-s,B_s)\,\mathrm{d}N(s)\right), \tag{2}$$

which allows for the dependency on space-dependent *random marks*, $(B_{s_i})_{i\in\mathbb{N}}$, taking values in some general measurable space. Hawkes-driven birth-death population processes have been studied in [19, 30, 34]. Another variant is the *ephemerally self-exciting point process*, as introduced in [17], in which the excitation caused by the i-th arrival vanishes after some stochastic time J_i. By considering this system as a birth-death process with lifetimes $(J_i)_{i\in\mathbb{N}}$, one could say that 'a particle excites as long as it is in the system'. Another variant of the classical Hawkes process is analyzed in [43], in which the excitation is dependent on the time since the last arrival, a phenomenon termed *age-dependency*. A process that describes behavior opposite to the Hawkes process, is the *self-correcting process* [27, 40, 41], in which any arrival *decreases* the conditional intensity, making more arrivals in the near future *less* likely. Recently, Hawkes processes allowing for both self-excitation and self-inhibition were studied; see [11] and the references therein.

A general observation, based on the cases dealt with in the literature, is that tractability tends to be preserved for models that admit a cluster process representation. This is the case for multivariate linear marked Hawkes point and birth-death processes (covering specific Hawkes-fed population

processes), and for the ephemerally self-exciting process. For these classes of processes one has succeeded in establishing analogs to results known for the classical Hawkes process. On the other hand, for processes having nonlinear intensity functions, age-dependent processes, Hawkes-fed single-server queues and self-correcting processes, there is no cluster representation, making such models considerably harder to analyze than the classical Hawkes process.

In this paper, we introduce a variant of the Hawkes process new to the literature, to the best of our knowledge. This variant is motivated as follows. Consider first a standard Hawkes-fed birth-death population process, or 'infinite-server queue with Hawkes input', denoted by $(Q(t))_{t\geqslant 0}$; see e.g., [19, 34]. Then, particles arrive at rate $\Lambda(t)$, and at the i-th *arrival* at time t_i, the conditional intensity process $\Lambda(t)$ jumps upwards by $B_i h(t-t_i)$, where $(B_i)_{i\in\mathbb{N}}$ are i.i.d. marks. The particle stays in the system for a duration J_i, where $(J_i)_{i\in\mathbb{N}}$ are i.i.d. lifetimes, or 'service times' in queueing terminology; after departure, the excitation effect is still present. By contrast, we define a process in which the conditional intensity does not jump at arrivals, but at *departures* from the system. More specifically, the intensity process $\Lambda(t)$ does not change at an arrival, but jumps upwards by $B_i h(t-t_i')$ at the i-th departure at time t_i'. In this situation, an arrival still increases the conditional intensity $\Lambda(t)$, but only after a *delay* equal to its lifetime (or service time, in queueing terms). For this reason, one may call the corresponding process $(Q(t))_{t\geqslant 0}$ a *delayed Hawkes birth-death process*, or a *delayed Hawkes infinite-server queue*; or, more briefly, a DH/G/∞ queue, using Kendall's notation. We refer to the counting process $(N(t))_{t\geqslant 0}$ as a *delayed Hawkes process* or briefly as *delayed Hawkes*. (The terms 'infinite-server queue' and 'birth-death population process' can be used interchangeably. In fact, one could argue that an infinite-server queue is not really a queue, since customers are always served directly, do not observe each other, and never wait. Using population processes terminology, one could refer to delayed Hawkes infinite-server queues as *birth-death processes exhibiting posthumous excitation*.)

A typical realization of the delayed Hawkes birth-death process can be found in Figure 1. Intuitively, one would expect this process to share some common features with the classical Hawkes process, but with a 'lower level of clustering' of events: one has to wait some time (distributed as the random variable J) for the excitation to start, so that arrivals induced by excitation are further away from the initial arrival than under the classical Hawkes process.

By setting the lifetimes J equal to zero and by keeping track of $N(\cdot)$, we recover the classical Hawkes process, entailing that the delayed Hawkes process constitutes a generalization of the classical Hawkes process. In the following examples, the delayed Hawkes birth-death process may provide a realistic and appealing probabilistic model.

- Word-of-mouth referrals: in a queueing context, customers who are satisfied about the service may excite other potential customers. Therefore, the arrival process may behave like a self-exciting process; however, a customer typically does not start exciting others during the service, but only starts doing so upon leaving/finishing the system/service.
- In epidemiology, the spread of infectious diseases often exhibits *delayed* self-exciting behavior. Indeed, when an individual becomes infected, there is typically an incubation period before the individual starts showing symptoms or becomes contagious. As more individuals become infected, start exhibiting symptoms and become contagious, the transmission rate increases, resulting in a (delayed) increase in the number of new cases.

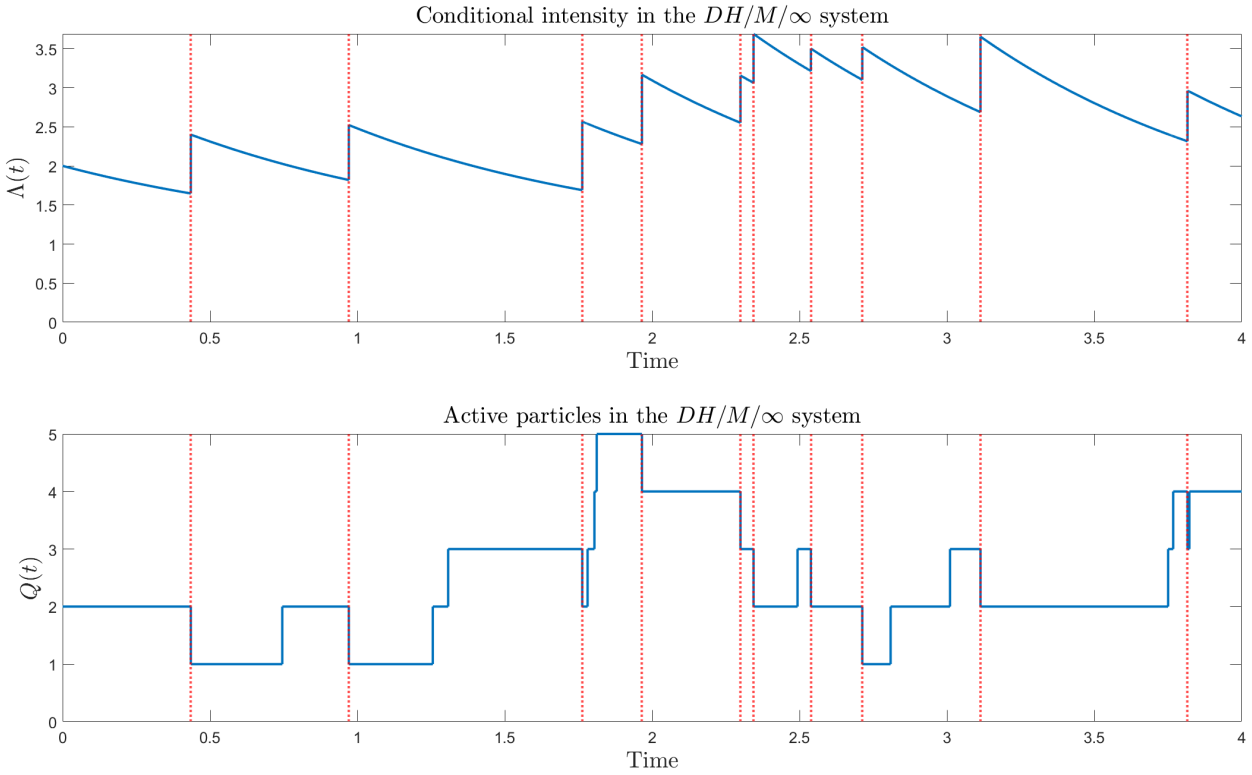

FIGURE 1. A realization of the Markovian DH/M/∞ queue, with $\lambda_0 = 1$, $h(t) = e^{-t}$, $B \sim \text{Beta}(3.5, 1.5)$ and $J \sim \text{Exp}(1)$. We start at $Q(0) = 2 = \Lambda(0)$. The vertical dotted lines correspond to departures, causing intensity increases.

- A financial order typically triggers more orders, but it may take time before an order is executed and therefore before it starts exciting. Even when the execution time is (very) small, as in liquid electronic markets, this delay changes the dynamics. Similar patterns arise in neuroscience.
- On social media platforms, the spread of content can exhibit self-exciting behavior with delay. When a popular post or topic emerges, it can trigger a cascade of user interactions. As it takes time for users to engage with the content and for the effects to ripple through their social networks, the propagation of these interactions can display a delayed response.

If we would like to model real-world phenomena, such as those described in the examples above, using the delayed Hawkes process, we need the process to be tractable, in order to understand its probabilistic structure. As it turns out, this novel process is remarkably tractable: its linear version admits a cluster process representation, and many results that are known for the classical Hawkes process have suitably modified counterparts for the delayed Hawkes process.

This work contributes to the literature in several ways. First, we introduce the delayed Hawkes process. In fact, more generally, we introduce a family of multivariate sojourn-time dependent point processes, containing the classical Hawkes, delayed Hawkes, and the ephemerally self-exciting point process [17] as special cases. This general family of models is formulated via a stochastic differential equation for the conditional intensity process, which we exploit to prove existence, uniqueness and stability results, leveraging methodology from [10, 36].

Second, we contribute to a rich literature on scaling limits for Hawkes processes, see e.g., [5, 25, 28, 29], by deriving a scaling limit that exhibits the effect of the delay for the delayed Hawkes process. Specifically, we show that our family of models obeys the same functional central limit theorem as the classical Hawkes process; however, in a scaling regime in which sojourn times are stretched out by a factor $\sqrt{T}$, after which time gets contracted by a factor T, and T is sent to ∞, the delayed Hawkes process behaves markedly differently from its classical counterpart.

Third, for the linear version of our family of models, we provide a cluster process representation, allowing us to derive fixed-point equations that enable transform characterizations. In addition, we employ these fixed-point equations to establish heavy-tailed asymptotics. We use the cluster representation of the Hawkes and delayed Hawkes processes to prove stochastic dominance results, which are typically proved by comparing sample paths. In essence, we couple sample paths only within generations, obtaining a complex genealogical coupling for both processes. From a methodological standpoint, the ideas underlying this approach have the potential to be fruitful in other contexts as well.

Finally, we generalize results of [34] for calculating moments of the Hawkes process in the univariate, Markovian setting, to a higher-dimensional, delayed Hawkes setting, also allowing for network effects. Interestingly, this analysis now involves a Clement-Kac-Sylvester matrix.

The remainder of this article is structured as follows. In Section 2, we introduce a general family of multivariate point process models encompassing classical Hawkes, delayed Hawkes and ephemeral Hawkes as special cases. In Section 3, existence, uniqueness and stability results are established for this general family of models. Section 4 studies scaling limits; in particular, we derive a scaling limit for delayed Hawkes highlighting the effect of the delay. In Section 5, we use cluster-representation based methods to describe fixed points in the transform domain; and exploit those fixed-point equations to derive heavy-tailed asymptotics. In Section 6, we compare Hawkes to delayed Hawkes systems using stochastic ordering. In Section 7, we study Markovian models, for which we describe recursive methods to calculate moments analytically. We provide a discussion and concluding remarks in Section 8. Various (lengthy) proofs and some additional results are relegated to the Appendix. In online Supplementary Material [4], we provide full details of the proofs from Section 4, and of the proof of Theorem 4 from Section 5.

2. Model definitions

In this section, we introduce and provide definitions for a family of multivariate point process models having sojourn-time dependent excitation, using both their conditional intensity processes and, in the linear case, their cluster process representation. Furthermore, we define a network of delayed Hawkes birth-death processes through conditional intensities.

The definitions in this section are ordered from most to least general. Definition 1 provides the most general framework, allowing for both nonlinear dynamics and non-Markovian, sojourn-time dependent excitation, and is considered in the existence results of Section 3. Definition 2 imposes linearity while still permitting a non-Markovian structure with sojourn-time dependent excitation; this linearity is used throughout Sections 4–6. Finally, Definition 3 restricts attention to delayed Hawkes networks that are both linear and Markovian, which are treated in Section 7.

We start by defining a family of models exhibiting sojourn-time dependent excitation, encompassing the classical Hawkes, the delayed Hawkes, and the ephemerally self-exciting [17] process. We first describe this family of models through a conditional intensity representation, allowing for nonlinearity. We then restrict attention to the linear case for which we also provide a cluster representation-based definition. The two definitions are equivalent for processes starting on an empty history, whenever the cluster representation exists (i.e., in the linear case). The conditional intensity-based definition allows for nonlinear effects, but we only use this in Section 3 when proving existence, uniqueness and stability results; in the rest of this article we focus on the linear case.

We denote the d-dimensional joint point (or counting), birth-death and conditional intensity process of the intended model (defined below) by the triple $(\boldsymbol{N}(t), \boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))_{t\geqslant 0}$, with $\boldsymbol{Z}(t) = [Z_1(t) \cdots Z_d(t)]^\top$, for $\boldsymbol{Z} \in \{\boldsymbol{N}, \boldsymbol{Q}, \boldsymbol{\Lambda}\}$. We write $t_1 < t_2 < \cdots$ for the a.s. increasing sequence of jump (or event) times of $\boldsymbol{N}(\cdot)$, and we denote events by triples (t_r, j, J), where $J \sim J_j$ if the r-th event occurs in coordinate j. We assume that an arrival in coordinate j at time t_r induces a *random* jump in the intensity in the i-th coordinate of size $h_{ij,J,\omega}(\cdot - t_r)$; the randomness in $h_{ij,J}$ is modeled by the ω-dependence.

Definition 1 (Conditional intensity for d-dimensional point processes with sojourn-time dependent excitation)**.** *Let $d \in \mathbb{N}$ denote the dimension. For $j \in [d]$, let J_j be the positive sojourn time random variable of coordinate j. For each $i, j \in [d]$, let $\omega \mapsto h_{ij,J_j,\omega}$ be a random J_j-dependent piecewise continuous function with support contained in $[0, \infty)$, for almost all (J_j, ω). Furthermore, suppose that for each $i, j \in [d]$, $\mathbb{E}_{J_j}\mathbb{E}_{\omega|J_j}\|h_{ij,J_j,\omega}\|_{L^\infty} < \infty$. Assume that the realizations of the random functions are conditionally (on J) cross-sectionally and serially independent. Suppose that the lifetimes are drawn at the time of arrival. Let $\boldsymbol{H_J}$ be the (random, $\boldsymbol{J}$-dependent) matrix consisting of elements $(\boldsymbol{H_J})_{ij} = h_{ij,J}$, where the j-th column is dependent on the same realization of J_j. Define the d-dimensional càdlàg point process $\boldsymbol{N} = (N_i(t))_{i\in[d],t\in\mathbb{R}}$ with sojourn-time dependent excitation through*

$$\begin{aligned}
\mathbb{P}(N_i(t+\Delta t) - N_i(t) = 0 \,|\, \mathcal{H}_t) &= 1 - \Lambda_i(t)\Delta t + o(\Delta t),\\
\mathbb{P}(N_i(t+\Delta t) - N_i(t) = 1 \,|\, \mathcal{H}_t) &= \Lambda_i(t)\Delta t + o(\Delta t),\\
\mathbb{P}(N_i(t+\Delta t) - N_i(t) \geqslant 2 \,|\, \mathcal{H}_t) &= o(\Delta t),
\end{aligned}$$

as $\Delta t \downarrow 0$, where $(\mathcal{H}_t)_{t\in\mathbb{R}} = \sigma(\boldsymbol{N}(s), (\boldsymbol{H_J}(\cdot))(s), \boldsymbol{J}(s) : s \leqslant t)_{t\in\mathbb{R}}$ is the natural filtration generated by $\boldsymbol{N}$ along with random lifetimes $\boldsymbol{J}(s)$ and excitation kernels $(\boldsymbol{H_J}(\cdot))(s)$ corresponding to an arrival at time s. In the linear case, we set

$$\boldsymbol{\Lambda}(t) = \boldsymbol{\lambda}_0 + \int_{-\infty}^{t} \boldsymbol{H_J}(t-s)\, \mathrm{d}\boldsymbol{N}(s), \tag{3}$$

where $\boldsymbol{\lambda}_0 \geqslant 0$, with at least one of the base rates being strictly positive, and the integral in (3) is understood to exclude t. In the nonlinear case, we take measurable L_i-Lipschitz functions $\phi_i : \mathbb{R} \to \mathbb{R}_+$, for each $i \in [d]$, and define the conditional intensity of the i-th coordinate via

$$\Lambda_i(t) = \phi_i\left(\sum_{j=1}^{d} \int_{-\infty}^{t} h_{ij,J}(t-s)\, \mathrm{d}N_j(s)\right). \tag{4}$$

From the point process, including the realizations of sojourn times, the birth-death process $(\boldsymbol{Q}(t))_{t\geqslant 0}$ can easily be constructed. The conditional intensity process $\boldsymbol{\Lambda}(\cdot)$ is taken left-continuous and is predictable; cf. [15, Example 7.2(b) and Ch. 14]. We interchangeably start the point process on a history on $(-\infty, 0)$, which typically refers to the stationary version of the point process, or on an empty history, in which case an integral $\int_{-\infty}^{t} \cdot \, \mathrm{d}\boldsymbol{N}(s)$ reduces to $\int_{0}^{t} \cdot \, \mathrm{d}\boldsymbol{N}(s)$.

Definition 1 encompasses the multivariate marked *classical* Hawkes process, the multivariate marked hybrid *ephemerally* self-exciting process (cf. [17]), and the multivariate marked *delayed* Hawkes process, which are defined by setting

$$h_{ij,J,\omega}(\cdot) = \begin{cases} B_{ij,\omega} h_{ij}(\cdot), & \text{(classical)} \\ B_{ij,\omega} h_{ij}(\cdot) \mathbf{1}\{\cdot < J\}, & \text{(ephemeral)} \\ B_{ij,\omega} h_{ij}(\cdot - J) \mathbf{1}\{\cdot > J\}, & \text{(delayed)} \end{cases} \tag{5}$$

respectively.

We highlight the richness of the family of processes introduced in Definition 1. Notably, this family includes processes where the degree of self-excitation depends, positively or negatively, upon the lifetimes of particles. For instance, one can define

$$h_{ij,J,\omega}(\cdot) = B_{ij,J,\omega} h_{ij}(\cdot).$$

One might *a priori* expect that such general models would be intractable; however, e.g., Theorem 4 in Section 5 demonstrates that calculations for these models can, in fact, be carried out effectively.

In Section 3, existence, uniqueness and stability for the nonlinear family of sojourn-time dependent point processes with intensities (4) is established. For the rest of this article, we focus on the linear case (3). The linear case admits a cluster representation, as follows.

Definition 2 (Cluster representation for d-dimensional point processes with sojourn-time dependent excitation)**.** *Let $d \in \mathbb{N}$ denote the dimension. For $j \in [d]$, let J_j be the positive sojourn time random variable of coordinate j. For each $i, j \in [d]$, let $h_{ij,J_j,\omega}$ be a random J_j-dependent piecewise continuous function with support contained in $[0, \infty)$, for almost all (J_j, ω). Let $K_{ij,J_j,\omega}$ be an inhomogeneous Poisson process of intensity $h_{ij,J_j,\omega}$. Furthermore, suppose that for each $i, j \in [d]$, $\mathbb{E}_{J_j}\mathbb{E}_{\omega|J_j}\|h_{ij,J_j,\omega}\|_{L^\infty} < \infty$, i.e., that $h_{ij,J,\omega}$ is a.s. bounded.*

Now let $T \in [0, \infty]$, and define a point process $\boldsymbol{N}(\cdot)$ through a sequence of events generated according to the following procedure:

(i) For $j \in [d]$, let $I_j(\cdot)$ be a homogeneous Poisson process of rate $\lambda_{j,0}$, generating immigration events

$$\{(t_r^{(0)}, j, J_r^{(0)})\}_{r=1}^{R_j^{(0)}(t)},$$

and where $J_r^{(0)} \overset{\text{iid}}{\sim} J_j$ and where $R_j^{(0)}(t)$ is the number of immigration events in component j up to time $t \in [0, T]$.

(ii) For each immigration event $(t_r^{(0)}, j, J_r^{(0)})$, in each target component $m \in [d]$, generate first-generation events

$$\{(t_r^{(1)}, m, J_r^{(1)})\}_{r=1}^{R_m^{(1)}(t)}$$

according to $K_{mj,J_r^{(0)},\omega}(t - t_r^{(0)})$, where $J_r^{(1)} \overset{\text{iid}}{\sim} J_m$.

(iii) Upon iterating the above rule, given an r-th event of the $(n-1)$-st generation in source component $m \in [d]$, for each target component $l \in [d]$, descendant $(t_r^{(n-1)}, m, J_r^{(n-1)})$ generates n-th generation events

$$\{(t_r^{(n)}, l, J_r^{(n)})\}_{r=1}^{R_l^{(n)}(t)},$$

according to $K_{lm,J_r^{(n-1)},\omega}(t - t_r^{(n-1)})$, and where $J_r^{(n)} \overset{\text{iid}}{\sim} J_l$.

Here, the Poisson processes are conditionally independent within and between each iteration, and the excitation functions are drawn conditionally independently. Then,

$$\boldsymbol{N}(t) = \bigcup_{n\geqslant 0}\left(\{(t_r^{(n)}, 1, J_r^{(n)})\}_{r=1}^{R_1^{(n)}(t)} \times \cdots \times \{(t_r^{(n)}, d, J_r^{(n)})\}_{r=1}^{R_d^{(n)}(t)}\right)$$

is the resulting multivariate sojourn-time dependent Hawkes process.

In Definition 2, having drawn lifetimes and excitation functions, one can construct the corresponding birth-death and conditional intensity processes, $\boldsymbol{Q}(\cdot), \boldsymbol{\Lambda}(\cdot)$, in a straightforward manner.

The cluster representation from Definition 2 exhibits the following useful properties. First, modulo the time shift corresponding to the arrival times, clusters generated by immigrants in the same coordinate are i.i.d. Second, cluster processes are generated independently across source components. Finally, within each source component, every event produces offspring using an identical iterative procedure, as each child represents a cluster, thus demonstrating *self-similarity*.

For later use, these cluster properties can be operationalized using notation borrowed from [30]. For an arrival in coordinate j, denote the d-dimensional counting, birth-death and intensity cluster process it generates by $\boldsymbol{S}_j^{\boldsymbol{N}}(\cdot), \boldsymbol{S}_j^{\boldsymbol{Q}}(\cdot), \boldsymbol{S}_j^{\boldsymbol{\Lambda}}(\cdot)$, respectively. Those have i-th coordinate $S_{i\leftarrow j}^{\bigstar}(\cdot)$ for $\bigstar \in \{\boldsymbol{N}, \boldsymbol{Q}, \boldsymbol{\lambda}\}$. Here, $S_{i\leftarrow j}^{\boldsymbol{N}}(u)$ records the number of events in component i up to time u with as oldest ancestor the arrival generating $\boldsymbol{S}_j^{\boldsymbol{N}}(\cdot)$, including the arrival itself when $i = j$. Similarly, $S_{i\leftarrow j}^{\boldsymbol{Q}}(u)$ records the number of nonexpired events in component i up to time u with as oldest ancestor the arrival generating $\boldsymbol{S}_j^{\boldsymbol{Q}}(\cdot)$, including the ancestor itself if $i = j$ and if the ancestor has not yet left the system. Finally, $S_{i\leftarrow j}^{\boldsymbol{\Lambda}}(u)$ records aggregated change in the intensity of component i caused by jumps with excitation functions $h_{im,J,\omega}$, following arrivals in component m with sojourn time J within the cluster $\boldsymbol{S}_j^{\boldsymbol{\Lambda}}(\cdot)$ generated by an arrival in component j.

Next, we define a d-dimensional network $(\boldsymbol{N}(t), \boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))_{t\geqslant 0}$ of (linear, Markovian) delayed Hawkes birth-death processes.

Definition 3 (Network of delayed Hawkes birth-death processes)**.** *Let $d \in \mathbb{N}$, and let $\mu_j, \mu_{ij} \geqslant 0$ for all $i, j \in [d]$, such that for each $j \in [d]$, either $\mu_j > 0$, or there is a sequence $(i_1, \ldots, i_k) \subset [d]$ such that $\mu_{i_k}\mu_{i_k i_{k-1}}\mu_{i_{k-1}i_{k-2}} \cdots \mu_{i_2 i_1}\mu_{i_1 j} > 0$. For each $i, j \in [d]$, let $\lambda_{i,0} \geqslant 0$, let $(B_{ij}(s))_{s\in\mathbb{R}}$ be a collection of cross-sectionally and serially independent distributed random marks, distributed as the generic random variable B_{ij}, which is assumed to be positive a.s., and let $h_i(t) = \exp(-r_i t)$, $r_i > 0$. Suppose that $\boldsymbol{N}(0) = \boldsymbol{Q}(0) = \boldsymbol{0}$ and $\boldsymbol{\Lambda}(0) = \boldsymbol{\lambda}_0 := [\lambda_{1,0} \cdots \lambda_{d,0}]^\top$. A network of delayed Hawkes birth-death processes involves a d-dimensional point process $\boldsymbol{N}(\cdot)$, taking values in $\mathbb{N}_0^d$, whose components $N_i(\cdot)$ satisfy, as $\Delta t \downarrow 0$,*

$$\mathbb{P}(N_i(t+\Delta t) - N_i(t) = 0 \,|\, \mathcal{H}_t) = 1 - \Lambda_i(t)\Delta t + o(\Delta t),$$
$$\mathbb{P}(N_i(t+\Delta t) - N_i(t) = 1 \,|\, \mathcal{H}_t) = \Lambda_i(t)\Delta t + o(\Delta t),$$

$$\mathbb{P}(N_i(t+\Delta t) - N_i(t) \geqslant 2 \,|\, \mathcal{H}_t) = o(\Delta t).$$

Suppose that the network of birth-death processes $\boldsymbol{Q}(\cdot)$ satisfies the following dynamics. (We write $\boldsymbol{e}_i$ for the i-th standard unit vector in $\mathbb{R}^d$.)

- *Arrivals, which are jumps upwards by $\boldsymbol{e}_i$, match jumps in $N_i(\cdot)$;*
- *Rerouting from coordinate j to i, that is, a jump by $\boldsymbol{e}_i - \boldsymbol{e}_j$, occurs with probability $\mu_{ij}Q_j(t)\Delta t + o(\Delta t)$ in $(t, t+\Delta t)$;*
- *Departures, which are jumps downwards by $\boldsymbol{e}_j$, occur with probability $\mu_j Q_j(t)\Delta t + o(\Delta t)$ in $(t, t+\Delta t)$.*

Now let $\boldsymbol{D}(\cdot)$ be the departure process, taking jumps upwards by $\boldsymbol{e}_j$ precisely when there is a departure from the system in coordinate j, i.e., when $\boldsymbol{Q}(\cdot)$ jumps downwards by $\boldsymbol{e}_j$. The intensity $\Lambda_i(\cdot)$ of component i is given by

$$\Lambda_i(t) = \lambda_{i,0} + \sum_{j=1}^{d} \int_{(0,t)} B_{ij}(s) h_i(t-s) \, \mathrm{d}D_j(s). \tag{6}$$

The $\mathcal{H}_t$-progressively measurable process $\boldsymbol{\Lambda}(\cdot)$ is called the conditional intensity process. (In (6)*, we may integrate over $(-\infty, t)$ in order to study the process in stationarity.)*

We note that the process from Definition 3 is a piecewise deterministic Markov process (e.g., [46]) with stochastically bounded jump rates, hence well-posedness can be readily verified, and stability follows under a spectral radius condition similar to the one given in Theorem 1 below.

A particle in coordinate j moves away at rate $\tilde{\mu}_j := \mu_j + \sum_{i=1}^d \mu_{ij}$, after which it leaves the system with probability $\mu_j/\tilde{\mu}_j$, and is rerouted to coordinate i with probability $\mu_{ij}/\tilde{\mu}_j$. Note that we do not assume that we have a feed-forward network: we allow for the possibility of loops. Although in natural applications one would typically set $\mu_{jj} = 0$, we do not make that assumption either. A particle creates excitation as soon as it leaves the system. Because of the possibility of rerouting, this is not necessarily in the coordinate where the particle arrived. It is possible to study a model where rerouting creates excitation as well: this yields similar results as those found in Section 7.1.

At a departure in coordinate j, there is a jump B_{ij} in each coordinate i, so that we have *mutual excitation*. We let $\boldsymbol{B}_j = [B_{1j} \cdots B_{dj}]^\top$ denote the vector of marks resulting from a departure in coordinate j.

3. Existence, uniqueness and stability

In this section, we prove that there exists a unique stationary distribution for the process $\boldsymbol{N}(\cdot)$ from Definition 1 having nonlinear sojourn-time dependent excitation, and we state conditions under which a transient process satisfying the given dynamics is shown to converge to this stationary distribution. In contrast to classical Hawkes, at each arrival a random excitation function is drawn, whose distribution depends on the sojourn time realization. It suffices to consider a model with i.i.d. random excitation functions h_{ij} having a distribution only depending on (i, j); the J-dependent randomness of the form $\omega|J$ occurs as a special case of this general randomness.

Let $(N_r, j_r, h_{1r}, \ldots, h_{dr})_{r\in\mathbb{Z}}$ be the events of a random-function marked point process $\boldsymbol{N}(\cdot)$, where N_r denotes the r-th event after time 0 for $r \geqslant 1$, and the $-(r+1)$ event before time 0 for $r \leqslant 0$; where j_r denotes the coordinate in which this event occurred; and where $h_{ir} \sim h_{ij_r}$ denotes

the excitation function for coordinate i associated to the r-th arrival. Let $(\Omega_{ij}, \mathcal{F}_{ij}, \mathbb{Q}_{ij})$ be the probability space on which h_{ij} is defined. Letting $\mathcal{H}_t^{\boldsymbol{N}}$ be the history of $\boldsymbol{N}(\cdot)$ up to time t, we assume that the model is driven by an $\mathcal{H}_t^{\boldsymbol{N}}$-progressively measurable intensity with i-th coordinate

$$\Lambda_i(t) = \phi_i\left(\sum_{j=1}^{d} \int_{(-\infty,t)\times\Omega_{ij}} h_{ij}(t-s,\omega_{ij}) N_j(\mathrm{d}s \times \mathrm{d}\omega_{ij})\right), \tag{7}$$

with the understanding that the random functions $\omega_{ij} \mapsto h_{ij}(\cdot, \omega_{ij})$ are drawn independently with common distribution h_{ij}, for all $i, j \in [d]$. We assume that $\phi_i : \mathbb{R} \to \mathbb{R}_+$ and $h_{ij} : \mathbb{R}_+ \to \mathbb{R}$, for all $i, j \in [d]$. In the linear case, which is the main focus of this paper, $\phi_i(x) = \lambda_{i,0} + x$ and $h_{ij} : \mathbb{R}_+ \to \mathbb{R}_+$. In the univariate case, (7) reduces to

$$\Lambda(t) = \phi\left(\int_{(-\infty,t)\times\Omega} h(t-s,\omega) N(\mathrm{d}s \times \mathrm{d}\omega)\right), \tag{8}$$

where $(\Omega, \mathcal{F}, \mathbb{Q})$ is the probability space on which the random functions $\omega \mapsto h(\cdot, \omega)$ are defined.

We construct an adapted point process: N on $\mathbb{R} \times \Omega$ with intensity $\Lambda(t)\mathbb{Q}(\mathrm{d}\omega)$ in the univariate case, and $\boldsymbol{N}$ on $\prod_{i=1}^{d}\left(\mathbb{R} \times \prod_{j=1}^{d} \Omega_{ji}\right)$ with intensity $\Lambda_i(t)\prod_{j=1}^{d}\mathbb{Q}_{ji}(\mathrm{d}\omega_{ji})$ in coordinate i in the multivariate case. In Appendix A, we present a proof for existence, uniqueness and stability of the univariate process having dynamics (8), leveraging the classical Picard proof for the existence of solutions to a differential equation, following the approach of [15], §14.3 and [10], Theorem 1 and using ideas from [36]. From this, the multivariate results can be proved along the lines of [10], Theorem 7, taking the randomness of the excitation functions into account in the same fashion as we do in the univariate case.

The conditional intensity specification (8) deals with i.i.d. random excitation functions, which can be seen to exist by invoking the Kolmogorov extension theorem. However, for a single random function, this construction only enables us to say something about the behavior of the function on a countable subset of $\mathbb{R}$, but in general this does not allow us to conclude anything about sample-path properties, such as measurability. To tackle this problem, we make additional assumptions on the generic random function $h(\cdot)$.

Definition 4. *A random function h is called separable with respect to a class $\mathcal{K}$ of subsets of $\mathbb{R}$ if there exists a countable subset $C \subset \mathbb{R}$ such that for each $K \in \mathcal{K}$ and each open interval $I \subset \mathbb{R}$ it holds that*

$$\bigcap_{t\in I\cap C}\{h(t) \in K\} = \bigcap_{t\in I}\{h(t) \in K\}, \quad a.s.$$

We typically assume that the random function h is a.s. piecewise continuous. In that case, h is separable with respect to the class of open subsets of $\mathbb{R}$, taking C to be any countable dense subset of $\mathbb{R}$, e.g., the set of rational numbers. By [37], §III.4, measurability of $(\omega, t) \mapsto h_\omega(t)$ can then be ensured. The feasibility of such a construction essentially comes down to the separability of the range space of the excitation functions.

In the following, we construct a univariate process having dynamics (8) upon a basis consisting of a bivariate Poisson process of unit rate marked by random functions $h(\cdot, \omega)$, for which we use Lemma 1 below. This is a well-known result underlying many simulation algorithms of point processes driven by conditional intensities, see e.g., [38]. Compare [36], Lemma 1. To state this

lemma, we define the *left-shift operator* S_t, $t \in \mathbb{R}$. For a univariate stochastic process X, we set $S_t X(A) = X(A + t)$, for all $A \in \mathcal{B}(\mathbb{R})$, with $\mathcal{B}$ the σ-algebra of Borel sets. Furthermore, we set

$$X_+ = \{X(A) : A \in \mathcal{B}([0, \infty))\}, \quad X_- = \{X(A) : A \in \mathcal{B}((-\infty, 0])\}.$$

With this notation, $S_t X_\pm$ can be interpreted as the future/history at time t. For a multivariate stochastic process, we assume that this shift is done with respect to the first variable, which is to be interpreted as time. In particular, the *history* at time t of a $(d+1)$-dimensional process Y is given by $S_t Y_- := \{Y(A) : A \in \mathcal{B}((-\infty, t] \times \mathbb{R}^d)\}$.

Lemma 1. *Let M be a marked Poisson process on $\mathbb{R} \times \mathbb{R}_+ \times \Omega$ with intensity $\mathrm{d}t \times \mathrm{d}s \times \mathbb{Q}(\mathrm{d}z)$, where the marks are defined on $(\Omega, \mathcal{F}, \mathbb{Q})$. Let $\mathcal{H}_t^M$ be a sigma-algebra containing the history of M at time t, such that $\mathcal{H}_s^M$ is independent of $S_t M_+$ for $s < t$. For some $\mathcal{H}_t^M$-predictable process $\Lambda(\cdot)$, define*

$$N(A \times B) = \int_{A \times \mathbb{R}_+ \times B} \mathbf{1}_{[0, \Lambda(t)]}(s) M(\mathrm{d}t \times \mathrm{d}s \times \mathrm{d}z), \quad A \times B \in \mathcal{B}(\mathbb{R}) \otimes \mathcal{F}. \tag{9}$$

Then N admits $\Lambda(t)\mathbb{Q}(\mathrm{d}z)$ as an $\mathcal{H}_t^M$-intensity.

The (lengthy) proof of the next result is postponed until Appendix A.

Theorem 1 (Existence, uniqueness and stability)**.** *Assume that $\phi_i : \mathbb{R} \to \mathbb{R}_+$ is L_i-Lipschitz for all $i \in [d]$. Suppose that for all $i, j \in [d]$, $h_{ij}(\cdot, \omega)$ is a random function defined on $(\Omega_{ij}, \mathcal{F}_{ij}, \mathbb{Q}_{ij})$, which is separable with respect to the class of open sets, such that $h_{ij}(t) \in L^1(\mathbb{Q}_{ij})$ for almost all $t \in \mathbb{R}$, and such that the $d \times d$ matrix $\|\boldsymbol{H}\| := (L_i \|\mathbb{E}|h_{ij}|\|_{L^1})_{i,j \in [d]}$ has spectral radius less than 1. Then there exists a stationary distribution for a process $\boldsymbol{N}(\cdot)$ satisfying the dynamics (7).*

In addition, assume that $\|\mathbb{E}|h_{ij}|\|_\infty < \infty$ for all $i, j \in [d]$. Then this stationary distribution is unique. Let

$$i_c(t) = \sum_{i,j \in [d]} \mathbb{E}_{h_{ij}} \left[\int_{t-c}^{t} \int_{(-\infty, 0) \times \Omega_{ij}} |h_{ij}(s - \tau, \omega_{ij})| \, N_j(\mathrm{d}\tau \times \mathrm{d}\omega_{ij}) \, \mathrm{d}s \right]. \tag{10}$$

Let M^d be a multivariate version of the marked Poisson process from Lemma 1, i.e., a marked Poisson process on $\mathbb{R} \times \mathbb{R}_+^d \times \prod_{i=1}^d \prod_{j=1}^d \Omega_{ji}$ with intensity $\mathrm{d}t \times \prod_{i=1}^d \left(\mathrm{d}s_i \times \prod_{j=1}^d \mathbb{Q}_{ji}(\mathrm{d}\omega_{ji}) \right)$. Suppose that $\boldsymbol{N}$ is defined w.r.t. M^d. If (i) for all $c > 0$, $\sup_{t \geqslant 0} i_c(t) < \infty$ and $\lim_{t \to \infty} i_c(t) = 0$, a.s., or (ii) for all $c > 0$, $\sup_{t \geqslant 0} \mathbb{E}_{M^d} i_c(t) < \infty$ and $\lim_{t \to \infty} \mathbb{E}_{M^d} i_c(t) = 0$, a.s., then for any $\tilde{\boldsymbol{N}}$ satisfying (10) with dynamics (7) on $\mathbb{R}_+$, we have $S_t \tilde{\boldsymbol{N}}_+ \xrightarrow{\mathcal{D}} \boldsymbol{N}_+$, as $t \to \infty$; i.e., we have stability in distribution.

Remark 1. *Both initial conditions (i) and (ii) say, in different ways, that the influence of the history at time 0, i.e., the behavior on $(-\infty, 0]$, on the future at time t, i.e., the behavior on $[t, \infty)$, vanishes, as $t \to \infty$.*

Remark 2. *Whereas existence, uniqueness and stability for the three specific processes given in Eqn. (5) is, in principle, already implied by [36], our Theorem 1 above is more explicit. To apply [36], consider, for example, the univariate delayed Hawkes process. We can define a point process*

on $\mathbb{R}\times\mathbb{R}_+\times\mathbb{R}_+$, where the coordinates represent time, marks and sojourn times, respectively. Then the conditional intensity can be written as

$$\Lambda(t,\mathrm{d}b,\mathrm{d}w)=\psi(S_tN_-)B(\mathrm{d}b)J(\mathrm{d}w),$$

where B denotes the mark distribution, J denotes the sojourn time distribution, and

$$\psi(S_tN_-)=\phi\left(\int_{(-\infty,t)\times\mathbb{R}_+\times\mathbb{R}_+}bh(t-s-w)\mathbf{1}_{[w,\infty)}(t-s)N(\mathrm{d}s\times\mathrm{d}b\times\mathrm{d}w)\right).$$

Theorem 1 may then be compared to [36, Theorems 2, 4]*: it gives more concrete conditions on $h(\cdot,\omega)$, and allows for a direct proof.*

4. Scaling limit with stretched sojourn times

Having formally introduced the general family of models having sojourn-time dependent excitation, we ask ourselves to what extent members of this family differ, statistically and probabilistically, from the classical Hawkes process. In fact, it turns out to be possible to distinguish between a Hawkes and a delayed Hawkes process from observed sample paths using statistical techniques from [3], as we outline in Appendix B.

In this section, we approach the problem of distinguishing between a Hawkes process and a delayed Hawkes process probabilistically, by analyzing asymptotic, macroscopic behavior through *scaling limits*. That is, we look for convergence at process level of some scaled version of the process. This convergence is weakly in $D[0,1]$, the space of càdlàg functions on the unit interval, equipped with the Skorokhod J_1-topology. (Note that the "J_1" in "J_1-topology" is unrelated to the random delay J.) We consider the linear, univariate case. In a typical scaling regime, one considers the compensated counting process (see [15], §7.2); one contracts time by a factor T; after which one divides by $\sqrt{T}$. This is the quantity studied in a *functional central limit theorem* (FCLT).

For the unmarked Hawkes process, a scaling limit of this type can be found in [5]. For a univariate model with immigration intensity λ_0 and excitation function h, their results imply that

$$\frac{N(T\cdot)-\mu T\cdot}{\sqrt{T}}\to\sigma B(\cdot),\tag{11}$$

as $T\to\infty$, weakly on $D[0,1]$ equipped with the Skorokhod J_1-topology, where B is a standard Brownian motion, and where

$$\mu:=\frac{\lambda_0}{1-\|h\|_{L^1}},\quad \sigma^2:=\frac{\lambda_0}{(1-\|h\|_{L^1})^3}.\tag{12}$$

On the other hand, for a model having sojourn-time dependent excitation, we can apply an existing FCLT for marked Hawkes random measures, as given in [25], Theorem 3.12. Since the scaling limit considers the counting process instead of the population process, we can, as in Section 3, replace the sojourn-time dependency of the random excitation function by general randomness. Letting $\mathbb{U}$ be a Lusin space modeling the randomness of the excitation functions, we use marks $\omega\in\mathbb{U}$ and excitation functions $h(t,\omega)$. It follows from [25], Theorem 3.12, that any two processes with random excitation functions *having the same expected L^1-norm* admit the same scaling limit of the FCLT type (i.e., take a compensated process; contract time by a factor T; divide by $\sqrt{T}$).

In particular, we can compare a Hawkes process to a delayed Hawkes process having the same parameters, corresponding to bivariate marks $\boldsymbol{\xi} \in \mathbb{R}_+^2$ whose coordinates represent 'actual' mark and sojourn time, respectively, and excitation functions

$$\phi_{\text{Hawkes}}(t, \boldsymbol{\xi}) = \xi_1 h(t), \quad \text{and} \quad \phi_{\text{delayed}}(t, \boldsymbol{\xi}) = \xi_1 h(t - \xi_2)\mathbf{1}\{t \geqslant \xi_2\},$$

to infer that they admit the same scaling limit, being the sum of a Gaussian white noise (contributed by the marks) and a correlated Brownian motion, having the same parameters for both models. Heuristically, if we contract time, deviations from the mean from the random excitation functions cancel each other out. For the delayed Hawkes process, if sojourn times stay the same, but if we contract time by a factor T, the delays are of length J/T, hence vanish, as $T \to \infty$.

A natural, subsequent question is whether the difference between two processes belonging to the family of processes having sojourn-time dependent excitation can be made visible in some scaling limit. To this end, we consider a univariate unmarked delayed Hawkes process with i.i.d. sojourn times $(J_i)_i \in \mathbb{N}$, which we compare to its nondelayed counterpart. The idea is to consider the compensated process on an interval $[0, T]$ with sojourn times stretched out from J_i to $T^\alpha J_i$, for some $\alpha \in [0, 1)$, after which we contract time by a factor of T, mapping $[0, T]$ onto $[0, 1]$. After rescaling our counting process by $T^{-1/2}$ and letting $T \to \infty$, we obtain a nondegenerate limit. By considering sojourn-time stretching of degree α with $0 \leqslant \alpha < \frac{1}{2}$, we will obtain the same scaling limit as given by (11)–(12), whereas if we set $\alpha = \frac{1}{2}$, the effect of the delays will become visible. The case $\alpha > \frac{1}{2}$, discussed in Remark 4 below, is less transparent.

Remark 3. *From an applied perspective, the scaling/stretching parameter α should not be interpreted as a model parameter to be estimated from a single dataset on a fixed time scale. Rather, it parametrizes how the characteristic delay scale (e.g., service times, incubation periods, execution lags, engagement times) grows relative to the observation horizon T in an asymptotic sequence of models indexed by T, and thereby determines which asymptotic* regime *provides an accurate approximation to the system. If typical delays remain of order $O(1)$ while the observation window $T \to \infty$, then effectively $\alpha = 0$ and the delayed Hawkes process falls into the classical Hawkes FLLN/FCLT regime, sharing the same FLLN/FCLT behavior as the classical Hawkes process. More generally, if delay distributions grow with the observational scale like T^α, then α quantifies that relative growth. The regime $\alpha = 1/2$ is distinguished because it marks the critical threshold at which delays stop averaging out under diffusion scaling and start contributing to second-order fluctuations. The results in this section therefore clarify under which conditions delayed excitation has a visible impact at the* macroscopic *level. Such regimes may arise in systems where waiting times grow with system size or congestion, for instance in overloaded service systems, epidemics with prolonged incubation or reporting delays, illiquid markets with slow execution, or online cascades with long-lived engagement.*

To gain insight into the scaling limits, we adapt and extend the arguments of [5] to our setting. Since several technical steps follow a similar overall strategy as in [5], we present full proofs of the scaling limits in the online Supplementary Material [4]. Here, we focus on the construction and the statement of the results.

Let $(N_\alpha^T(v))_{v\in[0,1]}$ be equal to $(\tilde{N}_\alpha^T(Tv))_{v\in[0,1]}$, with $\tilde{N}_\alpha^T(\cdot)$ the process having sojourn times $(T^\alpha J_i)_{i\in\mathbb{N}}$. Those sojourn times correspond to the increasing sequence of arrival times $(\tau_i)_{i\in\mathbb{N}}$,

where it is assumed that J_i is drawn at time τ_i. This process $N_\alpha^T(\cdot)$ has an arrival intensity $\Lambda_\alpha^T(\cdot)$ given by

$$\Lambda_\alpha^T(v) = \left(\lambda_0 + \sum_{\tau_i<t} h(Tv - \tau_i - T^\alpha J_i)\mathbf{1}\{Tv \geqslant \tau_i + T^\alpha J_i\}\right)\cdot T. \tag{13}$$

To derive our scaling limit, we impose the following three assumptions:

$$\|h\|_{L^1} < 1, \tag{A1}$$

$$\int_0^\infty t^{\frac{1}{2(1-\alpha)}} h(t)\,\mathrm{d}t < \infty, \tag{A2}$$

$$\mathbb{E}[J] < \infty. \tag{A3}$$

We assume (A1) throughout this section, whereas we use (A2)–(A3) only for $\alpha \leqslant \frac{1}{2}$.

In the proofs, which are postponed to the online Supplementary Material [4], we use the function $\bar{h}^T(\cdot)$, which can be seen as an average of h over the past, weighed according to the stretched sojourn times:

$$\bar{h}^T(t) := \int_0^{T^{-\alpha}t} h(t - T^\alpha w)\,\mathrm{d}\bar{\mathscr{F}}(w), \tag{14}$$

where $\bar{\mathscr{F}}(\cdot)$ denotes the CDF of the i.i.d. lifetimes J_i. We also define

$$\mathscr{H}^T := \sum_{k\geqslant 1} (\bar{h}^T)^{*k}, \tag{15}$$

where $*k$ denotes k-fold convolution. Note that for any $T > 0$, $\alpha \in [0,1)$,

$$\begin{aligned}\|\bar{h}^T\|_{L^1} &= \int_0^\infty \int_0^{T^{-\alpha}t} h(t - T^\alpha w)\,\mathrm{d}\bar{\mathscr{F}}(w)\,\mathrm{d}t = \int_0^\infty \int_{T^\alpha w}^\infty h(t - T^\alpha w)\,\mathrm{d}t\,\mathrm{d}\bar{\mathscr{F}}(w)\\ &= \int_0^\infty \int_0^\infty h(t)\,\mathrm{d}t\,\mathrm{d}\bar{\mathscr{F}}(w) = \|h\|_{L^1},\end{aligned} \tag{16}$$

and therefore, using that $\|(\bar{h}^T)^{*k}\|_{L^1} = \|\bar{h}^T\|_{L^1}^k$, as can easily be proved by induction,

$$\|\mathscr{H}^T\|_{L^1} = \sum_{k\geqslant 1} \|(\bar{h}^T)^{*k}\|_{L^1} = \sum_{k\geqslant 1} \|h\|_{L^1}^k = \frac{\|h\|_{L^1}}{1 - \|h\|_{L^1}} < \infty.$$

We are now equipped to state an FLLN for $\left(N_\alpha^T(\cdot)\right)_{T\geqslant 0}$. After the FLLN, we present our FCLT for α-stretched sojourn times. The proofs can be found in the online Supplementary Material [4].

Theorem 2 (FLLN)**.** *Let $\alpha \in [0,1)$. It holds that $\tilde{N}_\alpha^T(t) \in L^2(\mathbb{P})$, for all $T, t \geqslant 0$, and we have*

$$\sup_{v\in[0,1]} \left|T^{-1}N_\alpha^T(v) - \mu\right| \to 0 \text{ in } L^2(\mathbb{P}),\ \text{as } T\to\infty, \tag{17}$$

where μ is as defined in (12).

Define

$$\bar{\sigma} := \frac{\lambda_0 \mathbb{E}[J]\|h\|_{L^1}}{(1 - \|h\|_{L^1})^2}. \tag{18}$$

Theorem 3 (FCLT)**.** *Let μ and σ be as in (12), and let $\bar{\sigma}$ be as in (18). Let $\epsilon \in (0,1)$, and let B be a standard Brownian motion. For $\alpha = \frac{1}{2}$, we have*

$$\left(\frac{N_{1/2}^T(v) - \mu T v}{\sqrt{T}}\right)_{v\in[\epsilon,1]} \longrightarrow \left(-\bar{\sigma} + \sigma B(v)\right)_{v\in[\epsilon,1]}, \tag{19}$$

as $T \to \infty$, weakly on $D[\epsilon, 1]$ equipped with the Skorokhod J_1-topology. On the other hand, for $\alpha \in [0, \frac{1}{2})$, it holds that

$$\left(\frac{N_\alpha^T(v) - \mu T v}{\sqrt{T}} \right)_{v \in [0,1]} \longrightarrow (\sigma B(v))_{v \in [0,1]}, \tag{20}$$

as $T \to \infty$, weakly on $D[0, 1]$ equipped with the Skorokhod J_1-topology.

For $\alpha = \frac{1}{2}$, Theorem 3 yields convergence on intervals of the form $[\epsilon, 1]$, where $\epsilon > 0$ can be taken arbitrarily small. This is in contrast to the case $\alpha \in [0, \frac{1}{2})$ and to [5], Theorem 2, where we obtain convergence on the whole unit interval. For each $\alpha \in [0, \frac{1}{2}]$, both the centralising constant μ and the Brownian term are the same. A notable difference is that in Theorem 3 with $\alpha = \frac{1}{2}$ there is a 'correction term' $-\bar{\sigma}$ in the limit.

We can explain this result *heuristically*. In the limiting result (19), we start observing the process at time $T\epsilon$, for fixed $\epsilon > 0$. For large T, this means that the process approaches stationarity on $[0, T\epsilon)$. By Corollary 2 below — which covers the Markovian case — and the heuristic explanation given thereafter, there is good reason to believe that Hawkes and delayed Hawkes processes have the same stationary distributions. Therefore, we expect to find similar limits. However, in (19) delays were also stretched out by a factor of $T^{1/2}$, meaning that excitation takes more time to come into full effect, which causes $\mu T \cdot$ to overestimate the mean of $N_{1/2}^T(\cdot)$ on $[0, \epsilon)$. This is compensated for by the negative term $-\bar{\sigma}$ appearing in the limit.

Remark 4. *Under (A1), for any $\alpha \in [0, 1)$, it is possible to find a FCLT as in Theorem 3, stating that, as $T \to \infty$,*

$$\left(\frac{N_\alpha^T(v) - \mathbb{E}[N_\alpha^T(v)]}{\sqrt{T}} \right)_{v \in [0,1]} \longrightarrow (\sigma B(v))_{v \in [0,1]}, \tag{21}$$

weakly on $D[0, 1]$ equipped with the Skorokhod J_1-topology. When $\alpha \in (\frac{1}{2}, 1)$, in contrast to the case $\alpha \in [0, \frac{1}{2}]$, we cannot use Lemma A.III from the Online Appendix [4] *to replace $\mathbb{E}[N_\alpha^T(v)]$ in this expression, since we need $T^{(1-\alpha)p} \cdot T^{-1} = T^{-1/2}$, but Lemma A.III from the Online Appendix* [4] *only describes $p \leqslant 1$.*

For $\alpha = 1$, if we take sojourn times having support on $[1, \infty)$, the excitation would not be visible, since in the scaling limit we observe the process on (a subset of) $[0, 1]$. In this case, the unscaled process on $[0, T]$ would just be a homogeneous Poisson process of rate λ_0, for which an FCLT holds; e.g., use (11)–(12) with $h \equiv 0$. When $\alpha > 1$, we would see the same behavior. The situation where $\alpha = 1$ and *where the sojourn time attains values in $(0, 1)$ with positive probability is more delicate.*

5. Transform analysis and heavy-tailed asymptotics

In this section, we perform transform analysis for point processes having sojourn-time dependent excitation. First, in Section 5.1, we use cluster-representation based methods to describe fixed points in the transform domain, after which, in Section 5.2, those fixed-point equations are used to derive heavy-tailed asymptotics. A supplement to this section can be found in Appendix D, where we study cluster size distributions for gamma-distributed marks.

5.1. **Transform characterizations with sojourn-time dependent excitation.** In [30], multivariate non-Markovian Hawkes-fed birth-death processes were studied using cluster-representation based methods. In Definition 2, we gave a cluster representation for the d-dimensional birth-death process with sojourn-time dependent excitation, analogous to the one for the multivariate Hawkes-fed birth-death process. As it turns out, the cluster representation is the pivotal ingredient for the results from [30], §3–4: Definition 2 enables us to obtain analogous results for our general family of models having sojourn-time dependent excitation. The modifications needed in the respective proofs are relatively straightforward, and mostly come down to suitably replacing randomness of the form $B_{ij,\omega}h_{ij}$ by sojourn-time dependent randomness of the form $h_{ij,J,\omega}$. Therefore, to save space, we provide the proof of the next result in online Supplementary Material [4].

Theorem 4. *Consider the joint birth-death and intensity process $(\boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))$ from Definition 2. Under the regularity conditions given there, the joint Z- and Laplace transform of $(\boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))$ can be expressed as*

$$\mathbb{E}\left[z^{\boldsymbol{Q}(t)}e^{-\boldsymbol{s}^\top\boldsymbol{\Lambda}(t)}\right] = \prod_{j=1}^{d}\exp\left(-\lambda_{j,0}\left(t+s_j-\int_0^t \mathbb{E}\left[z^{\boldsymbol{S}_j^{\boldsymbol{Q}}(u)}e^{-\boldsymbol{s}^\top\boldsymbol{S}_j^{\boldsymbol{\Lambda}}(u)}\right]\,\mathrm{d}u\right)\right), \tag{22}$$

where the cluster processes $\boldsymbol{S}_j^{\boldsymbol{Q}}(\cdot)$, $\boldsymbol{S}_j^{\boldsymbol{\Lambda}}(\cdot)$ are defined in Section 2.

Combine the cluster processes for individual coordinates into a matrix $\boldsymbol{S}^\star(\cdot)$ with j-th column $\boldsymbol{S}_j^\star(\cdot)$, for $\star \in \{\boldsymbol{Q}, \boldsymbol{\Lambda}\}$. Then the joint vector-valued transform $\mathcal{J}_{\boldsymbol{S}^{\boldsymbol{Q}},\boldsymbol{S}^{\boldsymbol{\Lambda}}}(\cdot)$ of $\boldsymbol{S}^{\boldsymbol{Q}}(\cdot)$, $\boldsymbol{S}^{\boldsymbol{\Lambda}}(\cdot)$, which has as j-th component the joint transform

$$u \mapsto \mathbb{E}\left[z^{\boldsymbol{S}_j^{\boldsymbol{Q}}(u)}e^{-\boldsymbol{s}^\top\boldsymbol{S}_j^{\boldsymbol{\Lambda}}(u)}\right], \tag{23}$$

is the unique point of ϕ, which maps the space $\mathbb{J}^d$ of vector-valued d-dimensional joint Z- and Laplace transforms $\mathcal{J}(\cdot)$ to itself, and is defined by

$$\mathcal{J}(\cdot) = \begin{bmatrix}\mathcal{J}_1(\cdot)\\ \vdots \\ \mathcal{J}_d(\cdot)\end{bmatrix} \mapsto \begin{bmatrix}\phi_1(\mathcal{J}_1,\ldots,\mathcal{J}_d)(\cdot)\\ \vdots \\ \phi_d(\mathcal{J}_1,\ldots,\mathcal{J}_d)(\cdot)\end{bmatrix} = \begin{bmatrix}\phi_1(\mathcal{J})(\cdot)\\ \vdots \\ \phi_d(\mathcal{J})(\cdot)\end{bmatrix} = \phi(\mathcal{J})(\cdot), \tag{24}$$

where for $j \in [d]$

$$\begin{aligned}\phi_j(\mathcal{J})(u) &\equiv \phi_j(\mathcal{J})(u,\boldsymbol{s},\boldsymbol{z}) \qquad (25)\\ &= \mathbb{E}_{J,\omega}\left[z_j^{\mathbf{1}\{J>u\}}\prod_{i=1}^{d}e^{-s_i h_{ij,J,\omega}(u)}\prod_{m=1}^{d}\exp\left(-\int_0^u h_{mj,J,\omega}(v)\,(1-\mathcal{J}_m(u-v,\boldsymbol{s},\boldsymbol{z}))\,\mathrm{d}v\right)\right].\end{aligned}$$

Furthermore, for any $\mathcal{J}^{(0)}(\cdot) \in \mathbb{J}^d$, the sequence $(\mathcal{J}^{(n)}(u))_{n\in\mathbb{N}_0}$ of iterates of $\mathcal{J}^{(0)}(\cdot)$ under ϕ, defined inductively by $\mathcal{J}^{(n)}(\cdot) := \phi(\mathcal{J}^{(n-1)})(\cdot)$, converges pointwise on intervals $[0,t]$ to the fixed point $\mathcal{J}_{\boldsymbol{S}^{\boldsymbol{Q}},\boldsymbol{S}^{\boldsymbol{\lambda}}}(u)$. That is, as $n\to\infty$, for any $u\in[0,t]$,

$$\mathcal{J}^{(n)}(u) \equiv \mathcal{J}^{(n)}(u,\boldsymbol{s},\boldsymbol{z}) \to \mathcal{J}_{\boldsymbol{S}^{\boldsymbol{Q}},\boldsymbol{S}^{\boldsymbol{\lambda}}}(u,\boldsymbol{s},\boldsymbol{z}) \equiv \mathcal{J}_{\boldsymbol{S}^{\boldsymbol{Q}},\boldsymbol{S}^{\boldsymbol{\lambda}}}(u). \tag{26}$$

Remark 5. *It is possible to generalize Theorem 4 to a feedforward network in which a particle in coordinate $j\in[d]$ is sent to coordinate $j+1$ after service. Here, it is understood that when $j=d$, the particle leaves the system after service. Letting $J_j,\ldots,J_d$ be the sojourn times of the components visited by the particle that arrived in component j, and assuming that the excitation*

function $h_{ij,J,\omega}$ is dependent on the total *time $J := \sum_{l=j}^{d} J_l$ spent in the system, we can obtain a result analogous to Theorem 4. The operator appearing in the fixed-point equation now reads*

$$\phi_j(\mathcal{J})(u,\boldsymbol{s},\boldsymbol{z}) = \mathop{\mathbb{E}}_{J_j,\ldots,J_d,\omega}\left[c(u)\prod_{m=1}^{d}\exp\left(-\int_0^u h_{mj,J,\omega}(v)\,(1-\mathcal{J}_m(u-v,\boldsymbol{s},\boldsymbol{z}))\;\mathrm{d}v\right)\right],$$

where

$$c(u) = \prod_{l=j}^{d} z_l^{\mathbf{1}\left\{\sum_{m=j}^{l} J_m \leqslant u, \sum_{m=j}^{l} J_m > u\right\}} \prod_{i=1}^{d} e^{-s_i h_{ij,J,\omega}(u)}.$$

An analysis treating multiple parallel tandem systems, as conducted for shot-noise processes in [33]*, is hard in the non-Markovian (delayed) Hawkes case: in contrast to a network of shot-noise processes, the sample paths of parallel (delayed) Hawkes networks influence each other. In the Markovian case, however, we are able to characterize the transform of* any *irreducible d-dimensional network; see Section 7.1.*

5.2. **Heavy-tailed asymptotics.** In this subsection, we specify the non-Markovian model from Section 5.1 to the one-dimensional delayed Hawkes case, so that the randomness in the excitation function is of the form $h_{J,\omega}(\cdot) = B_\omega h(\cdot - J)\mathbf{1}\{\cdot > J\}$. We show that if the marks B_ω are heavy-tailed — in the sense of being regularly varying — the birth-death process will be so as well. Our proof uses (22) and the fixed-point equation for the transform, (25).

Definition 5. *Let $\alpha > 0$. An a.s. positive random variable X is called regularly varying of index $-\alpha$ if*

$$\mathbb{P}(X > x) = \ell(x)x^{-\alpha}, \quad x \geqslant 0, \tag{27}$$

where ℓ is a slowly varying function at infinity, meaning that $\ell(\gamma x) \sim \ell(x)$ as $x \to \infty$, for all $\gamma > 1$. We write $\mathscr{R}(-\alpha)$ for the class of regularly varying random variables of tail index α.

We also use the stronger notion of asymptotically power-law tails.

Definition 6. *An a.s. positive random variable X is said to have an asymptotically power-law tail (APT) if there exist $C > 0$ and $\gamma > 1$ such that*

$$\mathbb{P}(X > x)x^{\gamma} \to C, \tag{28}$$

as $x \to \infty$. In this case we write $X \in \mathrm{APT}(-\gamma)$ and we refer to γ as the tail index.

The next result may be compared to [34], Theorem 6.2. Its (lengthy) proof is postponed until Appendix C.

Theorem 5. *Consider the univariate delayed Hawkes birth-death process with general sojourn times. Assume the stability condition $\|h\|_{L^1} b_1 < 1$, where $b_1 := \mathbb{E}[B]$. Suppose that $B \in \mathscr{R}(-\alpha)$ with $\alpha \in (1,2)$. Then also $Q(t) \in \mathscr{R}(-\alpha)$.*

Remark 6. *Theorem 5 admits various extensions.*

(i) We can take sojourn-time dependent marks, i.e., $h_{J,\omega}(\cdot) = B_{J,\omega}h(\cdot - J)\mathbf{1}\{\cdot > J\}$. Suppose that $B|J$ is either light-tailed, or regularly varying of index $-\alpha$ for some $\alpha > 1$, $\bar{\mathscr{J}}$-a.s., in such a way that the infimum of the α for which $B_w \in \mathscr{R}(-\alpha)$, lies in $(1,2)$, and is attained with positive $\bar{\mathscr{J}}$-probability. Expanding $\beta_w := \mathbb{E}[e^{-sB} \mid J = w]$ in (59) using the Tauberian

theorem for w such that $B|J = w$ is regularly varying, and using a Taylor expansion for other w, we obtain an equivalent of (61), after which we proceed as in the proof of Theorem 5.

(ii) If $\alpha \in (k, k+1)$, $k \in \{2, 3, \ldots\}$, the Tauberian theorem for a higher-order expansion yields a more involved, but conceptually analogous, proof for $Q(t) \in \mathscr{R}(-\alpha)$.

(iii) Theorem 5 admits a multivariate generalization, by following the arguments from [30]*, §5.*

(iv) A proof analogous to the proof of Theorem 5 shows that if we have regularly varying marks, those marks propagate to the intensity $\Lambda(t)$ as well.

The following corollary describes heavy-traffic behavior in the heavy-tailed setting; its proof is in Appendix C.

Corollary 1. *Assume that we are in the heavy-tailed setting of Theorem 5, with $B \in \mathrm{APT}(-\alpha)$ for some $\alpha \in (1, 2)$. Let $\rho = \|h\|_{L^1} b_1 < 1$, and write (Q, Λ) for the stationary distribution of $(Q(\cdot), \Lambda(\cdot))$. Then it holds that $(1-\rho)Q$ converges in distribution to some nondegenerate, nondefective random variable X with $\mathbb{E}[X^\alpha] = \infty$, as $\rho \uparrow 1$.*

6. Comparisons using stochastic ordering

In this section, we consider a multivariate Hawkes-fed birth-death process $(\boldsymbol{N}(t), \boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))_{t\in\mathbb{R}_+}$ with intensity $\Lambda_i(\cdot)$ in component i given by

$$\Lambda_i(t) = \lambda_{i,0} + \sum_{j=1}^{d} \int_{-\infty}^{t} B_{ij}(s) h_{ij}(t-s)\, \mathrm{d}N_j(s), \tag{29}$$

where, for each $i, j \in [d]$, $(B_{ij}(s))_{s\in\mathbb{R}}$ is a collection of cross-sectionally and serially independently distributed random variables distributed as the a.s. positive random variable B_{ij}. We compare this process to the corresponding multivariate delayed Hawkes birth-death process $(\tilde{\boldsymbol{N}}(t), \tilde{\boldsymbol{Q}}(t), \tilde{\boldsymbol{\Lambda}}(t))_{t\in\mathbb{R}_+}$ having the same parameters; its intensity $\tilde{\Lambda}_i(\cdot)$ in coordinate i is given by

$$\tilde{\Lambda}_i(t) = \lambda_{i,0} + \sum_{j=1}^{d} \int_{-\infty}^{t} B_{ij}(s) h_{ij}(t-s)\, \mathrm{d}\tilde{D}_j(s), \tag{30}$$

where $\tilde{D}_j(\cdot)$ denotes the departure process of the j-th coordinate.

In the univariate case, both systems can be specified through conditional intensities of the form

$$\Lambda(t) = \lambda_0 + \sum_{t_i<t} B_i h(t - t_i), \tag{31}$$

the only difference being that in the former, classical case $(t_i)_{i\in\mathbb{N}}$ denote arrival times for the Hawkes-fed birth-death process, whereas in the latter, delayed case $(t_i)_{i\in\mathbb{N}}$ denote departure times for the delayed Hawkes birth-death process. To argue that the delayed Hawkes process is in a sense 'dominated' by the Hawkes process, we consider a comparison using stochastic ordering.

Definition 7. *Let X, Y be random variables. We say that X is larger than Y in the stochastic order, or, equivalently, that X dominates Y, if $F_X(z) \leqslant F_Y(z)$ for all $z \in \mathbb{R}$. We write $X \geqslant_{\mathrm{st}} Y$.*

Another way of representing both birth-death processes is by considering their respective cluster process representations, as given via Definition 2. Then, by an obvious coupling, the baseline intensity $\boldsymbol{\lambda}_0$ generates the same stream of immigrants, and therefore the same multiplicity of clusters,

for both birth-death processes. Coupling those clusters as well, it is clear that $\boldsymbol{N}(t) \geqslant_{\mathrm{st}} \tilde{\boldsymbol{N}}(t)$ for all $t \geqslant 0$, since the clusters produce the same offspring for both processes, but k-th generation children are counted k lifetimes later for $\tilde{N}(t)$ than for $N(t)$. This observation immediately raises the question whether we can compare $\boldsymbol{Q}(t)$ to $\tilde{\boldsymbol{Q}}(t)$ and $\boldsymbol{\Lambda}(t)$ to $\tilde{\boldsymbol{\Lambda}}(t)$ in the stochastic order as well. This question is answered affirmatively by the following theorem.

Theorem 6. *Let $(\boldsymbol{N}(t), \boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))_{t\in\mathbb{R}_+}$ be a multivariate Hawkes-fed birth-death process, with conditional intensity given by* (29)*. Furthermore, let $(\tilde{\boldsymbol{N}}(t), \tilde{\boldsymbol{Q}}(t), \tilde{\boldsymbol{\Lambda}}(t))_{t\in\mathbb{R}_+}$ denote the corresponding delayed Hawkes birth-death process having the same parameters, i.e., its conditional intensity satisfies* (30)*. Assume both systems have the same sojourn time distributions J_i, having CDF $\bar{\mathcal{J}}_i$. For both systems, let the $(B_{ij}(s))_{s\geqslant 0}$ be i.i.d., independent of other random variables driving the processes. In both cases, suppose that we start in an empty system with zero arrivals, and a conditional intensity equal to the baseline intensity $\boldsymbol{\lambda}_0$. Then we have for all $j \in [d]$ and for all $t \geqslant 0$, $N_j(t) \geqslant_{\mathrm{st}} \tilde{N}_j(t)$, $Q_j(t) \geqslant_{\mathrm{st}} \tilde{Q}_j(t)$ and $\Lambda_j(t) \geqslant_{\mathrm{st}} \tilde{\Lambda}_j(t)$.*

Proof. We first prove the result for univariate processes, after which we extend the arguments to multivariate processes. The proof relies on the cluster representation as given in Definition 2 with excitation functions specified in Eqn. 5.

Univariate case. Consider the conditional intensity processes $\Lambda(\cdot)$ and $\tilde{\Lambda}(\cdot)$. For Hawkes, define the generation-k arrival intensity by

$$\Lambda_{k\mathrm{G}}(t)) = \begin{cases} \lambda_0, & \text{if } k = 0, \\ \displaystyle\sum_{\substack{(t_i,B_i,w_i)\in N \text{ of generation } k-1 \\ t_i<t}} B_i h(t-t_i), & \text{if } k \in \mathbb{N}. \end{cases}$$

Similarly, for delayed Hawkes we define the generation-k arrival intensity by

$$\tilde{\Lambda}_{k\mathrm{G}}(t)) = \begin{cases} \lambda_0, & \text{if } k = 0, \\ \displaystyle\sum_{\substack{(t_i,B_i,w_i)\in \tilde{N} \text{ of generation } k-1 \\ t_i+w_i<t}} B_i h(t-t_i-w_i), & \text{if } k \in \mathbb{N}. \end{cases}$$

Note that $\Lambda_{0\mathrm{G}}(t) = \lambda_0 = \tilde{\Lambda}_{0\mathrm{G}}(t)$, i.e., immigrants arrive at the same intensity for both Hawkes and delayed Hawkes.

We say that a cluster *starts* when the excitation becomes active; for Hawkes this is at the birth of a particle, for delayed Hawkes at expiration of a particle. Note that for the delayed Hawkes process, at time t, excitation caused by immigration starts at rate $\int_0^t \lambda_0 \,\mathrm{d}\bar{\mathcal{J}}(s) = \lambda_0 \bar{\mathcal{J}}(t) \leqslant \lambda_0$. Let $(\Omega, \mathcal{F}, \mathbb{Q}) = (\mathbb{R}_+^2, \mathcal{F}_B \otimes \mathcal{F}_J, \mathcal{Q} \otimes \bar{\mathcal{J}})$ be the probability space on which the marks B and lifetimes J are defined jointly.

For $k \geqslant 0$, define the k-th cluster of a delayed Hawkes process $\tilde{N}$ recursively w.r.t. i.i.d. Poisson random measures $(\tilde{M}_k)_{k\in\mathbb{N}_0}$ on $\mathbb{R}\times\mathbb{R}_+\times\Omega$ with intensity $\mathrm{d}t \times \mathrm{d}s \times \big(\mathcal{Q}(\mathrm{d}z) \otimes \mathrm{d}\bar{\mathcal{J}}(w)\big)$ by

$$\tilde{N}_{k\mathrm{G}}(A\times B) = \int_{A\times\mathbb{R}_+\times B} \mathbf{1}_{[0,\tilde{\Lambda}_{k\mathrm{G}}(t)]}(s)\, \tilde{M}_k(\mathrm{d}t\times\mathrm{d}s\times\mathrm{d}(z,w)), \quad A\times B \in \mathcal{B}(\mathbb{R})\otimes\mathcal{F}, \tag{32}$$

where $\mathrm{d}(z,w)$ denotes integration with respect to a mark z and a lifetime w. Let $(M_k)_{k\in\mathbb{N}_0}$ and $(M'_k)_{k\in\mathbb{N}_0}$ be independent copies of the sequence of Poisson random measures $(\tilde{M}_k)_{k\in\mathbb{N}_0}$.

We inductively construct a coupling between the Hawkes and delayed Hawkes clusters. We construct the Hawkes immigration cluster $N_{0\mathrm{G}}$ by shifting immigration particles of delayed Hawkes by their lifetimes and adding an independent residual part, so that the increase in first-generation arrival intensity resulting from the coupled particles coincides. More specifically, we set

$$\begin{aligned} N_{0\mathrm{G}}(A\times B) &:= \int_{A\times B} \mathbf{1}_A(u+w)\ \tilde{N}_{0\mathrm{G}}(\mathrm{d}u\times \mathrm{d}(z\times w)) \\ &\quad + \int_{A\times\mathbb{R}_+\times B} \mathbf{1}_{[0,(1-\bar{\mathcal{F}}(t))\tilde{\Lambda}_{0\mathrm{G}}(t)]}(s)\ M_0(\mathrm{d}t\times \mathrm{d}s\times \mathrm{d}(z,w)) \\ &=: N_{0\mathrm{G},1}(A\times B) + N_{0\mathrm{G},2}(A\times B), \quad A\times B\in\mathcal{B}(\mathbb{R})\otimes\mathcal{F}, \end{aligned} \tag{33}$$

with $\bar{\mathcal{F}}(\cdot)$ the CDF of the i.i.d. lifetimes J_i. Because of the shift, the first term $N_{0\mathrm{G},1}$ is a time-inhomogeneous Poisson process of intensity $\bar{\mathcal{F}}(t)\lambda_0$, which creates an increase in first-generation intensity for the Hawkes process matching the first-generation intensity for the delayed Hawkes process. Since $N_{0\mathrm{G},2}$ is non-negative, the 'above-baseline' intensity caused by immigrant arrivals of the Hawkes process dominates that of the delayed Hawkes process along this sample path:

$$\Lambda_{1\mathrm{G}}(t) \geqslant \tilde{\Lambda}_{1\mathrm{G}}(t), \quad t\geqslant 0. \tag{34}$$

For $k\geqslant 1$, we refine this coupling inductively. Suppose that $N_{(k-1)G}$ has been decomposed into $N_{(k-1)G,1}$ and $N_{(k-1)G,2}$, such that the increase in k-th generation intensity for Hawkes resulting from $N_{(k-1)\mathrm{G},1}$ exactly matches the k-th generation intensity $\tilde{\Lambda}_{kG}$ for delayed Hawkes. Hence, offspring originating from $N_{(k-1)\mathrm{G},1}$ may be constructed from the k-th generation $\tilde{N}_{kG}$ of delayed Hawkes. We denote the intensity created by the remainder stream by

$$\Lambda_{k\mathrm{G},2}(t) := \sum_{\substack{(t_i,B_i,w_i)\in N_{(k-1)\mathrm{G},2}\\ t_i<t}} B_i h(t-t_i). \tag{35}$$

We define the k-th generation cluster of the Hawkes process through right-shifting the corresponding cluster for delayed Hawkes by the particles' lifetimes (which creates a matched increase in $(k+1)$-th generation arrival intensity for Hawkes and delayed Hawkes), plus remainder terms. More specifically, we set

$$\begin{aligned} N_{k\mathrm{G}}(A\times B) &= \int_{A\times B} \mathbf{1}_A(u+w)\ \tilde{N}_{k\mathrm{G}}(\mathrm{d}u\times \mathrm{d}(z\times w)) \\ &\quad + \int_{A\times\mathbb{R}_+\times B} \mathbf{1}_{[0,\tilde{\Lambda}_{k\mathrm{G}}(t)-\int_0^t \tilde{\Lambda}_{k\mathrm{G}}(t-w)\ \bar{\mathcal{F}}(\mathrm{d}w)]}(s)\ M_k(\mathrm{d}t\times \mathrm{d}s\times \mathrm{d}(z,w)) \\ &\quad + \int_{A\times\mathbb{R}_+\times B} \mathbf{1}_{[0,\Lambda_{k\mathrm{G},2}(t)]}(s)\ M_k'(\mathrm{d}t\times \mathrm{d}s\times \mathrm{d}(z,w)) \\ &=: N_{k\mathrm{G},1}(A\times B) + N_{k\mathrm{G},2}(A\times B), \quad A\times B\in\mathcal{B}(\mathbb{R})\otimes\mathcal{F}, \end{aligned} \tag{36}$$

where $N_{k\mathrm{G},1}$ is defined as the first integral from (36), whereas $N_{k\mathrm{G},2}$ is defined as the sum of the second and third terms. Due to the right-shifting of delayed Hawkes particles by their lifetimes, the increase in $(k+1)$-th generation intensity for Hawkes resulting from $N_{k\mathrm{G},1}$ exactly matches the $(k+1)$-th generation intensity $\tilde{\Lambda}_{(k+1)G}$ for delayed Hawkes. Since $N_{k\mathrm{G},2}$ is a non-negative stream of particles increasing the arrival intensity for $(k+1)$-th generation Hawkes particles, it follows

that along the sample path induced by this coupling,

$$\Lambda_{(k+1)\mathrm{G}}(t) = \tilde{\Lambda}_{(k+1)\mathrm{G}}(t) + \Lambda_{(k+1)\mathrm{G},2}(t) \geqslant \tilde{\Lambda}_{(k+1)\mathrm{G}}(t), \quad t \geqslant 0. \tag{37}$$

Repeating the argument of the previous paragraph inductively, we obtain a single coupling such that along this sample path, $\Lambda_{k\mathrm{G}}(t) \geqslant \tilde{\Lambda}_{k\mathrm{G}}(t), t \geqslant 0$, for all $k \in \mathbb{N}$. Hence, it follows that $\Lambda_{k\mathrm{G}}(t) \geqslant_{\mathrm{st}} \tilde{\Lambda}_{k\mathrm{G}}(t), t \geqslant 0$, for all $k \in \mathbb{N}$. Furthermore, we also deduce from this coupling that

$$\sum_{k=0}^{n} \Lambda_{k\mathrm{G}}(t) \geqslant_{\mathrm{st}} \sum_{k=0}^{n} \tilde{\Lambda}_{k\mathrm{G}}(t), \quad t \geqslant 0, \quad \text{for all } n \geqslant 0. \tag{38}$$

In any step, the stochastic ordering for the k-th generation clusters starts from the comparison of the Hawkes cluster with the (pathwise) right-shifted delayed Hawkes cluster. Hence, it implicitly uses the *genealogical order*.

Note that $\sum_{k=0}^{n} \Lambda_{k\mathrm{G}}(t) \xrightarrow{\mathcal{D}} \Lambda(t)$ as $n \to \infty$, for all $t \geqslant 0$, and similarly for $\tilde{\Lambda}(t)$. Hence, for every continuity point x of $F_{\Lambda(t)}$,

$$\lim_{n\to\infty} F_{\sum_{k=0}^{n} \Lambda_{k\mathrm{G}}(t)}(x) = F_{\Lambda(t)}(x),$$

and similarly for every continuity point x of $F_{\tilde{\Lambda}(t)}$,

$$F_{\sum_{k=0}^{n} \tilde{\Lambda}_{k\mathrm{G}}(t)}(x) \to F_{\tilde{\Lambda}(t)}(x).$$

Since any distribution function has at most countably many discontinuities, and using the fact that $F_{\sum_{k=0}^{n} \Lambda_{k\mathrm{G}}(t)}(x) \leqslant F_{\sum_{k=0}^{n} \tilde{\Lambda}_{k\mathrm{G}}(t)}(x)$ for all $x \in \mathbb{R}$ and $n \geqslant 0$ by the stochastic ordering we established above, it immediately follows that $F_{\Lambda(t)}(x) \leqslant F_{\tilde{\Lambda}(t)}(x)$ for all but at most countably many x. By right-continuity of distribution functions, if this inequality does not hold at z, it does not hold for a continuum of values $x \in [z, z + \epsilon]$. Hence, this inequality holds for all $x \in \mathbb{R}$, and we conclude that $\Lambda(t) \geqslant_{\mathrm{st}} \tilde{\Lambda}(t)$.

From this, we can decompose the conditional intensity $\Lambda(\cdot)$ of a Hawkes process as the sum of the intensity $\tilde{\Lambda}(\cdot)$ of a delayed Hawkes process with the same parameters, and the non-negative process $(\Lambda - \tilde{\Lambda})(\cdot)$ consisting of the remaining intensity. By coupling arrivals and setting sojourn times equal, it follows that $Q(t) \geqslant_{\mathrm{st}} \tilde{Q}(t)$ and $N(t) \geqslant_{\mathrm{st}} \tilde{N}(t)$, as claimed.

Multivariate case. Suppose that an immigrant in coordinate i_0 produces offspring in coordinate i_1, which in turn produces offspring in coordinate i_2, and so on, until there is a child in coordinate i_n. Write $i_0 i_1 \dots i_n$ for the path indicating this order of visited coordinates. By analogy to the univariate case, let $\Lambda_{i_0 i_1 \dots i_n}$ and $\tilde{\Lambda}_{i_0 i_1 \dots i_n}$ be the arrival intensities of n-th generation offspring in coordinate i_n through the order $i_0 i_1 \dots i_n$ for the Hawkes and the delayed Hawkes process, respectively. As in the univariate case, it can be argued that $\Lambda_{i_0 i_1 \dots i_n}(t) \geqslant_{\mathrm{st}} \tilde{\Lambda}_{i_0 i_1 \dots i_n}(t)$ for each such path $i_0 i_1 \dots i_n$, By using the conditional independency structure inherent in the cluster representation and by summing over all possible paths $i_0 i_1 \dots i_n$, $n \in \mathbb{N}_0$, $i_k \in [d]$, it follows that for each $j \in [d]$, $\Lambda_j(t) \geqslant_{\mathrm{st}} \tilde{\Lambda}_j(t)$. By coupling arrivals and setting sojourn times equal, the other claims follow. □

We conclude this section by considering two univariate delayed Hawkes birth-death processes having different parameters that dominate each other, and indicate when one process dominates the other. Denote those delayed Hawkes birth-death processes by $(N^{(j)}(\cdot), Q^{(j)}(\cdot), \Lambda^{(j)}(\cdot))$, $j = 1, 2$,

in which we have arrivals generated by conditional intensities of the form

$$\Lambda^{(j)}(t) = \lambda_0^{(j)} + \sum_{t_i^{(j)} < t} B_i^{(j)} h^{(j)}(t - t_i^{(j)}), \tag{39}$$

where $(t_i^{(j)})$ denote departure times from system j, and where $B_i^{(j)} \overset{\text{iid}}{\sim} B^{(j)}$. For system j, we have i.i.d. departures distributed as $J^{(j)}$. If the baseline intensity, mark distribution or excitation function of system 1 dominates that of system 2, or if the sojourn time of system 2 dominates that of system 1, we would expect system 1 to stochastically dominate system 2. Those conjectures are confirmed by the next theorem.

Theorem 7. *Suppose that system* 1 *and system* 2 *satisfy the following conditions:*

(i) $\lambda_0^{(1)} \geqslant \lambda_0^{(2)}$;
(ii) $B^{(1)} \geqslant_{\text{st}} B^{(2)}$;
(iii) $h^{(1)}(t) \geqslant h^{(2)}(t)$ *for almost all* t;
(iv) $J^{(1)} \leqslant_{\text{st}} J^{(2)}$.

Then $N^{(1)}(t) \geqslant_{\text{st}} N^{(2)}(t)$, $Q^{(1)}(t) \geqslant_{\text{st}} Q^{(2)}(t)$ *and* $\Lambda^{(1)}(t) \geqslant_{\text{st}} \Lambda^{(2)}(t)$ *for all* $t \geqslant 0$.

Proof. It suffices to consider the case where just one of the conditions (i)-(iv) holds strictly. For example, if (i) and (ii) hold strictly, we select an intermediate process $N^{(3)}$ with $\lambda_0^{(3)} = \lambda_0^{(2)}$ and $B^{(3)} = B^{(1)}$, and use our arguments to arrive at $N^{(1)}(t) \geqslant_{\text{st}} N^{(3)}(t) \geqslant_{\text{st}} N^{(2)}(t)$.

For (i)–(iii), the proof is straightforward: it uses the cluster representation, and relies on an easy coupling argument, by partly coupling the parameter of system 1 to the corresponding one of system 2, with the remaining part generating a positive stream. For (iv), we argue as in the univariate case of the proof of Theorem 6. □

In the next corollary to Theorem 6, we describe the steady-state delayed Hawkes birth-death process $(\tilde{Q}(\infty), \tilde{\Lambda}(\infty))$; its proof is in Appendix E.

Corollary 2. *Using the notation of Theorem 6, in steady state we have* $Q(\infty) \overset{\mathcal{D}}{=} \tilde{Q}(\infty)$ *and* $\Lambda(\infty) \overset{\mathcal{D}}{=} \tilde{\Lambda}(\infty)$.

In stationarity, the distribution of population sizes and intensities *at a fixed time instant* are the same for Hawkes and delayed Hawkes. It should be borne in mind, however, that the dynamics of the two processes *are* different in stationarity, since an arrival does not increase the intensity instantaneously for delayed Hawkes.

Corollary 2 has an appealing *informal* explanation. In stationarity, the stream of particles entering and leaving a Hawkes-fed birth-death process are 'in equilibrium'. Hence, starting in the stationary distribution of the Hawkes-fed birth-death process, excitation caused by arriving particles (as we have for Hawkes) equals excitation caused by departing particles (as we have for delayed Hawkes). For the Hawkes process, under the stationary distribution, the inward stream in intensity (caused by excitation) equals the outward stream (caused by decay). Hence, if the delayed Hawkes process starts in the stationary distribution of Hawkes, increase in intensity caused by departures (equals increase in intensity that we would see for Hawkes) equals the decrease caused by decay. This indicates that this distribution is also stationary for delayed Hawkes.

Corollary 2 allows us to describe heavy-traffic behavior for the *Markovian* delayed Hawkes birth-death process, assuming marks having finite second moments; cf. Corollary 1 and see Appendix E for the proof.

Corollary 3. *Consider a delayed Hawkes birth-death process as in Corollary 2, now assumed to be Markovian. Suppose that* $b_2 := \mathbb{E}[B^2] < \infty$. *Then we have, as* $\rho = b_1/r \uparrow 1$,

$$(1-\rho)\tilde{\Lambda}(\infty) \xrightarrow{\mathcal{D}} \Gamma\left(\frac{2r\lambda_0}{b_2}, \frac{2r}{b_2}\right), \quad (1-\rho)\tilde{Q}(\infty) \xrightarrow{\mathcal{D}} \Gamma\left(\frac{2r\lambda_0}{b_2}, \frac{2r\mu}{b_2}\right).$$

7. Networks of Markovian delayed Hawkes birth-death processes

Next, we specify to (networks of) the Markovian delayed Hawkes process, which allows us to set up a more concrete characterization of the transform than the one found in Section 5, and to formulate a recursive procedure for calculating the joint moments of $(\boldsymbol{Q}(t), \boldsymbol{\Lambda}(t))$; see Section 7.1. In the univariate case, this leads to a system of ODEs involving a Clement-Kac-Sylvester matrix, which can be solved explicitly; see Section 7.2.

7.1. Networks of birth-death processes.

Networks of birth-death processes with shot-noise driven arrival rates have been studied in [33]. Although networks of Hawkes processes have been introduced in [21], to the best of our knowledge, there is no account in the literature of the exact transient behavior of such processes. In this subsection, we analyze transient behavior for a network of Markovian delayed Hawkes birth-death processes. After obvious modifications, this analysis can be adapted to networks of (classical) Hawkes-fed birth-death processes. Furthermore, by setting $\mu_{ij} = 0$ for all $i, j \in [d]$, see Definition 3, our analysis applies to the multivariate delayed Hawkes (point) process as well.

We first characterize the distribution of the Markovian network process from Definition 3 by deriving a PDE for the joint Z- and Laplace transform of $(\boldsymbol{Q}(\cdot), \boldsymbol{\Lambda}(\cdot))$, given by

$$\zeta(t, \boldsymbol{z}, \boldsymbol{s}) = \mathbb{E}\left[\boldsymbol{z}^{\boldsymbol{Q}(t)} e^{-\boldsymbol{s}^\top \boldsymbol{\Lambda}(t)}\right] = \mathbb{E}\left[\prod_{j=1}^{d} z_j^{Q_j(t)} e^{-s_j \Lambda_j(t)}\right], \tag{40}$$

where $\boldsymbol{z} \in [-1,1]^d, \boldsymbol{s} \in \mathbb{R}_+^d$. For $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{n} \in \mathbb{N}_0^d$, write $\boldsymbol{x}^{\boldsymbol{n}} = \prod_{j=1}^d x_j^{n_j}$. By analogy to the univariate, nondelayed case, see [34], we derive a PDE for ζ, to which we apply the method of characteristics to reduce it to a system of ODEs. Furthermore, this PDE can be used to derive a system of ODEs for the joint moments. The proofs of the following two results can be found in Appendix E.

Theorem 8. *For all* $i, j \in [d]$, *assume that* $h_{ij}(t) = e^{-r_i t}$, *where* $r_i > 0$, *and assume that* $B_{ij} > 0$ *a.s. Consider the (now Markovian) network of delayed Hawkes birth-death processes from Definition 3. Then the multivariate joint Z- and Laplace transform* $\zeta(t, \boldsymbol{z}, \boldsymbol{s})$ *satisfies the following PDE:*

$$\frac{\partial \zeta(t,\boldsymbol{z},\boldsymbol{s})}{\partial t} + \sum_{j=1}^{d} (r_j s_j + z_j - 1) \frac{\partial \zeta(t,\boldsymbol{z},\boldsymbol{s})}{\partial s_j} + \sum_{j=1}^{d} \mu_j (z_j - \beta_j(\boldsymbol{s})) \frac{\partial \zeta(t,\boldsymbol{z},\boldsymbol{s})}{\partial z_j}$$
$$+ \sum_{j=1}^{d} \sum_{i=1}^{d} \mu_{ij} (z_j - z_i) \frac{\partial \zeta(t,\boldsymbol{z},\boldsymbol{s})}{\partial z_j} = -\zeta(t,\boldsymbol{z},\boldsymbol{s}) \sum_{j=1}^{d} r_j \lambda_{j,0} s_j, \tag{41}$$

where $\beta_j(\boldsymbol{s}) = \mathbb{E}[e^{-\boldsymbol{s}^\top \boldsymbol{B}_j}]$ is the multivariate Laplace transform of $\boldsymbol{B}_j$.

Furthermore, given initial conditions $\boldsymbol{Q}(0) = \boldsymbol{0}$ and $\boldsymbol{\Lambda}(0) = \boldsymbol{\lambda}_0$, we have

$$\zeta(t, \boldsymbol{z}, \boldsymbol{s}) = \prod_{j=1}^{d} \exp\left(-\lambda_{j,0}\left(s_j(t) + r_j \int_0^t s_j(u)\,\mathrm{d}u\right)\right), \tag{42}$$

where $s_j(\cdot)$, $j \in [d]$, solve the system of ODEs

$$\begin{aligned} s_j'(u) &= -r_j s_j(u) - z_j(u) + 1; \\ z_j'(u) &= \mu_j(\beta_j(s(u)) - z_j(u)) + \sum_{i=1}^{d} \mu_{ij}(z_i(u) - z_j(u)), \quad 0 \leqslant u \leqslant t, \end{aligned} \tag{43}$$

with boundary conditions $s_j(0) = s_j$ and $z_j(0) = z_j$.

Theorem 9. *For $q, Q \in \mathbb{N}_0$, let $\bar{Q}^q := Q(Q-1)(Q-q+1)$ be the falling factorial, with $\bar{Q}^0 := 1$ and $\bar{Q}^{-1} := 0$. Write $b_{kj} = \mathbb{E}[B_{kj}]$. Next, for $\boldsymbol{g}, \boldsymbol{\ell} \in \mathbb{N}_0^d$, write*

$$\binom{\boldsymbol{g}}{\boldsymbol{\ell}} := \prod_{j=1}^{d} \binom{g_j}{l_j}. \tag{44}$$

Furthermore, for $\boldsymbol{q} \in \mathbb{N}_0^d, \boldsymbol{Q} \in \mathbb{N}_0^d$, write $\bar{\boldsymbol{Q}}^{\boldsymbol{q}} := \prod_{j=1}^{d} \bar{Q}_j^{q_j}$. Let $\circ$ be the Hadamard product. Then we have the following differential equation for the joint moments of $\boldsymbol{Q}(t), \boldsymbol{\Lambda}(t)$:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}}(t)\right] + \|\boldsymbol{g} \circ \boldsymbol{r} + \boldsymbol{q} \circ \boldsymbol{\mu}\|_1 \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}}(t)\right] - \sum_{j=1}^{d} q_j \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}-\boldsymbol{e}_j}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}+\boldsymbol{e}_j}(t)\right] \\ &- \sum_{j=1}^{d}\sum_{k=1}^{d} \mu_j g_k b_{kj} \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}+\boldsymbol{e}_j}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}-\boldsymbol{e}_k}(t)\right] \\ &+ \sum_{j=1}^{d}\sum_{i=1}^{d} \mu_{ij}\left(q_j \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}}(t)\right] - q_i \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}+\boldsymbol{e}_j-\boldsymbol{e}_i}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}}(t)\right]\right) \\ &= \sum_{j=1}^{d} g_j r_j \lambda_{j,0} \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}}(t)\boldsymbol{\Lambda}^{\boldsymbol{g}-\boldsymbol{e}_j}(t)\right] + \sum_{j=1}^{d} \mu_j \sum_{\substack{\boldsymbol{0} \leqslant \boldsymbol{\ell} \leqslant \boldsymbol{g} \\ \|\boldsymbol{\ell}\|_1 \leqslant \|\boldsymbol{g}\|_1 - 2}} \binom{\boldsymbol{g}}{\boldsymbol{\ell}} \mathbb{E}\left[\boldsymbol{B}_j^{\boldsymbol{g}-\boldsymbol{\ell}}\right] \mathbb{E}\left[\bar{\boldsymbol{Q}}^{\boldsymbol{q}+\boldsymbol{e}_j}(t)\boldsymbol{\Lambda}^{\boldsymbol{\ell}}(t)\right]. \end{aligned} \tag{45}$$

Eqn. (45) allows us to devise a recursive procedure to find the joint moments of arbitrary order. Indeed, the left-hand side of (45) expresses a joint moment of order $n = \|(\boldsymbol{q}, \boldsymbol{g})\|_1$ as a linear ODE dependent on joint moments of equal order, whereas the right-hand side contains a forcing term, consisting of lower-order moments only. In general, we can find the $(n+1)$-th order moments by solving a linear system of ODEs with forcing constant dependent on the moments of order up to n. Since the system for the first-order moments does not contain unknown quantities, this provides us with a recursive procedure for expressing the moments of a network of delayed Hawkes birth-death processes in the moments of the mark random variables, in the exponential decay rates r_i, and in the departure and rerouting rates μ_i, μ_{ij}.

Remark 7. *To find the moments of order n, we need to solve a system of ODEs of dimension $\binom{n+2d-1}{n}$. The ODEs are found by substituting all possible $(\boldsymbol{q}, \boldsymbol{g})$ into* (45) *satisfying $\|(\boldsymbol{q}, \boldsymbol{g})\|_1 = n$.*

7.2. **Transient behavior of the univariate delayed Hawkes birth-death process.** We now specify to the univariate case with $h(t) = e^{-rt}$, since in this setting we can be more specific about the moments of $(Q(t), \Lambda(t))$. Specifying (45) to the univariate case $d = 1$, we obtain the following ODE:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\bar{Q}^q(t)\Lambda^g(t)\right] + (gr + q\mu)\mathbb{E}\left[\bar{Q}^q(t)\Lambda^g(t)\right] - q\mathbb{E}\left[\bar{Q}^{q-1}(t)\Lambda^{g+1}(t)\right]\\ &= \mathbf{1}\{g \geqslant 1\}gr\lambda_0\mathbb{E}\left[\bar{Q}^q(t)\Lambda^{g-1}(t)\right] + \mathbf{1}\{g \geqslant 1\}\mu\sum_{j=0}^{g-1}\binom{g}{j}\mathbb{E}\left[B^{g-j}\right]\mathbb{E}\left[\bar{Q}^{q+1}(t)\Lambda^j(t)\right]. \qquad (46)\end{aligned}$$

We wish to derive a system of ODEs for the joint moments of order $n \in \mathbb{N}$, which we accomplish by taking a combination of indices $g = k$, $q = n - k$, $k \in \{0, 1, \ldots, n\}$, for which (46) reads

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\bar{Q}^{n-k}(t)\Lambda^k(t)\right] + (kr + (n-k)\mu)\mathbb{E}\left[\bar{Q}^{n-k}(t)\Lambda^k(t)\right] - (n-k)\mathbb{E}\left[\bar{Q}^{n-k-1}(t)\Lambda^{k+1}(t)\right]\\ &- \mu k b_1\mathbb{E}\left[\bar{Q}^{n-k+1}(t)\Lambda^{k-1}(t)\right]\\ &= \mathbf{1}\{k \geqslant 1\}kr\lambda_0\mathbb{E}\left[\bar{Q}^{n-k}(t)\Lambda^{k-1}(t)\right] + \mathbf{1}\{k \geqslant 2\}\mu\sum_{j=0}^{k-2}\binom{k}{j}\mathbb{E}\left[B^{k-j}\right]\mathbb{E}\left[\bar{Q}^{n-k+1}(t)\Lambda^j(t)\right], \qquad (47)\end{aligned}$$

where $b_1 = \mathbb{E}[B]$. Letting

$$Z^{(n+1)}(t) := \begin{bmatrix}\mathbb{E}\left[\bar{Q}^n(t)\right] & \mathbb{E}\left[\bar{Q}^{n-1}(t)\Lambda(t)\right] & \cdots & \mathbb{E}\left[\bar{Q}^1(t)\Lambda^{n-1}(t)\right] & \mathbb{E}\left[\Lambda^n(t)\right]\end{bmatrix}^\top,$$

$$A^{(n+1)} = \begin{bmatrix}-a_0^{(n)} & n & 0 & \cdots & 0 & 0\\ \mu b_1 & -a_1^{(n-1)} & n-1 & \cdots & 0 & 0\\ 0 & 2\mu b_1 & -a_2^{(n-2)} & \ddots & 0 & 0\\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -a_{n-1}^{(1)} & 1\\ 0 & 0 & 0 & \cdots & n\mu b_1 & -a_n^{(0)}\end{bmatrix}, \quad C^{(n+1)}(t) = \begin{bmatrix}c_0^{(n)}(t)\\ c_1^{(n-1)}(t)\\ c_2^{(n-2)}(t)\\ \vdots\\ c_{n-1}^{(1)}(t)\\ c_n^{(0)}(t)\end{bmatrix},$$

where $a_k^{(n-k)} = kr + (n-k)\mu = n\mu + k(r-\mu)$ and

$$c_k^{(n-k)}(t) = \mathbf{1}\{k \geqslant 1\}kr\lambda_0\mathbb{E}\left[\bar{Q}^{n-k}(t)\Lambda^{k-1}(t)\right] + \mathbf{1}\{k \geqslant 2\}\mu\sum_{j=0}^{k-2}\binom{k}{j}\mathbb{E}\left[B^{k-j}\right]\mathbb{E}\left[\bar{Q}^{n-k+1}(t)\Lambda^j(t)\right]$$

it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}Z^{(n+1)}(t) = A^{(n+1)}Z^{(n+1)}(t) + C^{(n+1)}(t). \qquad (48)$$

Note that $A^{(n+1)}$ is a *generalized Clement-Kac-Sylvester matrix*. To solve this ODE, we need $C^{(n+1)}(t)$, which is a vector dependent on moments of order at most $n-1$, meaning that we can solve for the transient moments of the delayed Hawkes birth-death process recursively. The proofs of the next two results are in Appendix E.

Theorem 10. *The solution to the ODE (48) is*

$$Z^{(n+1)}(t) = e^{A^{(n+1)}t}Z^{(n+1)}(0) + \int_0^t e^{A^{(n+1)}(t-s)}C^{(n+1)}(s)\,\mathrm{d}s, \qquad (49)$$

where $Z^{(n+1)}(0) = \lambda_0^n \boldsymbol{e}_{n+1}$*, with* $\boldsymbol{e}_{n+1}$ *the last standard unit vector in* $\mathbb{R}^{n+1}$*. The matrix exponential* $e^{A^{(n+1)}t}$ *can be calculated explicitly by*

$$e^{A^{(n+1)}t} = \sum_{k=0}^{n} e^{\lambda_k^{(n+1)}t} \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{A^{(n+1)} - \lambda_j^{(n+1)} I_{n+1}}{\lambda_k^{(n+1)} - \lambda_j^{(n+1)}}, \tag{50}$$

where I_{n+1} *is the* $(n+1)\times(n+1)$ *identity matrix and where*

$$\lambda_k^{(n+1)} = -\frac{n}{2}(\mu + r) + \frac{n-2k}{2}\sqrt{(\mu - r)^2 + 4\mu b_1}, \quad k = 0, 1, \ldots, n, \tag{51}$$

are the eigenvalues of $A^{(n+1)}$*. This implies that we have a stable system — i.e., with* $Z^{(n+1)}(t)$ *converging, as* $t \to \infty$ *— if and only if the stability condition* $b_1/r < 1$ *holds.*

Remark 8. *The moment recursion obtained in Theorem 10 scales combinatorially with the order of the joint moments and the dimension of the system. Specifically, for a given moment order n, the size of the system of ODEs grows like* $\binom{n+d-1}{d-1}$ *with the dimension d, reflecting all possible combinations of powers of* Q_i *and* Λ_i*. The recursion is fully tractable for* either *low-dimensional systems* or *low-order moments, whereas it becomes computationally prohibitive for large dimensions combined with high-order moments, due to both the rapidly growing number of ODEs and the complexity of computing the matrix exponentials.*

We are able to find the first-order moments in the stationary regime, by letting $t \to \infty$.

Theorem 11. *Let* $b_1 := \mathbb{E}[B]$*. If the stability condition* $b_1/r < 1$ *holds, then, as* $t \to \infty$*,*

$$\begin{bmatrix} \mathbb{E}[Q(t)] \\ \mathbb{E}[\Lambda(t)] \end{bmatrix} \to \frac{r\lambda_0}{r - b_1} \begin{bmatrix} 1/\mu \\ 1 \end{bmatrix}. \tag{52}$$

Remark 9. *By combining Corollary 2 in the Markovian setting with* [34]*, Corollaries 3.8 and 3.9, one could also find* $\mathrm{Var}(\tilde{N}(\infty))$*,* $\mathrm{Cov}(\tilde{N}(\infty), \tilde{\Lambda}(\infty))$*, and* $\mathbb{E}[\tilde{\Lambda}^g(\infty)]$ *for any* $g \in \mathbb{N}$*.*

8. Discussion and concluding remarks

We have formally introduced the delayed Hawkes process, and a rich family of point processes having sojourn-time dependent excitation, containing Hawkes, delayed Hawkes and the ephemerally self-exciting process as special cases. The delayed Hawkes process arises naturally in applications and has turned out to be remarkably tractable, admitting a cluster process representation in the linear case enabling transform characterizations by a fixed-point equation and the analysis of heavy-tailed asymptotics. The effect of delays has been made visible in a scaling limit that is markedly different from its classical, non-delayed counterpart. Furthermore, using a method that one can describe as *genealogical coupling*, we have demonstrated that the delayed Hawkes birth-death process is stochastically dominated by a comparable Hawkes-fed birth-death process. In the Markovian case, we have provided a recursive procedure to calculate the moments of a network of delayed Hawkes birth-death processes explicitly.

In future research, several directions can be envisioned.

- In Theorem 3, we could only state our FCLTs on an interval bounded away from 0. One may want to study the (complex) behavior on an interval $[0, \epsilon]$ including 0 as well.

- As discussed in Remark 4, in the scaling limit for $\alpha \in (\frac{1}{2}, 1)$, it would be interesting to identify $\mathbb{E}[N_\alpha^T(v)]$. In the same remark, we saw that for $\alpha \in (\frac{1}{2}, 1)$, we still find a Brownian limit, whereas for $\alpha = 1$ and sojourn times taking values in the unit interval, the situation is more involved; in particular, one may ask whether it is reasonable to expect short-range dependence. If there is non-Gaussian behavior for $\alpha = 1$, one may want to look for a scaling limit in which one multiplies sojourn times by $T^{\alpha(T)}$, before one contracts time by a factor T; here, $\alpha(T) \to 1$ as $T \to \infty$. This setting bears some similarities with the one considered in [28], although in our case quantities unscaled by $1 - \alpha(T)$ do not diverge, but instead become smaller and, in some sense, 'collapse' to a Poisson process for $\alpha > 1$. A similar regime that may be interesting is the one where $\alpha = 1/2$, but where we have sojourn times $J_T = T \cdot J$, for some positive random variable J, so that (A3) is not satisfied in the limit $T \to \infty$.
- An interesting line of study concerns statistical inference for delayed Hawkes processes. A considerable amount of literature exists on this topic for classical Hawkes processes, but it is open to what extent these results extend to delayed Hawkes processes. In the Markovian case our closed-form expressions for the moments can be used to identify moment estimators, whereas the non-Markovian case is anticipated to be substantially more challenging. In this direction, the goodness-of-fit results reported in Appendix B are promising.
- It would be interesting to analyze the effect of delays on cluster durations for delayed Hawkes, following the recent results by Daw [16]. We have not succeeded in extending Daw's arguments to our setting.
- Our general family of models having sojourn-time dependent excitation encompasses the Hawkes, the delayed Hawkes, and the ephemerally self-exciting processes as special cases. It would be interesting to identify other relevant models belonging to this family.

Acknowledgements

We are very grateful to the Editor-in-Chief, the Editor and to two anonymous referees for their careful reading and valuable comments, which have substantially improved the clarity and quality of the paper.

References

[1] Y. Aït-Sahalia, J. A. Cacho-Diaz, and R. J. A. Laeven (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics* **117**, pp. 585–606.

[2] Y. Aït-Sahalia, R. J. A. Laeven, and L. Pelizzon (2014). Mutual excitation in Eurozone sovereign CDS. *Journal of Econometrics* **183**, pp. 151–167.

[3] J. Baars, S. U. Can, and R. J. A. Laeven (2025). Asymptotically distribution-free goodness-of-fit testing for point processes. Preprint. Available at `https://arxiv.org/abs/2503.24197v1`.

[4] J. Baars, R. J. A. Laeven, and M. Mandjes (2026). Online supplement to "Delayed Hawkes birth-death processes".

[5] E. Bacry, S. Delattre, M. Hoffmann, and J. F. Muzy (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications* **123**, pp. 2475–2499.

[6] E. Bacry and J. F. Muzy (2014). Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance* **14**, pp. 1147–1166.

[7] L. Bauwens and N. Hautsch (2009). Modelling financial high frequency data using point processes. In book: *Handbook of Financial Time Series*, pp. 953–979.

[8] N. H. Bingham, C. M. Goldie, and J. L. Teugels (1989). *Regular Variation*. Cambridge University Press **27**, Cambridge.

[9] V. C. Borkar and M. A. Salman (2016). The exact methods to compute the matrix exponential. *IOSR Journal of Mathematics* **12**, pp. 72–86.

[10] P. Brémaud and L. Massoulié (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability* **24**, pp. 1563–1588.

[11] P. Cattiaux, L. Colombani, and M. Costa (2022). Limit theorems for Hawkes processes including inhibition. *Stochastic Processes and their Applications* **149**, pp. 404–426.

[12] W. Chiang, X. Liu, and G. Mohler (2022). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International Journal of Forecasting* **38**, pp. 505–520.

[13] W. Chu (2010). Fibonacci polynomials and Sylvester determinant of tridiagonal matrix. *Applied Mathematics and Computation* **216**, pp. 1018–1023.

[14] L. R. Cui, A. G. Hawkes, and H. Yi (2020). An elementary derivation of moments of Hawkes processes. *Advances in Applied Probability* **52**, pp. 102–137.

[15] D. J. Daley and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes*, Vol I and II, 2nd ed. Springer-Verlag, New York.

[16] A. Daw (2023). Conditional uniformity and Hawkes processes. *Mathematics of Operations Research*, Articles in Advance.

[17] A. Daw and J. Pender (2022). An ephemerally self-exciting point process. *Advances in Applied Probability* **54**, pp. 340–403.

[18] A. Daw and J. Pender (2023). Matrix calculations for moments of Markov processes. *Advances in Applied Probability* **55**, pp. 126–150.

[19] A. Daw and J. Pender (2018). Queues driven by Hawkes processes. *Stochastic Systems* **8**, pp. 192–229.

[20] A. Dassios and H. Zhou (2011). A dynamic contagion process. *Advances in Applied Probability* **43**, pp. 814–846.

[21] S. Delattre, N. Fournier, and M. Hoffmann (2016). Hawkes processes on large networks. *The Annals of Applied Probability* **26**, pp. 216–261.

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

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

[24] R. Van der Hofstad and M. Keane (2008). An elementary proof of the hitting time theorem. *The American Mathematical Monthly* **115**, pp. 753–756.

[25] U. Horst and W. Xu (2021). Functional limit theorems for marked Hawkes point measures. *Stochastic Processes and their Applications* **134**, pp. 94–131.

[26] M. Ikefuji, R. J. A. Laeven, J. R. Magnus and Y. Yue (2022). Earthquake risk embedded in property prices: Evidence from five Japanese cities. *Journal of the American Statistical Association* **117**, pp. 82–93.

[27] V. Isham and M. Westcott (1979). A self-correcting point process. *Stochastic Processes and their Applications* **8**, pp. 335–347.

[28] T. Jaisson and M. Rosenbaum (2015). Limit theorems for nearly unstable Hawkes processes. *The Annals of Applied Probability* **25**, pp. 600–631.

[29] T. Jaisson and M. Rosenbaum (2016). Rough fractional diffusions as scaling limits of nearly unstable heavy tailed Hawkes processes. *The Annals of Applied Probability* **26**, pp. 2860–2882.

[30] R. Karim, R. J. A. Laeven, and M. Mandjes (2021). Exact and asymptotic analysis of general multivariate Hawkes processes and induced population processes. Preprint. Available at `https://arxiv.org/abs/2106.03560`.

[31] R. Karim, R. J. A. Laeven and M. Mandjes (2025). Compound multivariate Hawkes processes: Large deviations and rare event simulation. *Bernoulli* **31**, pp. 3113–3138.

[32] M. Kirchner (2017). An estimation procedure for the Hawkes process. *Quantitative Finance* **17**, pp. 571–595.
[33] D. T. Koops, O. J. Boxma, and M. Mandjes (2017). Networks of $\cdot/G/\infty$ server queues with shot-noise-driven arrival intensities. *Queueing Systems* **86**, pp. 301–325.
[34] D. T. Koops, M. Saxena, O. J. Boxma, and M. Mandjes (2018). Infinite-server queues with Hawkes input. *Journal of Applied Probability* **55**, pp. 920–943.
[35] J. C. Louzada Pinto, T. Chahed and E. Altman (2016). A framework for information dissemination in social networks using Hawkes processes. *Performance Evaluation* **103**, pp. 86–107.
[36] L. Massoulié (1998). Stability results for a general class of interacting point processes dynamics, and applications. *Stochastic Processes and their Applications* **75**, pp. 1–30.
[37] J. Neveu (1965). *Mathematical Foundations of the Calculus of Probability*, 1st ed. Holden-Day series in probability and statistics.
[38] Y. Ogata (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory* **27**, pp. 23–31.
[39] Y. Ogata (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* **83**, pp. 9–27.
[40] Y. Ogata and D. Vere-Jones (1984). Inference for earthquake model. *Stochastic Processes and their Applications* **17**, pp. 337–347.
[41] Y. Ogata and D. Vere-Jones (1984). On the moments of a self-correcting process. *Journal of Applied Probability* **21**, pp. 335–342.
[42] J. Olinde and M. Short (2020). A self-limiting Hawkes process: Interpretation, estimation, and use in crime modeling. *2020 IEEE International Conference on Big Data*.
[43] M. B. Raad, S. Ditlevsen, and E. Löcherbach (2020). Stability and mean-field limits of age dependent Hawkes processes. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* **56**, pp. 1958–1990.
[44] P. Reynaud-Bouret, R. Lambert, C. Tuleau-Malot, T. Bessaih, V. Rivoirard, Y. Bouret, and N. Leresche (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of Neuroscience Methods* **297**, pp. 9–21.
[45] M. Rizoiu, Y. Lee, S. Mishra, and L. Xie (2017). A tutorial on Hawkes processes for events in social media. In book: *Frontiers of Multimedia Research*, pp. 191–218.
[46] T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels (1999). *Stochastic Processes for Insurance and Finance*, John Wiley, New York.
[47] D. Vere-Jones (1978). Earthquake prediction - a statistician's view. *Journal of Physics of the Earth* **26**, pp. 129–146.
[48] L. Zhu (2013). Central limit theorem for nonlinear Hawkes processes. *Journal of Applied Probability* **50**, pp. 760–771.

Appendix A: Relegated proofs of Section 3

Proof of Theorem 1. We prove the theorem in the univariate case. From there, the multivariate result can be proved along the lines of [10], Theorem 7, taking the randomness of the excitation functions into account in the same fashion as we do in the univariate case. To avoid repetition, we exclude the proof.

The proof uses the idea of the Picard proof method for the existence of differential equations, and follows [5], Theorem 1. It is structured as follows. We can assume, w.l.o.g., that $L = 1$, by writing $\phi(\cdot) = \phi(L^{-1}L\,\cdot)$. First, we prove the 'existence' part. We take a bivariate Poisson process M marked with random functions. With the aid of Lemma 1, we use Picard iteration, starting from the empty process, to construct a stationary process N with finite mean intensity satisfying the desired dynamics. Second, we prove the 'uniqueness' part, by proving that *any* stationary process $\tilde{N}$ with finite mean intensity satisfying the desired dynamics also satisfies condition (ii) in the theorem.

This means that we have stability, from which we deduce $\tilde{N} \overset{\mathcal{D}}{=} N$. Third, we prove stability under condition (i). Next, under condition (ii), we can take expectations with respect to M in the proof of the stability part below, after which the proof is analogous to the one under condition (i); therefore it is omitted.

Existence. We construct the process N upon a basis being a product probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ of (i) the canonical space of bivariate point processes on $\mathbb{R} \times \mathbb{R}_+$, with a probability measure $\mathbb{P}_M$ such that the identity mapping is a bivariate Poisson process of unit rate, and (ii) $(\Omega, \mathcal{F}) = L^1(\mathbb{R}_+) \cap L^\infty(\mathbb{R}_+)$, with a probability measure $\mathbb{Q}$ denoting the distribution of the random functions h. Such a random function exists by Kolmogorov's extension theorem. We denote the resulting marked Poisson process on $\mathbb{R} \times \mathbb{R}_+ \times \Omega$ by M. Write $(\mathcal{A}_t)_{t\in\mathbb{R}}$ for the filtration induced by M, i.e., $\mathcal{A}_t = \sigma(S_t M_-)$. Write $\mathcal{P}(\mathcal{A}_t) := \bigvee_{s<t} \mathcal{A}_s$ for the corresponding predictable σ-algebra. As indicated in Section 3, we treat the first coordinate of M as time.

We say that a point process N is *compatible* w.r.t. the left-shift operator S_t if for all $t \in \mathbb{R}$, $\tilde{\omega} \in \mathcal{A}$, $S_t N(\tilde{\omega}) = N(S_t \tilde{\omega})$, where $S_t \tilde{\omega}$ means that time is shifted in the basis space, meaning that the first coordinate of M is shifted.

We approximate the desired process $(N(\cdot), \Lambda(\cdot))$ using Picard iteration. More specifically, we set $\Lambda_0 \equiv 0$, and for $n \in \mathbb{N}_0$,

$$\begin{aligned} N_n(A \times B) &= \int_{A\times\mathbb{R}_+\times B} \mathbf{1}_{[0,\Lambda_n(\tau)]}(s) M(\mathrm{d}\tau \times \mathrm{d}s \times \mathrm{d}\omega), && A \times B \in \mathcal{B}(\mathbb{R}) \otimes \mathcal{F}, \\ \Lambda_{n+1}(t) &= \phi\left(\int_{(-\infty,t)\times\Omega} h(t-\tau,\omega) N_n(\mathrm{d}\tau \times \mathrm{d}\omega) \right), && t \in \mathbb{R}. \end{aligned} \tag{53}$$

By induction, for every $n \in \mathbb{N}_0$, N_n is adapted to $(\mathcal{A}_t)_{t\in\mathbb{R}}$, while Λ_n is adapted to $(\mathcal{P}(\mathcal{A}_t))_{t\in\mathbb{R}}$. Note that if $\phi(0) = 0$, the zero solution is stationary; we typically work with functions such that $\phi(0) > 0$. By construction, the processes (N_n), (Λ_n) are S_t-compatible and increasing in n. Since the basis space on which the process is constructed is time-invariant, it follows that (N_n), (Λ_n) are stationary. Since ϕ is Lipschitz, for $n \geqslant 1$ it holds that

$$\mathbb{E}|\Lambda_{n+1}(0) - \Lambda_n(0)| \leqslant \mathbb{E} \int_{(-\infty,0)\times\Omega} |h(-\tau,\omega)| \, (N_n - N_{n-1})(\mathrm{d}\tau \times \mathrm{d}\omega),$$

where the first coordinate of $N_n - N_{n-1}$ counts the number of points between $t \mapsto \Lambda_n(t)$ and $t \mapsto \Lambda_{n-1}(t)$. By Lemma 1, this process has $\Lambda_n - \Lambda_{n-1}$ as an $\mathcal{H}_t^M$-intensity. Hence,

$$\begin{aligned} \mathbb{E}[\Lambda_{n+1}(0) - \Lambda_n(0)] &\leqslant \mathbb{E} \int_{(-\infty,0)\times\Omega} |h(-\tau,\omega)| \, (N_n - N_{n-1})(\mathrm{d}\tau \times \mathrm{d}\omega) \\ &= \int_{(-\infty,0)\times\Omega} |h(-\tau,\omega)| \, \mathbb{E}(N_n - N_{n-1})(\mathrm{d}\tau \times \mathrm{d}\omega) \\ &= \int_{(-\infty,0)\times\Omega} |h(-\tau,\omega)| \, (\mathrm{d}\tau \times \mathrm{d}\omega) \mathbb{E}[\Lambda_n(0) - \Lambda_{n-1}(0)] \\ &= \| \mathbb{E}|h| \|_{L^1} \mathbb{E}[\Lambda_n(0) - \Lambda_{n-1}(0)], \end{aligned}$$

where we use Fubini's theorem, and where the second equality follows by stationarity of (Λ_n). It follows that

$$\sum_{n\geqslant 0} \mathbb{E}[\Lambda_{n+1}(0) - \Lambda_n(0)] \leqslant \frac{\phi(0)}{1 - \| \mathbb{E}|h| \|_{L^1}} < \infty,$$

hence (Λ_n) converges in L^1 to some limit process Λ. Using the same bounds, Markov's inequality gives

$$\mathbb{P}\left(\Lambda_{n+1}(0) - \Lambda_n(0) \geqslant \|\mathbb{E}|h|\|_{L^1}^{n/2}\right) \leqslant \phi(0)\|\mathbb{E}|h|\|_{L^1}^{n/2},$$

and since $\sum_{n\geqslant 0} \|\mathbb{E}|h|\|_{L^1}^{n/2} < \infty$, an application of Borel-Cantelli gives that (Λ_n) converges a.s. as well, to the same limit Λ.

Next, since $N_n - N_{n-1}$ is a point process itself, for any bounded $A \in \mathcal{B}(\mathbb{R})$ of Lebesgue measure $\mathrm{Leb}(A) < \infty$, and $B \in \mathcal{F}$,

$$\begin{aligned}\sum_{n\geqslant 0} \mathbb{P}\left(\int_{A\times B} (N_{n+1} - N_n)(\mathrm{d}\tau \times \mathrm{d}\omega) \neq 0\right) &\leqslant \sum_{n\geqslant 0} \mathbb{E}\int_{A\times B} (N_{n+1} - N_n)(\mathrm{d}\tau \times \mathrm{d}\omega) \\ &= \mathrm{Leb}(A)\mathbb{Q}(B)\sum_{n\geqslant 0}\mathbb{E}[\Lambda_{n+1}(0) - \Lambda_n(0)],\end{aligned}$$

which is finite, using that $\mathbb{Q}$ is a probability measure. Hence, by Borel-Cantelli, N_n is a.s. eventually constant on any bounded $A \times B \in \mathcal{B}(\mathbb{R}) \times \mathcal{F}$, whence it converges to some process N. The left-shift operator is continuous, whence

$$S_t N(\tilde{\omega}) = S_t \lim_{n\to\infty} N_n(\tilde{\omega}) = \lim_{n\to\infty} S_t N_n(\tilde{\omega}) = \lim_{n\to\infty} N_n(S_t\tilde{\omega}) = N(S_t\tilde{\omega}),$$

i.e., N inherits the S_t-compatibleness of $(N_n)_{n\geqslant 0}$.

To finish the proof of the existence part, we verify that the limit processes N, Λ satisfy the stated dynamics. First, by Fatou's lemma, for all $A \in \mathcal{B}(\mathbb{R})$, $B \in \mathcal{F}$ of bounded measure, it holds that

$$\begin{aligned}&\mathbb{E}\int_{A\times B} |N(\mathrm{d}\tau \times \mathrm{d}\omega) - M(\mathrm{d}\tau \times [0, \Lambda(\tau)] \times \mathrm{d}\omega)| \\ &\leqslant \liminf_{n\to\infty} \mathbb{E}\int_{A\times B} |M(\mathrm{d}\tau \times [0, \Lambda_n(\tau)] \times \mathrm{d}\omega) - M(\mathrm{d}\tau \times [0, \Lambda(\tau)] \times \mathrm{d}\omega)| \\ &= \mathrm{Leb}(A)\mathbb{Q}(B) \liminf_{n\to\infty} \mathbb{E}|\Lambda_n(0) - \Lambda(0)| = 0,\end{aligned}$$

where we use stationarity of the intensity processes as we did before. Note that the limits in the previous display actually exists, so that we can replace the limit inferiors by limits. Hence, N is a modification of a process with conditional intensity $\Lambda(\cdot)$. For the process $\Lambda(\cdot)$, note that

$$\begin{aligned}&\mathbb{E}\left|\Lambda(0) - \phi\left(\int_{(-\infty,0)\times\Omega} h(-\tau,\omega)\, N(\mathrm{d}\tau \times \mathrm{d}\omega)\right)\right| \\ &\leqslant \mathbb{E}|\Lambda(0) - \Lambda_{n+1}(0)| + \mathbb{E}\int_{(-\infty,0)\times\Omega} h(-\tau,\omega)\, (N - N_n)(\mathrm{d}\tau \times \mathrm{d}\omega) \\ &= \mathbb{E}|\Lambda(0) - \Lambda_{n+1}(0)| + \|\mathbb{E}|h|\|_{L^1}\mathbb{E}|\Lambda(0) - \Lambda_n(0)|,\end{aligned}$$

where we apply Lemma 1, the Lipschitz condition, the triangle inequality, and Fubini's theorem. Hence, by letting $n \to \infty$ and by using stationarity, we see that $\Lambda(\cdot)$ is a modification of the process satisfying dynamics (8).

Uniqueness. To prove uniqueness of the stationary solution $\tilde{N}$ with finite mean intensity $\tilde{\Lambda}$, we show that such a process satisfies initial condition (ii) given in the theorem. From the stability part, it then follows that $S_t\tilde{N} \xrightarrow{\mathcal{D}} N$. By stationarity, $S_t\tilde{N} \overset{\mathcal{D}}{=} \tilde{N}$, so $\tilde{N} \overset{\mathcal{D}}{=} N$.

Indeed, by a change of variables,

$$\mathbb{E}_M i_c(t) = \tilde{\Lambda} \int_{t-c}^{t} \int_{(-\infty,0)\times\Omega} |h(s-\tau,\omega)| \ (\mathrm{d}\tau \times \mathrm{d}\omega) \ \mathrm{d}s \leqslant c\tilde{\Lambda} \int_{t-c}^{\infty} \mathbb{E}|h(\tau,\omega)| \ \mathrm{d}\tau;$$

note that this upper bound tends to 0 as $t \to \infty$ by dominated convergence and Fubini, and that we have $\mathbb{E}_M i_c(t) \leqslant c\tilde{\Lambda}\|\mathbb{E}|h|\|_{L^1}$ for all $t \in \mathbb{R}$. This verifies initial condition (ii).

Stability. Let $\tilde{N}$ be a bivariate point process marked by random functions with dynamics (8) on $\mathbb{R}_+$, satisfying initial condition (i). In particular, we do *not* assume that it also satisfies dynamics (8) on $\mathbb{R}_-$. We prove that the finite-dimensional distributions of $S_t\tilde{N}$ converge to those of S_tN. Then [15], Theorem 11.1.VII gives stability: $S_t\tilde{N}_+ \xrightarrow{\mathcal{D}} N_+$, as $t \to \infty$.

We prove convergence of finite-dimensional distributions by proving that for every $c \in (0,t)$,

$$\mathbb{P}\left(N\{\tau\} \neq \tilde{N}\{\tau\} \text{ for some } \tau \in (t-c,t) \middle| \mathcal{H}_0^{\tilde{N}}\right) \to 0$$

as $t \to \infty$. Here, we assume that N and $\tilde{N}$ are constructed using the same marked bivariate Poisson process M of unit rate. This is justified as follows. It can be proved that the $\mathcal{H}_t^{\tilde{N}}$-intensity

$$\tilde{\Lambda}(t) = \phi\left(\int_{(-\infty,t)\times\Omega} h(t-\tau,\omega) \ \tilde{N}(\mathrm{d}\tau \times \mathrm{d}\omega)\right) \tag{54}$$

of $\tilde{N}$ is such that $t \mapsto \mathbb{E}[\tilde{\Lambda}(t)|\mathcal{H}_0^{\tilde{N}}]$ is a.s. locally integrable; this is proved in the same way as in [10], Theorem 1. For this we need the assumption $\|\mathbb{E}|h|\|_{L^\infty} < \infty$. It follows that $\tilde{N}$ is nonexplosive, a.s. Then [36], Lemma 2, implies existence of some marked bivariate Poisson process M of unit rate from which $\tilde{N}$ can be constructed using Lemma 1.

In order to prove convergence of finite-dimensional distributions, we consider

$$f(t) = \mathbb{E}\left[|\Lambda(t) - \tilde{\Lambda}(t)| \,\middle|\, \mathcal{H}_0^{\tilde{N}}\right] \mathbf{1}\{t \geqslant 0\},$$

which is a.s. locally integrable because $t \mapsto \mathbb{E}[\tilde{\Lambda}(t)|\mathcal{H}_0^{\tilde{N}}]$ is. Here, $\mathcal{H}_t^{\tilde{N}}$ is the sigma-algebra generated by the history of $\tilde{N}$ up to time t. Also consider the integrated version of f:

$$\begin{aligned} F(t) &:= \int_{t-c}^{t} f(\tau) \ \mathrm{d}\tau = \mathbb{E}\left[\int_{t-c}^{t} \mathrm{d}|N-\tilde{N}|(\tau) \,\middle|\, \mathcal{H}_0^{\tilde{N}}\right] \\ &\geqslant \mathbb{P}\left(N\{\tau\} \neq \tilde{N}\{\tau\} \text{ for some } \tau \in (t-c,t) \,\middle|\, \mathcal{H}_0^{\tilde{N}}\right), \end{aligned}$$

where again $c \in (0,t)$. By the last inequality, it suffices to prove that $F(t) \to 0$ as $t \to \infty$.

With Λ denoting the average intensity of $\Lambda(t)$, it holds for $t \geqslant 0$ that

$$\begin{aligned} f(t) \leqslant &\int_{(-\infty,0)\times\Omega} |h(t-\tau,\omega)| \ \tilde{N}(\mathrm{d}\tau \times \mathrm{d}\omega) + \Lambda \int_{(-\infty,0)\times\Omega} |h(t-\tau,\omega)| \ (\mathrm{d}\tau \times \mathrm{d}\omega) \\ &+ \int_{(0,t)\times\Omega} |h(t-\tau,\omega)| f(\tau) \ (\mathrm{d}\tau \times \mathrm{d}\omega). \end{aligned}$$

Integrating from $t-c$ to t gives, after some more bounding,

$$F(t) \leqslant j_c(t) + \int_0^t \mathbb{E}|h(\tau)| F(t-\tau) \ \mathrm{d}\tau,$$

where $j_c(t) = i_c(t) + c\Lambda \int_{t-c}^{\infty} \mathbb{E}|h(\tau)|\, \mathrm{d}\tau$. This is a Volterra integral inequality of the second kind. Since $\|\mathbb{E}|h|\|_{L^1} < 1$, Picard iteration gives

$$F(t) \leqslant \int_0^t j_c(t-\tau) \left(\sum_{n \geqslant 0} (\mathbb{E}|h|)^{n*}(\tau) \right) \mathrm{d}\tau.$$

Note that $\sum_{n\geqslant 0}(\mathbb{E}|h|)^{n*}(\tau)$ can be bounded in L^1 by Young's convolution inequality. Also, by our assumption (i), it follows that $j_c(t)$ is bounded a.s. and converges to 0 as $t \to \infty$. By dominated convergence, $F(t) \to 0$ as $t \to \infty$, finishing the proof of the stability part. □

Appendix B: Relegated details of Section 4

We claim at the start of Section 4 that it is possible to distinguish between a Hawkes and a delayed Hawkes process using statistical techniques. In particular, suppose that one generates realizations on $[0, T] \ni t$, with $T = 50{,}000$, of a univariate, linear, exponential delayed Hawkes pure-birth process $N(t)$ having conditional intensity

$$\Lambda(t) = \lambda_0 + \sum_{t_i < t} \alpha e^{-r(t-t_i)}, \tag{55}$$

where $t_i - J_i$ are the event times of $N(t)$, with $J_i \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\mu)$. In other words, (t_i) correspond to the death times of the birth-death process Q associated with N. We choose parameters $(\lambda_0, \alpha, r, \mu) = (1/6, 3, 3.6, 1/6)$, where the first three parameters imply that the expected stationary arrival intensity of N equals 1.

We fit N to a parametric null hypothesis consisting of univariate, linear, exponential Hawkes processes. In particular, for the parameter space $\Theta = (0, 10)^3$, we consider the parametric null hypothesis

$$H_0^{\mathrm{Exp}} : N \overset{d}{=} N_\theta^{\mathrm{Exp}} \text{ for some } \theta \in \{(\lambda_0, \alpha, r) \in \Theta : \alpha < r\}, \tag{56}$$

where $N_\theta^{\mathrm{Exp}} = N_{\lambda_0,\alpha,r}^{\mathrm{Exp}}$ is a univariate, linear, exponential Hawkes process having intensity

$$\lambda_\theta^{\mathrm{Exp}}(t) = \lambda_{\lambda_0,\alpha,r}^{\mathrm{Exp}}(t) = \lambda_0 + \sum_{t_i < t} \alpha e^{-r(t-t_i)}, \tag{57}$$

where t_i denote the event times of N_θ^{Exp}.

We apply the asymptotically correct goodness-of-fit test described in [3], Algorithm 1, using $n = \mathrm{ceil}(\sqrt{T}/4)$ and an Andersen-Darling test in step (v) of their algorithm; these choices are motivated in [3]. Out of 1,000 simulated sample paths, we reject 494, 768 and 949 times using significance levels of 0.01, 0.05, and 0.20, respectively. Hence, we can clearly detect the deviation of the delayed Hawkes process from the non-delayed null hypothesis empirically.

Appendix C: Relegated proofs of Section 5

Proof of Theorem 5. In this proof, we first assume that $B \in \mathrm{APT}(-\alpha)$, so that B is of class $\mathscr{R}(-\alpha)$ with $\ell(x)$ a function converging to a positive constant. Under this assumption, we prove that also $Q(t) \in \mathscr{R}(-\alpha)$. Then we argue that essentially the same proof holds to show that $B \in \mathscr{R}(-\alpha)$ implies $Q(t) \in \mathscr{R}(-\alpha)$, and we indicate what needs to be changed in the proof.

By specifying (22) to the univariate case, and setting $s = 0$, we express the Z-transform of $Q(t)$ as

$$\mathbb{E}\left[z^{Q(t)}\right] = \exp\left(-\lambda_0 \int_0^t (1-\eta(u,z))\,\mathrm{d}u\right), \tag{58}$$

where $\eta(u,z) := \mathbb{E}[z^{S^Q(u)}]$, the Z-transform of the birth-death cluster process S^Q. It satisfies

$$\eta(u,z) = \mathscr{J}(u)z + \int_0^u \beta\left(\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right)\mathrm{d}\bar{\mathscr{F}}(w), \tag{59}$$

which follows by specifying (25) to the univariate delayed Hawkes setting, and where $\mathscr{J}$, $\bar{\mathscr{F}}$ denote the survival function and CDF, respectively, of the generic sojourn time random variable J. In the remainder of the proof, we invoke a Tauberian theorem to relate the behavior of a regularly varying function at infinity to the behavior of its Laplace-Stieltjes transform at 0. This relation for β is substituted into (59), after which we analyze expansions for $\eta(u,z)$ and $\mathbb{E}[z^{Q(t)}]$. By invoking the Tauberian theorem in the reverse direction, we conclude that $Q(t)$ is also of class $\mathscr{R}(-\alpha)$.

As indicated, we first assume that $\mathbb{P}(B > x)x^\alpha \to C$ for some $C > 0$. Then it follows from the Tauberian theorem [8], Theorem 8.1.6, that $\beta(s) - 1 + sb_1 \sim -C\Gamma(1-\alpha)s^\alpha$ as $s \downarrow 0$. Hence, as $z \uparrow 1$,

$$\begin{aligned}
&\beta\left(\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right) - 1 + b_1\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s \\
&\sim -C\Gamma(1-\alpha)\left(\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha.
\end{aligned} \tag{60}$$

Substituting this into (59) yields, as $z \uparrow 1$,

$$\begin{aligned}
1-\eta(u,z) &\sim 1 - \mathscr{J}(u)z - \int_0^u \left\{1 - b_1\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right. \\
&\quad \left. - C\Gamma(1-\alpha)\left(\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha\right\}\mathrm{d}\bar{\mathscr{F}}(w) \\
&= \mathscr{J}(u)(1-z) + \int_0^u \left\{b_1\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right. \\
&\quad \left. + C\Gamma(1-\alpha)\left(\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha\right\}\mathrm{d}\bar{\mathscr{F}}(w).
\end{aligned} \tag{61}$$

Next, expand $1-\eta(u,z) = \mathbb{E}[S(u)](1-z) + o(1-z)$, as $z \uparrow 1$. Write $\mathbb{E}[S(u)] = R_1(u)$ for the leading term. Substituting this into (61) and comparing terms of order $1-z$, we see that R_1 satisfies

$$\begin{aligned}
R_1(u) &= \mathscr{J}(u) + b_1\int_0^u\int_w^u h(s-w)R_1(u-s)\,\mathrm{d}s\,\mathrm{d}\bar{\mathscr{F}}(w) \\
&= \mathscr{J}(u) + b_1\int_0^u R_1(u-s)\int_0^s h(s-w)\,\mathrm{d}\bar{\mathscr{F}}(w)\,\mathrm{d}s \\
&= \mathscr{J}(u) + b_1(R_1 * \bar{h})(u),
\end{aligned} \tag{62}$$

where $\bar{h}$ is defined by $\bar{h}(s) := \int_0^s h(s-w)\,\mathrm{d}\bar{\mathscr{F}}(w)$, and where $*$ denotes the convolution operator. This is a Volterra equation of the second kind, and by Picard iteration we obtain, for $u \geqslant 0$,

$$R_1(u) = \sum_{n\geqslant 0} b_1^n(\bar{h}^{n*} * \mathscr{J})(u). \tag{63}$$

The next term in the expansion of $1-\eta(u,z)$ is of the form $R_\alpha(u)(1-z)^\alpha$. When we substitute $1-\eta(u,z)=\mathbb{E}[S(u)](1-z)+R_\alpha(u)(1-z)^\alpha+o((1-z)^\alpha)$ into (61) and compare terms of order $(1-z)^\alpha$, we obtain

$$R_\alpha(u)=b_1(R_\alpha*\bar{h})(u)+C\Gamma(1-\alpha)\int_0^u\left(\int_w^u h(s-w)R_1(u-s)\,\mathrm{d}s\right)^\alpha\,\mathrm{d}\bar{\mathscr{F}}(w). \tag{64}$$

This is again a Volterra equation of the second kind; by Picard iteration we obtain

$$R_\alpha(u)=C\Gamma(1-\alpha)\sum_{n\geqslant 0}b_1^n\left(\bar{h}^{n*}*\left(\int_0^\cdot\left(\int_w^\cdot h(s-w)R_1(\cdot-s)\,\mathrm{d}s\right)^\alpha\,\mathrm{d}\bar{\mathscr{F}}(w)\right)\right)(u). \tag{65}$$

From (16) with $\alpha=0$, we infer that $\|\bar{h}\|_{L^1}=\|h\|_{L^1}$. Hence, by applying Young's convolution inequality n times with $r=p=\infty, q=1$ to each term of (63), and by recognizing a geometric series, $\|\bar{h}\|_{L^1}b_1=\|h\|_{L^1}b_1<1$ implies that R_1 is a bounded function of u. Since $\|h\|_{L^1}<\infty$, the inner integral in (65) is finite, whence R_α is also a bounded function of u.

We now substitute the expansion $1-\eta(u,z)=\mathbb{E}[S(u)](1-z)+R_\alpha(u)(1-z)^\alpha+o((1-z)^\alpha)$ into (58), which gives, after expanding the exponential functions,

$$\begin{aligned}\mathbb{E}\left[z^{Q(t)}\right]&\sim\exp\left(-\lambda_0\int_0^t\left(R_1(u)(1-z)+R_\alpha(u)(1-z)^\alpha\right)\,\mathrm{d}u\right)\\&=1-\lambda_0(1-z)\int_0^t R_1(u)\,\mathrm{d}u-\lambda_0(1-z)^\alpha\int_0^t R_\alpha(u)\,\mathrm{d}u+o((1-z)^\alpha).\end{aligned} \tag{66}$$

By using the Tauberian theorem [8], Theorem 8.1.6, the other way around, it then follows that $Q(t)\in\mathscr{R}(-\alpha)$, as claimed.

We now indicate what we have to change in the proof if we assume that $B\in\mathscr{R}(-\alpha)$, so that $\mathbb{P}(B>x)=\ell(x)x^{-\alpha}$ for some slowly varying function ℓ. Note that the constant $-C\Gamma(1-\alpha)$ in (61) should in that case be replaced by $\ell(1/I(u,z;w))$, where $I(u,z;w)=\int_w^u h(s-w)(1-\eta(u-s,z))\,\mathrm{d}s$.

For small $\delta\in(0,\alpha-1)$, we use Potter's Theorem (i.e., [8], Theorem 1.5.6) to conclude that for z sufficiently close to 1 and for some $A>1$,

$$\ell\left(\frac{1}{I(u,z;w)}\right)\Big/\ell\left(\frac{1}{1-z}\right)\leqslant A\max\left\{\left(\frac{1-z}{I(u,z;w)}\right)^\delta,\left(\frac{1-z}{I(u,z;w)}\right)^{-\delta}\right\}. \tag{67}$$

Our assumptions on h imply that given $\epsilon>0$, there exists $K>0$ such that

$$\sup_{v\in[0,u]}\mathbb{P}(S^Q(v)>K)<\epsilon, \tag{68}$$

whence for $0<z<1$ and $0\leqslant v\leqslant u$ we have $\eta(v,z)=\mathbb{E}[z^{S^Q(v)}]\geqslant(1-\epsilon)z^K$. Hence, we have, as $z\uparrow 1$,

$$\frac{1}{I(u,z;w)}\geqslant\frac{1}{(1-(1-\epsilon)z^K)\int_0^u h(s)\,\mathrm{d}s}\to\frac{1}{\epsilon\int_0^u h(s)\,\mathrm{d}s}.$$

Given some threshold $D>0$ such that (67) holds for $\ell(x)/\ell(y)$ for all $x,y\geqslant D$, cf. [8], Theorem 1.5.6, we choose $\epsilon>0$ sufficiently small to assure that $\epsilon\int_0^u h(s)\,\mathrm{d}s\leqslant 1/(2D)$, so that we can find some $z^*\in(0,1)$ such that $z\geqslant z^*$ implies that $1/I(u,z;w)\geqslant D$ for all s,w. We also have $(z-1)^{-1}\geqslant D$ for $z\geqslant 1-1/D$.

When we have $B \in \mathscr{R}(-\alpha)$ instead of $B \in \mathrm{APT}(-\alpha)$, we replace in (60) the factor $-C\Gamma(1-\alpha)$ by $\ell(1/I(u,z;w))$. For $z > \max\{z^*, 1-\frac{1}{D}\}$, we apply the bound (67) and a similar Potter bound for

$$\ell\left(\frac{1}{1-z}\right) \Big/ \ell\left(\frac{1}{I(u,z;w)}\right); \tag{69}$$

then we have an upper and a lower bound for the asymptotic expansion of $1-\eta(u,z)$, to which we conduct an analysis analogous to the case $B \in \mathrm{APT}(-\alpha)$, yielding $Q(t) \in \mathscr{R}(-\alpha)$ both when we use the upper bound as if it were the true expansion, and when we use the lower bound. We conclude that $Q(t) \in \mathscr{R}(-\alpha)$. □

Proof of Corollary 1. Letting $t \to \infty$ in (66), we have

$$1 - \mathbb{E}\left[z^Q\right] \sim \lambda_0(1-z)\int_0^\infty R_1(u)\,\mathrm{d}u + \lambda_0(1-z)^\alpha \int_0^\infty R_\alpha(u)\,\mathrm{d}u. \tag{70}$$

By applying Young's convolution inequality to each term of (63) and by recognizing a geometric series, we observe that $\int_0^\infty R_1(u)\,\mathrm{d}u$ is of order $(1-\rho)^{-1}$, as $\rho \uparrow 1$. Similarly, we use (65) to conclude that $\int_0^\infty R_\alpha(u)\,\mathrm{d}u$ is of order $(1-\rho)^{-\alpha-1}$.

Note that $\mathbb{E}[Q] = \lambda_0 \int_0^\infty R_1(u)\,\mathrm{d}u$, so $\mathbb{E}[Q] = O\left((1-\rho)^{-1}\right)$, as $\rho \uparrow 1$, i.e., $(1-\rho)Q$ stays bounded as $\rho \uparrow 1$. More specifically, using $1 - z^{1-\rho} = (1-\rho)(1-z) + O\left((1-z)^2\right)$, as $z \uparrow 1$, we have, up to $O\left((1-z)^2\right)$ terms,

$$\begin{aligned} 1 - \mathbb{E}\left[z^{(1-\rho)Q}\right] &\sim \lambda_0\left(1 - z^{(1-\rho)}\right)\int_0^\infty R_1(u)\,\mathrm{d}u + \lambda_0\left(1 - z^{(1-\rho)}\right)^\alpha \int_0^\infty R_\alpha(u)\,\mathrm{d}u \\ &= (1-\rho)(1-z)\lambda_0 \int_0^\infty R_1(u)\,\mathrm{d}u + (1-\rho)^\alpha(1-z)^\alpha\lambda_0 \int_0^\infty R_\alpha(u)\,\mathrm{d}u. \end{aligned} \tag{71}$$

From this expansion, it is clear that $(1-\rho)\lambda_0 \int_0^\infty R_1(u)\,\mathrm{d}u < \infty$ as $\rho \uparrow 1$. The second term in (71) diverges, as $\rho \uparrow 1$, which implies that $X := \lim_{\rho\uparrow 1}(1-\rho)Q$ satisfies $\mathbb{E}\left[X^\alpha\right] = \infty$. □

Appendix D: Supplement to Section 5: Cluster size distributions for gamma-distributed marks

In this appendix, we study the distribution of the cluster size of the delayed Hawkes process, that is, the total number of descendants of a single immigrant, including the immigrant itself. Note that the offspring size is given by a Poisson random variable with parameter equal $B\int_J^\infty h(t-J)\,\mathrm{d}t = B\varrho$, where J is the sojourn time of the parent, and where we set $\varrho := \|h\|_{L^1}$. In particular, the offspring distribution is the same as the one for a Hawkes process having the same parameters. This implies that the total size of a cluster is the same for both processes, and is given by the total progeny size of a Galton-Watson branching process, which can be determined with the aid of the hitting time theorem, see, e.g., [24].

Lemma 2 (Hitting time theorem)**.** *The total progeny size Z of a Galton-Watson branching process with offspring distribution X has a distribution with probability mass function*

$$\mathbb{P}(Z=n) = \frac{1}{n}\mathbb{P}\left(\sum_{k=1}^n X_k = n-1\right), \tag{72}$$

where $(X_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of random variables having the same distribution as X.

For unmarked Hawkes processes, it is a well-known result that $Z \sim \text{Borel}(\varrho)$, i.e.,

$$\mathbb{P}(Z = n) = \frac{e^{-\varrho n}(\varrho n)^{n-1}}{n!}.$$

Even without the probabilistic context, it can be proved that those Borel probabilities sum to unity by setting $x = -\varrho e^{-\varrho} \in (-e^{-1}, 0)$ for $\varrho \in (0, 1)$, and by considering the Taylor expansion around 0 of the principal branch of the Lambert W function.

We now consider a marked (delayed) Hawkes process under the stability condition $\mathbb{E}[B]\varrho < 1$. In this case, the offspring size follows a mixed-Poisson type distribution. To make use of Lemma 2, we want this distribution to be such that i.i.d. sums belong to a well-known parametric family. This is the case for gamma-distributed marks. In fact, the assumption of gamma-distributed marks is not too restrictive, for the set of mixtures of gamma distributions is dense in the set of continuous probability distributions on $[0, \infty)$.

Proposition 1. *Let $\alpha, c > 0$ be such that $\alpha\varrho/c < 1$. Consider a (delayed) Hawkes process with $\Gamma(\alpha, c)$ distributed marks, i.e., the marks admit a density*

$$f_B(x) = \frac{c^\alpha x^{\alpha-1} e^{-cx}}{\Gamma(\alpha)}.$$

Then the total cluster size Z is finite a.s. and has probability mass function

$$\mathbb{P}(Z = n) = \frac{1}{n}\binom{(\alpha+1)n - 2}{n-1}\left(\frac{c}{c+\varrho}\right)^{\alpha n}\left(\frac{\varrho}{c+\varrho}\right)^{n-1}, \quad n \in \mathbb{N}, \tag{73}$$

where, for $x, y \in \mathbb{R}$, with $x > y - 1$, we use the generalized binomial coefficient

$$\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}. \tag{74}$$

Proof. Let X denote the offspring random variable. Since, in self-evident notation, $X \mid B \sim \text{Pois}(B\varrho)$,

$$\begin{aligned}\mathbb{P}(X = n) = \mathbb{E}[\mathbb{P}(X = n|B)] &= \int_0^\infty \frac{e^{-\varrho x}(\varrho x)^n}{n!}\frac{c^\alpha x^{\alpha-1}e^{-cx}}{\Gamma(\alpha)}\,\mathrm{d}x = \frac{c^\alpha \varrho^n}{\Gamma(\alpha)n!}\int_0^\infty e^{-(c+\varrho)x}x^{\alpha+n-1}\,\mathrm{d}x \\ &= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)n!}\frac{c^\alpha\varrho^n}{(c+\varrho)^{\alpha+n}} = \binom{\alpha+n-1}{n}p^\alpha(1-p)^n,\end{aligned}$$

where $p := c/(c+\varrho)$. Hence, $X \sim \text{NB}(\alpha, p)$, i.e., X follows the *generalized* negative binomial distribution; note that $\alpha > 0$ is not necessarily integer. It follows that if $X_1, \ldots, X_n$ are i.i.d. copies of X, then $\sum_{k=1}^n X_k \sim \text{NB}(\alpha n, p)$. The result now follows by an application of Lemma 2. □

When $\alpha = 1$, the gamma distribution reduces to an exponential distribution, and we obtain

$$\mathbb{P}(Z = n) = C_{n-1}p^n(1-p)^{n-1}, \tag{75}$$

where $C_n = \binom{2n}{n}/(n+1)$ is the nth Catalan number and p as defined in the proof of Proposition 1. Note that the ephemerally self-exciting process with intensity jump ϱ and expiration rate c has the same progeny distribution, see [17], Proposition 3.3. This is no coincidence. Letting B be the expiration time of the ephemeral excitation, $X|B \sim \text{Pois}(B\varrho)$, where $B \sim \text{Exp}(c)$, showing that the offspring random variable X has the same probabilistic behavior under the ephemerally self-exciting process and the (delayed) Hawkes process with exponentially distributed marks.

Appendix E: Relegated proofs of Sections 6–7

Proof of Corollary 2. We know from Theorem 6 that $Q(t) \geqslant_{\mathrm{st}} \tilde{Q}(t)$ and $\Lambda(t) \geqslant_{\mathrm{st}} \tilde{\Lambda}(t)$ for all $t \geqslant 0$. Furthermore, from Theorem 4 with $d = 1$, by (i) differentiating w.r.t. z and setting $z = 1$ and $s = 0$; and by (ii) differentiating w.r.t. s and setting $z = 1$ and $s = 0$, we find that

$$\mathbb{E}[\tilde{Q}(\infty)] = \lambda_0 \int_0^\infty \mathbb{E}\left[S^{\tilde{Q}}(u)\right] \,\mathrm{d}u,$$
$$\mathbb{E}[\tilde{\Lambda}(\infty)] = \lambda_0 \left(1 + \int_0^\infty \mathbb{E}\left[S^{\tilde{\Lambda}}(u)\right] \,\mathrm{d}u\right),$$

where $S^{\tilde{Q}}$ and $S^{\tilde{\Lambda}}$ denote the cluster birth-death and intensity processes corresponding to a single immigrant. Those processes were introduced in Section 2. Note that $\int_0^\infty \mathbb{E}[S^{\tilde{Q}}(u)]\,\mathrm{d}u$ measures the total active time over all particles in the cluster, whereas $\int_0^\infty \mathbb{E}[S^{\tilde{\Lambda}}(u)]\,\mathrm{d}u$ corresponds to the limit of the compensator of the conditional intensity of the cluster. Since clusters are a.s. finite, since we integrate over $[0,\infty)$, and since integration is invariant over right shifts, both quantities do not depend on possible delays of the particles. Thus, it follows that $\mathbb{E}[Q(\infty)] = \mathbb{E}[\tilde{Q}(\infty)]$ and $\mathbb{E}[\Lambda(\infty)] = \mathbb{E}[\tilde{\Lambda}(\infty)]$.

Hence, to prove the claim, it suffices to prove that if $X(\cdot), Y(\cdot)$ are stochastic processes on $[0,\infty)$ such that for all $t \geqslant 0$, $X(t) \geqslant_{\mathrm{st}} Y(t)$, while

$$X(t) \xrightarrow{\mathcal{D}} X, \quad Y(t) \xrightarrow{\mathcal{D}} Y,$$

as $t \to \infty$, with $\mathbb{E}[X] = \mathbb{E}[Y]$, then $X \overset{\mathcal{D}}{=} Y$.

Indeed, let $F_{X(t)}$ be the CDF of $X(t)$ and $F_{Y(t)}$ be the CDF of $Y(t)$. Then, for all $t \geqslant 0, z \in \mathbb{R}$, $F_{X(t)}(z) \leqslant F_{Y(t)}(z)$ since $X(t) \geqslant_{\mathrm{st}} Y(t)$. Furthermore, for each continuity point z of F_X, $F_{X(t)}(z) \to F_X(z)$; similarly, for each continuity point z of F_Y, $F_{Y(t)}(z) \to F_Y(z)$. Hence, for all but at most countably many points z,

$$F_X(z) = \lim_{t\to\infty} F_{X(t)}(z) \leqslant \lim_{t\to\infty} F_{Y(t)}(z) = F_Y(z).$$

If this inequality does not hold for some z, then by right-continuity it does not hold for a continuum of values $[z, z+\epsilon]$. By contradiction, $F_X(z) \leqslant F_Y(z)$ for all $z \in \mathbb{R}$. Since $X, Y \geqslant 0$, it follows that

$$0 = \mathbb{E}[X] - \mathbb{E}[Y] = \int_0^\infty (1 - F_X(z))\,\mathrm{d}z - \int_0^\infty (1 - F_Y(z))\,\mathrm{d}z = \int_0^\infty (F_Y(z) - F_X(z))\,\mathrm{d}z.$$

Since the integrand is non-negative for all $z \geqslant 0$, it follows that $F_X(z) = F_Y(z)$ for almost all $z \geqslant 0$. Inequality at a point would again imply inequality on an interval of positive measure. Hence, $F_X = F_Y$, i.e., $X \overset{\mathcal{D}}{=} Y$. □

Proof of Corollary 3. The result follows from Corollary 2, in combination with [34], Theorems 6.4 and 6.6. □

Proof of Theorem 8. For $\boldsymbol{k} \in \mathbb{N}_0^d$, $\boldsymbol{\lambda} \in \mathbb{R}_+^d$, let

$$F(t,\boldsymbol{k},\boldsymbol{\lambda}) = \mathbb{P}(\boldsymbol{Q}(t) = \boldsymbol{k}, \boldsymbol{\Lambda}(t) \leqslant \boldsymbol{\lambda}), \quad f(t,\boldsymbol{k},\boldsymbol{\lambda}) = \frac{\partial^d F(t,\boldsymbol{k},\boldsymbol{\lambda})}{\partial\lambda_1\cdots\lambda_d},$$
$$\xi(t,\boldsymbol{k},\boldsymbol{s}) = \int_{\mathbb{R}_+^d} e^{-\boldsymbol{s}^\top\boldsymbol{\lambda}} f(t,\boldsymbol{k},\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{\lambda}.$$

Note that, with this notation,

$$\zeta(t, \boldsymbol{z}, \boldsymbol{s}) = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d} \boldsymbol{z}^{\boldsymbol{k}} \xi(t, \boldsymbol{k}, \boldsymbol{s}).$$

Let $\circ$ be the Hadamard product, and let $\boldsymbol{e}_j$ be the j-th standard unit vector in $\mathbb{R}^d$. We consider the Markovian dynamics between times t and $t + \Delta t$. Let $\boldsymbol{k} \in \mathbb{N}_0^d$ and $\boldsymbol{\lambda} \in \mathbb{R}_+^d$. Write $[\boldsymbol{0}, \boldsymbol{\lambda}] := [0, \lambda_1] \times \cdots \times [0, \lambda_d]$. Note that we may enter state $\boldsymbol{k}$ either due to an arrival in coordinate j, leaving state $\boldsymbol{k} - \boldsymbol{e}_j$; due to a departure in coordinate j, leaving state $\boldsymbol{k} + \boldsymbol{e}_j$; or due to rerouting from coordinate j to i, leaving state $\boldsymbol{k} + \boldsymbol{e}_j - \boldsymbol{e}_i$. Therefore, as $\Delta t \downarrow 0$,

$$\begin{aligned}
F(t + \Delta t, \boldsymbol{k}, \boldsymbol{\lambda} - \boldsymbol{r} \circ (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)\Delta t) = &\sum_{j=1}^d \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} y_j \Delta t f(t, \boldsymbol{k} - \boldsymbol{e}_j, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&+ \sum_{j=1}^d (k_j + 1)\mu_j \Delta t \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} \mathbb{P}(\boldsymbol{B}_j \leqslant \boldsymbol{\lambda} - \boldsymbol{y}) f(t, \boldsymbol{k} + \boldsymbol{e}_j, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&+ \sum_{j=1}^d \sum_{i=1}^d (k_j + 1)\mu_{ij} \Delta t F(t, \boldsymbol{k} + \boldsymbol{e}_j - \boldsymbol{e}_i, \boldsymbol{\lambda}) \\
&+ F(t, \boldsymbol{k}, \boldsymbol{\lambda}) \left(1 - \sum_{j=1}^d k_j \mu_j \Delta t - \sum_{j=1}^d \sum_{i=1}^d k_j \mu_{ij} \Delta t \right) \\
&- \sum_{j=1}^d \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} y_j \Delta t f(t, \boldsymbol{k}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} + o(\Delta t).
\end{aligned}$$

Subtracting $F(t, \boldsymbol{k}, \boldsymbol{\lambda})$ from both sides, dividing by Δt and taking the limit as $\Delta t \downarrow 0$ gives us

$$\begin{aligned}
&\frac{\partial F(t, \boldsymbol{k}, \boldsymbol{\lambda})}{\partial t} - \left[\frac{\partial F(t, \boldsymbol{k}, \boldsymbol{\lambda})}{\partial \lambda_1}, \cdots, \frac{\partial F(t, \boldsymbol{k}, \boldsymbol{\lambda})}{\partial \lambda_d} \right] (\boldsymbol{r} \circ (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)) \\
&= \sum_{j=1}^d \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} y_j f(t, \boldsymbol{k} - \boldsymbol{e}_j, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} + \sum_{j=1}^d (k_j + 1)\mu_j \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} \mathbb{P}(\boldsymbol{B}_j \leqslant \boldsymbol{\lambda} - \boldsymbol{y}) f(t, \boldsymbol{k} + \boldsymbol{e}_j, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&+ \sum_{j=1}^d \sum_{i=1}^d (k_j + 1)\mu_{ij} F(t, \boldsymbol{k} + \boldsymbol{e}_j - \boldsymbol{e}_i, \boldsymbol{\lambda}) - \sum_{j=1}^d k_j \mu_j F(t, \boldsymbol{k}, \boldsymbol{\lambda}) \\
&- \sum_{j=1}^d \sum_{i=1}^d k_j \mu_{ij} F(t, \boldsymbol{k}, \boldsymbol{\lambda}) - \sum_{j=1}^d \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} y_j f(t, \boldsymbol{k}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.
\end{aligned}$$

Next, we take the partial derivative with respect to the intensity λ_j of each coordinate; i.e., we apply the differential operator $\frac{\partial^d}{\partial \lambda_1 \cdots \lambda_d}$ to both sides of the last equation. Here we apply Leibniz' integral rule and we use our assumption $\mathbb{P}(B_{ij} \leqslant 0) = 0$ for all $i, j \in [d]$. This yields

$$\begin{aligned}
&\frac{\partial f(t, \boldsymbol{k}, \boldsymbol{\lambda})}{\partial t} - \sum_{j=1}^d r_j \frac{\partial}{\partial \lambda_j} (\lambda_j f(t, \boldsymbol{k}, \boldsymbol{\lambda})) + \sum_{j=1}^d r_j \lambda_{j,0} \frac{\partial f(t, \boldsymbol{k}, \boldsymbol{\lambda})}{\partial \lambda_j} \\
&= \sum_{j=1}^d \lambda_j f(t, \boldsymbol{k} - \boldsymbol{e}_j, \boldsymbol{\lambda}) + \sum_{j=1}^d (k_j + 1)\mu_j \int_{[\boldsymbol{0},\boldsymbol{\lambda}]} \frac{\partial^d \mathbb{P}(\boldsymbol{B}_j \leqslant \boldsymbol{\lambda} - \boldsymbol{y})}{\partial \lambda_1 \cdots \partial \lambda_d} f(t, \boldsymbol{k} + \boldsymbol{e}_j, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}
\end{aligned} \tag{76}$$

$$+\sum_{j=1}^{d}\sum_{i=1}^{d}(k_j+1)\mu_{ij}f(t,\boldsymbol{k}+\boldsymbol{e}_j-\boldsymbol{e}_i,\boldsymbol{\lambda})-\sum_{j=1}^{d}(k_j\mu_j+\lambda_j)f(t,\boldsymbol{k},\boldsymbol{\lambda})-\sum_{j=1}^{d}\sum_{i=1}^{d}k_j\mu_{ij}f(t,\boldsymbol{k},\boldsymbol{\lambda}).$$

Our next step is transforming to $\xi(t,\boldsymbol{k},\boldsymbol{s})$ by applying the integral operator $\int_{\mathbb{R}_+^d} e^{-\boldsymbol{s}^\top\boldsymbol{\lambda}}\cdot\,\mathrm{d}\boldsymbol{\lambda}$ to both sides of (76). We can do this term by term; the calculations rely on integration by parts, Tonelli's theorem, and swapping the order of differentiation and integration. We obtain

$$\begin{aligned}
&\frac{\partial\xi(t,\boldsymbol{k},\boldsymbol{s})}{\partial t}+\sum_{j=1}^{d}\left((r_js_j-1)\frac{\partial\xi(t,\boldsymbol{k},\boldsymbol{s})}{\partial s_j}+\frac{\partial\xi(t,\boldsymbol{k}-\boldsymbol{e}_j,\boldsymbol{s})}{\partial s_j}\right)+\sum_{j=1}^{d}r_j\lambda_{j,0}s_j\xi(t,\boldsymbol{k},\boldsymbol{s})\\
&=\sum_{j=1}^{d}(k_j+1)\mu_j\beta_j(\boldsymbol{s})\xi(t,\boldsymbol{k}+\boldsymbol{e}_j,\boldsymbol{s})-\sum_{j=1}^{d}k_j\mu_j\xi(t,\boldsymbol{k},\boldsymbol{s})\\
&+\sum_{j=1}^{d}\sum_{i=1}^{d}(k_j+1)\mu_{ij}\xi(t,\boldsymbol{k}+\boldsymbol{e}_j-\boldsymbol{e}_i,\boldsymbol{s})-\sum_{j=1}^{d}\sum_{i=1}^{d}k_j\mu_{ij}\xi(t,\boldsymbol{k},\boldsymbol{s}).
\end{aligned}\tag{77}$$

Now multiplying by $\boldsymbol{z}^{\boldsymbol{k}}$ and summing over $\boldsymbol{k}\in\mathbb{N}_0^d$ gives (41).

The final part of the theorem follows by the method of characteristics, similarly as in [34], Theorem 3.1, by parametrising s_j and z_j by $u\in[0,t]$, with $s_j(t)=s_j$ and $z_j(t)=z_j$, after which we change variables to $u'=t-u$. □

Proof of Theorem 9. We rewrite (41) to the joint transform by substituting its definition

$$\zeta(t,\boldsymbol{z},\boldsymbol{s})=\mathbb{E}\left[\boldsymbol{z}^{\boldsymbol{Q}(t)}e^{-\boldsymbol{s}^\top\boldsymbol{\Lambda}(t)}\right]=\mathbb{E}\left[\prod_{j=1}^{d}z_j^{Q_j(t)}e^{-s_j\Lambda_j(t)}\right].$$

This gives us the PDE

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\prod_{l=1}^{d}z_l^{Q_l(t)}e^{-s_l\Lambda_l(t)}\right]-\sum_{j=1}^{d}(r_js_j+z_j-1)\mathbb{E}\left[\prod_{l=1}^{d}z_l^{Q_l(t)}\Lambda_j(t)e^{-s_l\Lambda_l(t)}\right]\\
&\quad+\sum_{j=1}^{d}\mu_j(z_j-\beta_j(\boldsymbol{s}))\mathbb{E}\left[Q_j(t)z_j^{Q_j(t)-1}e^{-s_j\Lambda_j(t)}\prod_{\substack{l=1\\l\neq j}}^{d}z_l^{Q_l(t)}e^{-s_l\Lambda_l(t)}\right]\\
&\quad+\sum_{j=1}^{d}\sum_{i=1}^{d}\mu_{ij}(z_j-z_i)\mathbb{E}\left[Q_j(t)z_j^{Q_j(t)-1}e^{-s_j\Lambda_j(t)}\prod_{\substack{l=1\\l\neq j}}^{d}z_l^{Q_l(t)}e^{-s_l\Lambda_l(t)}\right]\\
&=-\sum_{j=1}^{d}r_j\lambda_{j,0}s_j\mathbb{E}\left[\prod_{l=1}^{d}z_l^{Q_l(t)}e^{-s_l\Lambda_l(t)}\right].
\end{aligned}\tag{78}$$

We differentiate (78) $\boldsymbol{g}\in\mathbb{N}_0^d$ times with respect $\boldsymbol{s}$, meaning that we differentiate g_j times with respect to s_j, for each $j\in[d]$. After this, we set $\boldsymbol{s}=\boldsymbol{0}$. Similarly, we differentiate $\boldsymbol{q}\in\mathbb{N}_0^d$ times with respect to $\boldsymbol{z}$, after which we set $\boldsymbol{z}=\boldsymbol{1}$. This yields the following ODE for $\mathbb{E}\left[\prod_{l=1}^{d}\bar{Q}_l^{q_l}(t)\Lambda_l^{g_l}(t)\right]$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\prod_{l=1}^{d}\bar{Q}_l^{q_l}(t)\Lambda_l^{g_l}(t)\right]+\sum_{j=1}^{d}(g_jr_j+q_j\mu_j)\mathbb{E}\left[\prod_{l=1}^{d}\bar{Q}_l^{q_l}(t)\Lambda_l^{g_l}(t)\right]$$

$$
\begin{aligned}
&- \sum_{j=1}^{d} q_j \mathbb{E}\left[\bar{Q}_j^{q_j-1}(t)\Lambda_j^{g_j+1}(t)\prod_{\substack{l=1\\l\neq j}}^{d}\bar{Q}_l^{q_l}(t)\Lambda_l^{g_l}(t)\right]\\
&- \sum_{j=1}^{d}\mu_j\left(\sum_{k=1}^{d} g_k b_{kj}\mathbb{E}\left[\bar{Q}_j^{q_j+1}(t)\Lambda_k^{g_k-1}(t)\prod_{\substack{l=1\\l\neq j}}^{d}\bar{Q}_l^{q_l}(t)\prod_{\substack{l=1\\l\neq k}}^{d}\bar{\Lambda}_l^{g_l}(t)\right]\right)\\
&+ \sum_{j=1}^{d}\sum_{i=1}^{d}\mu_{ij}\left(q_j\mathbb{E}\left[\prod_{l=1}^{d}\bar{Q}_l^{q_l}(t)\Lambda_l^{g_l}(t)\right] - q_i\mathbb{E}\left[\bar{Q}_i^{q_i-1}(t)\bar{Q}_j^{q_j+1}(t)\prod_{\substack{l=1\\l\neq i,j}}^{d}\bar{Q}_l^{q_l}(t)\prod_{l=1}^{d}\bar{\Lambda}_l^{g_l}(t)\right]\right)\\
&= \sum_{j=1}^{d} g_j r_j \lambda_{j,0}\mathbb{E}\left[\bar{Q}_j^{q_j}(t)\Lambda_j^{g_j-1}(t)\prod_{\substack{l=1\\l\neq j}}^{d}\bar{Q}_l^{q_l}(t)\Lambda_l^{g_l}(t)\right]\\
&+ \sum_{j=1}^{d}\mu_j\sum_{\substack{\mathbf{0}\leqslant\boldsymbol{\ell}\leqslant\boldsymbol{g}\\\|\boldsymbol{\ell}\|_1\leqslant\|\boldsymbol{g}\|_1-2}}\prod_{l=1}^{d}\binom{g_l}{\ell_l}\mathbb{E}\left[B_{lj}^{g_l-\ell_l}\right]\mathbb{E}\left[\bar{Q}_j^{q_j+1}(t)\Lambda_j^{\ell_j}(t)\prod_{\substack{k=1\\k\neq j}}^{d}\bar{Q}_k^{q_k}(t)\bar{\Lambda}_k^{\ell_k}(t)\right], \qquad (79)
\end{aligned}
$$

We can write (79) more compactly as (45). □

Proof of Theorem 10. First, (49) is immediate from (48). In order to calculate $e^{A^{(n+1)}}$, we need to exploit the structure of $A^{(n+1)}$: the superdiagonal is of the form $c_1[n:1]$, the subdiagonal of the form $c_2[1:n]$, and the diagonal of the form $c_3[0:n]+c_4[n:0] = c_4 n + (c_3-c_4)[0:n]$. Here, we write $[k:m] := \{k, k\pm1, k\pm2, \ldots, m\mp1, m\}$ for the set of integers between $k, m \in \mathbb{Z}$.

In fact, $A^{(n+1)}$ is a generalization of the Clement-Kac-Sylvester matrix. Using [13], §3, its characteristic polynomial $p_{n+1}(w) = \det\left(A^{(n+1)} - wI_{n+1}\right)$ is given by

$$
p_{n+1}(w) = \prod_{k=0}^{n}\left(-w - n\mu + \frac{n(\mu-r)}{2} + \frac{n-2k}{2}\sqrt{(\mu-r)^2+4\mu b_1}\right),
$$

hence the eigenvalues of $A^{(n+1)}$ are given by (51). Finally, formula (50) follows from Lagrange-Sylvester interpolation; see [9], Theorem 8.1.

For stability, take some $n \in \mathbb{N}$, and note that we have convergence of the moments $Z^{(n+1)}(t)$ if and only if $e^{A^{(n+1)}t} \to 0$ as $t \to \infty$, which holds if and only if

$$
\lambda_{\max}^{(n+1)} = \lambda_0^{(n+1)} = -\frac{n}{2}\left(\mu + r - \sqrt{(\mu-r)^2+4b_1}\right) = -\frac{n}{2}\left(\mu + r - \sqrt{(\mu+r)^2-4\mu(r-b_1)}\right) < 0,
$$

which in turn holds if and only if $4\mu(r-b_1) > 0$ so if and only if $b_1/r < 1$. □

Proof of Theorem 11. This can be proved by solving the system of ODEs for $n = 1$ by Theorem 10, using

$$
A^{(2)} = \begin{bmatrix} -\mu & 1 \\ \mu b_1 & -r \end{bmatrix}, \quad C^{(2)}(s) = \begin{bmatrix} 0 \\ r\lambda_0 \end{bmatrix}, \quad Z^{(2)}(0) = \begin{bmatrix} 0 \\ \lambda_0 \end{bmatrix}, \qquad (80)
$$

and letting $t \to \infty$. Alternatively, use (48), set the derivative equal to 0, and solve for $Z^{(2)}(\infty)$. □