J. Appl. Prob. **55,** 920–943 (2018)
*doi:*10.1017/jpr.2018.58
© *Applied Probability Trust* 2018

INFINITE-SERVER QUEUES WITH HAWKES INPUT

D. T. KOOPS,* ** *University of Amsterdam*
M. SAXENA *** **** AND
O. J. BOXMA,*** ***** *Eindhoven University of Technology*
M. MANDJES,* ****** *University of Amsterdam*

Abstract

In this paper we study the number of customers in infinite-server queues with a self-exciting (Hawkes) arrival process. Initially we assume that service requirements are exponentially distributed and that the Hawkes arrival process is of a Markovian nature. We obtain a system of differential equations that characterizes the joint distribution of the arrival intensity and the number of customers. Moreover, we provide a recursive procedure that explicitly identifies (transient and stationary) moments. Subsequently, we allow for non-Markovian Hawkes arrival processes and nonexponential service times. By viewing the Hawkes process as a branching process, we find that the probability generating function of the number of customers in the system can be expressed in terms of the solution of a fixed-point equation. We also include various asymptotic results: we derive the tail of the distribution of the number of customers for the case that the intensity jumps of the Hawkes process are heavy tailed, and we consider a heavy-traffic regime. We conclude by discussing how our results can be used computationally and by verifying the numerical results via simulations.

Keywords: Self-exciting process; Hawkes process; infinite-server queue; branching process; heavy-tailed distribution; heavy traffic

2010 Mathematics Subject Classification: Primary 60K25; 60G55; 60J80

1. Introduction

A common assumption in queueing theory is that the customer arrival process is a Poisson process with a deterministic rate. However, various empirical studies have revealed that arrival processes may display *overdispersion*, i.e. the variance of the number of arrivals in a given interval exceeds the corresponding expected value; see, e.g. [16] and the references therein. Overdispersion therefore indicates that the standard Poisson assumption (under which the above mentioned variance coincides with the expected value) is not valid. This has led to the study of queueing systems with overdispersed arrival processes; see [16], [17], [19], and [22]. Our aim in the current paper is to contribute to this line of research.

Received 12 September 2017; revision received 1 May 2018.
* Postal address: Korteweg-de Vries Institute, University of Amsterdam, PO Box 94248, 1090GE, Amsterdam, The Netherlands.
** Email address: d.t.koops@uva.nl
*** Postal address: Eurandom and Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600MB, Eindhoven, The Netherlands.
**** Email address: m.mayank@tuc.nl
***** Email address: o.j.boxma@tue.nl
****** Email address: m.r.h.mandjes@uva.nl

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Our object of study is the *infinite-server queue* with *Hawkes input*.

- An infinite-server queue is a service system in which each customer is taken into service immediately upon arrival. Customers are served independently of each other and have independent and identically distributed (i.i.d.) service requirements. There is a large body of literature on infinite-server queues, but typically some regularity properties are assumed such as Poisson (or renewal) arrivals.
- A Hawkes process is a point process with an *exogenous* component and an *endogenous* component. The exogenous component generates arrivals according to a homogeneous Poisson process. The endogenous component entails that arrival epochs coincide with jumps in the arrival rate. The arrival intensity subsequently behaves deterministically over time according to an *excitation function*. This type of point process was originally studied by Hawkes [13], who coined the name *self-exciting process*; nowadays the name *Hawkes process* is frequently used as well.

Hawkes processes have been used to model various phenomena, including the firing of neurons in the brain, earthquakes, criminality, and riots; see, e.g. [3] or [20] and the references therein. Very recently, Hawkes models have also been used to study trending social media; see [9], [27], and [26]. Interestingly, Daw and Pender [9] added a queueing aspect: they modeled the arrival process of visitors to a website as a Hawkes process, thus trying to capture the viral behavior of such an arrival process, taking into account that customers leave after a time which has a phase-type distribution. They studied the number of visitors on the website at any time t. Another queueing application with Hawkes processes is high-frequency transaction processes in limit order books; see [3], [7], and [28]. Limit order books are in essence queueing systems on financial markets: the books keep track of buy and sell orders that are waiting until an order arrives that matches the desired execution price above or below a certain limit. It is well known that trades tend to trigger other trades, which makes self-exciting models a natural choice.

Our work also combines queueing with Hawkes arrival processes, an area that is still largely unexplored. In, e.g. [7] and [10] scaling limits were derived for queueing systems that allow for Hawkes input. Scaling limits for infinite-server queues designed specifically for Hawkes input were derived in [10]; the authors state that exact and numerical analysis of this model is 'challenging'. To the best of the authors' knowledge, only [9] has addressed the exact (i.e. non-asymptotic) analysis of queues driven by Hawkes processes. In [9] the focus was on exact analysis of an infinite-server queue driven by an unmarked Markovian Hawkes process (i.e. the jump sizes in the arrival intensity are deterministic). The model we consider is a general version of the one studied in [9]: in our case the driving process is a *marked* Hawkes process, i.e. the intensity jumps are stochastic. In addition, we use a branching process representation to cover a non-Markovian setting in which the Hawkes intensity does not need to decay exponentially, and the job distributions are general. Furthermore, we discuss various novel asymptotic results and show how the obtained results can be used computationally.

Our work is also related to [19]; the main difference with the model studied there is that in [19] the arrival rate to the infinite-server queue is a Cox process, and, hence, not self exciting.

The paper is organized as follows. After introducing Hawkes processes in Section 2, we assume in Section 3 that the service requirements are exponentially distributed, and the excitation function is an exponentially decreasing function. As a consequence, the number of customers jointly with the Hawkes arrival rate is a Markov process; note that (marginally) the evolution of the number of customers is not Markov. We derive a partial differential equation (PDE) which characterizes the joint Laplace and z-transform of the Hawkes intensity and

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

number of customers. We show that transient moments (of arbitrary order) satisfy a specific system of ordinary differential equations (ODEs), and we provide explicit first and second transient moments. Furthermore, by using the characteristic method, we simplify the PDE to a set of ODEs that characterizes the joint distribution, which we later use for numerical analysis.

Subsequently, in Section 4, we lift the exponentiality assumptions on the excitation function and the service requirements. The price to be paid comes in the form of less explicit results. Similar to the approach followed in [15], the main idea is that the Hawkes process is represented in terms of a branching process, where in our model each node (representing a customer) in this branching process is served in an infinite-server queue. The analysis yields a fixed-point equation for the z-transform of the number of customers in the system (Section 4). By performing a finite number of iterations, we can find numerical approximations of the probability mass function (PMF) of the number of customers, akin to the approach proposed in [2]; explicit bounds on the error are also derived. A numerical example is provided in Section 5, where the methods from Sections 3 and 4 are also verified by simulations. In Section 6 the fixed-point equation also leads to new asymptotic results:

(i) in the situation that the intensity jumps of the Hawkes process are heavy-tailed, we derive the tail of the distribution of the number of customers, and

(ii) we find the asymptotics of the number of customers in the heavy-traffic regime.

Section 7 contains suggestions for further research.

2. Preliminaries

In this section we first formally define the Hawkes process (see, e.g. [3], [8], and [20] for similar descriptions). Below we consider two equivalent definitions of the Hawkes process, presented in the same generality as they are used in the remainder of this paper.

Definition 2.1. (*Conditional intensity.*) Consider a counting process $(M(t))_{t\geq 0}$, with associated filtration $(\mathcal{F}(t))_{t\geq 0}$, that satisfies

$$\mathbb{P}(M(t+\Delta t)-M(t)=m \mid \mathcal{F}(t)) = \begin{cases} \Lambda(t)\Delta t + o(\Delta t), & m=1, \\ o(\Delta t), & m>1, \\ 1-\Lambda(t)\Delta t + o(\Delta t), & m=0, \end{cases} \quad \text{as } \Delta t \downarrow 0,$$

where the conditional intensity has the form

$$\Lambda(t) = \lambda_\infty + \sum_{t_i<t} B_i h(t-t_i), \tag{2.1}$$

where $t_1, t_2, \ldots$ denote arrival epochs for a set of i.i.d. random variables B_i with a nonnegative support, and for some *reversion level* $\lambda_\infty > 0$ and some function $h\colon [0,\infty) \to [0,\infty)$ which are called the *background intensity* and *excitation function*, respectively. The summand in (2.1) is called a *kernel*. The process $M(\cdot)$ as defined above is called a *self-exciting* or *Hawkes process*.

Note that an arrival increases the future arrival intensity which, in turn, increases the probability of another arrival in the future, which explains the name 'self-exciting process'.

Remark 2.1. There exist definitions of varying generality (see, e.g. [3], [8], and [20]). For example, there are multidimensional definitions (also referred to as mutually-exciting Hawkes

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

processes), there is a distinction between *marked* and *unmarked* Hawkes processes, and the initial intensity can be taken as not equal to the reversion level. The version defined above is considered to be a marked Hawkes process with a multiplicative kernel. It is 'marked' because the kernel depends on a random variable B_i (which is called a mark) that is sampled at each event. The kernel is multiplicative, since the size of the increase B_i and the time effect $h(t-t_i)$ are multiplied. Furthermore, note that $\Lambda(0) = \lambda_\infty$ by (2.1). In general, one could consider the $\Lambda(0) = \lambda_0 \neq \lambda_\infty$ case. However, this case is hardly more general but introduces more cumbersome notation. Indeed, the additional contribution of the initial intensity can be handled independently and in the same way as the rate increase due to other events.

Next, we consider an alternative definition of Hawkes processes which is based on a representation of Hawkes processes as branching processes with immigration. The observation that this is possible was made by Hawkes and Oakes [15], and it is now standard in the literature.

Definition 2.2. (*Cluster representation.*) Let $\{B_i\}$ be a set of i.i.d. random variables with nonnegative support. Consider a (possibly infinite) $T > 0$ and define a sequence of events $\{t_m \leq T\}$ according to the following procedure.

- Consider a set of *immigrant events* $\{t_m^{(0)} \leq T\}$ that arrive according to a homogeneous Poisson process with rate λ_∞ in the interval $[0, T]$.
- Set $n = 0$. For each arrival labeled by $t_{m'}^{(n)}$, generate a sequence of next generation events $\{t_m^{(n+1)} \leq T\}$ in the interval $[t_{m'}^{(n)}, T]$ by sampling $B_{m'}$ and then sampling from the resulting Poisson process with time-dependent rate $B_{m'}h(t - t_{m'})$.
- Iterate the above rule for $n = 1, 2, \ldots$, until no more events are generated in $[0, T]$, so as to obtain the event sequence $E_n(T) := \{t_m^{(n)} \leq T\}$ for $n = 0, 1, \ldots$.

Define $M(t) := |\bigcup_{n=0}^{\infty} E_n(t)|$ as the number of events in $[0, t]$, then $(M(t))_{0\leq t\leq T}$ is called a *self-exciting* or *Hawkes process* (on the interval $[0, T]$).

In the remainder of this paper we will use both definitions to analyze infinite-server queues driven by Hawkes arrival processes. Definition 2.1 plays a key role in Section 3, in which we consider Markovian Hawkes queues. The approach in Section 3 does not apply for non-Markovian queues; in that case we resort to the interpretation as in Definition 2.2, which we discuss thoroughly in Section 4.

3. A Markovian Hawkes-fed infinite-server queue

In the previous section we introduced the Hawkes process which serves as the input process of our infinite-server queue. In this section we suppose that the arrival process is a Markovian Hawkes process, i.e. the excitation function is exponential and the service requirements are exponentially distributed. In particular, we consider the situation that $h(t) = \mathrm{e}^{-rt}$, $r > 0$, and the service requirements are $J \sim \exp(\mu)$. In addition, we assume that B (distributed as the B_i featuring in the definition of the Hawkes process) is a random variable such that $\mathbb{P}(B > 0) = 1$. The Hawkes process acts as an input process to an infinite-server system, i.e. we could call this model Hawkes/M/∞ in Kendall's notation. In Subsection 3.1 we characterize the joint transform of the Hawkes intensity $\Lambda(t)$ and the number of customers $N(t)$ of the Hawkes/M/∞ queue at time t in terms of the solution of an ODE. In Subsection 3.2 we develop a recursive procedure that yields the transient moments of $(\Lambda(t), N(t))$. Steady-state moments are briefly discussed in Subsection 3.3.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

3.1. Characterization of the queueing process

In this section our main objective is to characterize the double transform

$$\zeta(t, z, s) := \mathbb{E}z^{N(t)}\mathrm{e}^{-s\Lambda(t)},$$

which uniquely defines the joint transient distribution of $(N(t), \Lambda(t))$.

Theorem 3.1. *Let the arrival process be a Markovian Hawkes process. Then, given* $\Lambda(0) = \lambda_\infty$ *and* $N(0) = 0$,

$$\zeta(t, z, s) = \mathrm{e}^{-s(t)\lambda_\infty} \exp\left(-\lambda_\infty r \int_0^t s(u)\,\mathrm{d}u\right), \tag{3.1}$$

where, with $\beta(s) := \mathbb{E}\mathrm{e}^{-sB}$, $s(\cdot)$ *solves the ODE*

$$s'(u) + rs(u) + (1 + (z-1)\mathrm{e}^{-\mu u})\beta(s(u)) - 1 = 0, \qquad 0 \le u \le t, \tag{3.2}$$

with boundary condition $s(0) = s$.

Remark 3.1. Note that this result is closely related to results about Hawkes counting processes. For example, a version of [8, Theorem 3.1] is retrieved when $\mu = 0$ is substituted in (3.2) (note: set $\rho = 0$ in [8], as we do not consider a shot-noise background process here). Note that $\mu = 0$ corresponds to infinitely long service times, and, hence, the number-of-customers process reduces to a counting process as studied in [8]. In the situation where $\mu = 0$, it is possible to derive an explicit equation for the probability generating function of the counting process by separation of variables. Unfortunately, in our situation the variables $s(u)$ and u in (3.2) are not separable due to the additional factor $\mathrm{e}^{-\mu u}$, and, hence, we cannot solve it analytically.

Proof of Theorem 3.1. We derive the joint distribution of $(N(t), \Lambda(t))$. Define

$$F(t, k, \lambda) = \mathbb{P}(N(t) = k, \Lambda(t) \le \lambda), \qquad f(t, k, \lambda) = \frac{\partial F(t, k, \lambda)}{\partial \lambda}.$$

Considering the evolution of the Markovian system between t and $t + \Delta t$ yields

$$\begin{aligned} F(t + \Delta t, k, \lambda - r(\lambda - \lambda_\infty)\Delta t) = {} & \int_0^\lambda y\Delta t\mathbb{P}(B \le \lambda - y) f(t, k-1, y)\,\mathrm{d}y \\ & + (k+1)\mu\Delta t F(t, k+1, \lambda) + F(t, k, \lambda)(1 - k\mu\Delta t) \\ & - \int_0^\lambda y\Delta t f(t, k, y)\,\mathrm{d}y. \end{aligned}$$

After elementary manipulations and letting $\Delta t \downarrow 0$, it follows that

$$\begin{aligned} \frac{\partial F(t, k, \lambda)}{\partial t} - r(\lambda - \lambda_\infty)\frac{\partial F(t, k, \lambda)}{\partial \lambda} = {} & \int_0^\lambda y\mathbb{P}(B \le \lambda - y) f(t, k-1, y)\,\mathrm{d}y \\ & + (k+1)\mu F(t, k+1, \lambda) - k\mu F(t, k, \lambda) \\ & - \int_0^\lambda y f(t, k, y)\,\mathrm{d}y. \end{aligned}$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Since we assumed that $\mathbb{P}(B \le 0) = 0$, differentiating with respect to λ yields

$$\frac{\partial f(t,k,\lambda)}{\partial t} - \frac{\partial}{\partial \lambda}[r\lambda f(t,k,\lambda)] + r\lambda_\infty \frac{\partial}{\partial \lambda} f(t,k,\lambda) = \int_0^\lambda y f(t,k-1,y)\,\mathrm{d}_y \mathbb{P}(B \le \lambda - y) + (k+1)\mu f(t,k+1,\lambda) - (k\mu + \lambda) f(t,k,\lambda). \tag{3.3}$$

The next step consists of transforming (3.3) with respect to the Hawkes intensity λ, and to that end we define the transform

$$\xi(t,k,s) := \int_0^\infty \mathrm{e}^{-s\lambda} f(t,k,\lambda)\,\mathrm{d}\lambda.$$

After transformation, we find that

$$\frac{\partial \xi(t,k,s)}{\partial t} + rs\frac{\partial \xi(t,k,s)}{\partial s} + rs\lambda_\infty \xi(t,k,s) = -\frac{\partial \xi(t,k-1,s)}{\partial s}\beta(s) + (k+1)\mu\xi(t,k+1,s) - k\mu\xi(t,k,s) + \frac{\partial \xi(t,k,s)}{\partial s},$$

which we can rewrite as the PDE

$$\frac{\partial \xi(t,k,s)}{\partial t} + (rs-1)\frac{\partial \xi(t,k,s)}{\partial s} + \frac{\partial \xi(t,k-1,s)}{\partial s}\beta(s) + rs\lambda_\infty \xi(t,k,s) = (k+1)\mu\xi(t,k+1,s) - k\mu\xi(t,k,s).$$

Next, we transform this equation in terms of the number of customers variable k, for which we use the transform

$$\zeta(t,z,s) = \sum_{k=0}^\infty z^k \xi(t,k,s) = \mathbb{E} z^{N(t)} \mathrm{e}^{-s\Lambda(t)}.$$

This yields

$$\frac{\partial \zeta(t,z,s)}{\partial t} + (rs + z\beta(s) - 1)\frac{\partial \zeta(t,z,s)}{\partial s} + \mu(z-1)\frac{\partial \zeta(t,z,s)}{\partial z} = -rs\lambda_\infty \zeta(t,z,s). \tag{3.4}$$

Now let s and z be parametrized by u. Then the characteristic equations are

$$-s'(u) + rs(u) + z(u)\beta(s(u)) - 1 = 0, \tag{3.5}$$

$$-z'(u) + \mu(z(u) - 1) = 0 \tag{3.6}$$

with the boundary conditions $s(t) = s$ and $z(t) = z$. The solution of (3.6) is

$$z(u) = 1 + C\mathrm{e}^{\mu u}$$

with C determined by $z(t) = z = C\mathrm{e}^{\mu t} + 1$, i.e. $C = (z-1)\mathrm{e}^{-\mu t}$. We thus find that $z(u) = 1 + (z-1)\mathrm{e}^{-\mu(t-u)}$, which we can substitute into (3.5) so as to obtain the ODE

$$-s'(u) + rs(u) + (1 + (z-1)\mathrm{e}^{-\mu(t-u)})\beta(s(u)) - 1 = 0.$$

By substituting t for $t - u$, we obtain the ODE

$$s'(t) + rs(t) + (1 + (z-1)\mathrm{e}^{-\mu t})\beta(s(t)) - 1 = 0$$

with the boundary condition $s(0) = s$. The result follows. □

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Numerically, one can obtain the PMF of $N(t)$ by first solving the differential equation, and then applying a Fourier inversion algorithm. Numerical results can be found in Section 5. Another powerful feature of Theorem 3.1 is that it allows us to obtain moments, as is presented in Subsections 3.2 and 3.3.

3.2. Transient moments

In this section we discuss the computation of the joint transient moments of the type $\mathbb{E}\Lambda^g(t)N^q(t)$ for some integers g and q under the additional assumption that $\mathbb{E}B^g < \infty$. The marginal moments of the Hawkes process are known and can be found in, e.g. [8, Lemma 3.1, Theorem 3.6], but we include them here for completeness. The key idea is to use PDE (3.4) to derive ODEs for the joint moments. To this end, we write (3.4) as

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}z^{N(t)}\mathrm{e}^{-s\Lambda(t)} - (rs + z\beta(s) - 1)\mathbb{E}\Lambda(t)z^{N(t)}\mathrm{e}^{-s\Lambda(t)} + \mu(z-1)\mathbb{E}N(t)z^{N(t)-1}\mathrm{e}^{-s\Lambda(t)}\\ &\quad = -rs\lambda_\infty \mathbb{E}z^{N(t)}\mathrm{e}^{-s\Lambda(t)}. \end{aligned} \tag{3.7}$$

We begin by differentiating the above equation $g \in \mathbb{N}$ times with respect to s and then inserting $s = 0$, which expresses the $(g+1)$th moment of $\Lambda(t)$ in terms of the first up to the gth moment of $\Lambda(t)$ through the following ODE (where we have assumed that $\mathbb{E}B^g < \infty$):

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Lambda^g(t)z^{N(t)} + g(r - z\mathbb{E}B)\mathbb{E}\Lambda^g(t)z^{N(t)} + \mu(z-1)\mathbb{E}\Lambda^g(t)N(t)z^{N(t)-1}\\ &\quad = (z-1)\mathbb{E}\Lambda^{g+1}(t)z^{N(t)} + \mathbf{1}_{\{g\geq 1\}}\, g\lambda_\infty r\mathbb{E}\Lambda^{g-1}(t)z^{N(t)}\\ &\qquad + z\,\mathbf{1}_{\{g\geq 2\}}\sum_{j=0}^{g-2}\binom{g}{j}\mathbb{E}B^{g-j}\mathbb{E}\Lambda^{j+1}(t)z^{N}(t), \qquad g = 0, 1, \ldots, \end{aligned} \tag{3.8}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

So as to obtain a relation between the joint moments and the moments that correspond to the number of customers, we differentiate (3.8) $q+1$ times with respect to z and insert $z = 1$. This gives us the key relation, for $g = 0, 1, \ldots, q = 0, 1, \ldots,$

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Lambda^g(t)\bar{N}^q(t) + ((q+1)\mu + g(r - \mathbb{E}B))\mathbb{E}\Lambda^g(t)\bar{N}^q(t)\\ &\quad = (q+1)\mathbb{E}\Lambda^{g+1}(t)\bar{N}^{q-1}(t) + \mathbf{1}_{\{g\geq 1\}}\, g\lambda_\infty r\mathbb{E}\Lambda^{g-1}(t)\bar{N}^q(t)\\ &\qquad + (q+1)g\mathbb{E}B\mathbb{E}\Lambda^g(t)\bar{N}^{q-1}(t)\\ &\qquad + \mathbf{1}_{\{g\geq 2\}}\sum_{j=0}^{g-2}\binom{g}{j}\mathbb{E}B^{g-j}[(q+1)\mathbb{E}\Lambda^{j+1}(t)\bar{N}^{q-1}(t) + \mathbb{E}\Lambda^{j+1}(t)\bar{N}^q(t)], \end{aligned} \tag{3.9}$$

where $\bar{N}^q(t) := N(t)(N(t)-1)\cdots(N(t)-q)$ and $\bar{N}^{-1}(t) := 1$.

Now, by (3.9), we can obtain a system of first-order ODEs. To this end, we substitute into (3.9) a combination of indices $g := k$ and $q := q-k$ for values of $k \in \{0, \ldots, q+1\}$. Denoting

$$Z^{(q+2)}(t) := [\mathbb{E}\bar{N}^q(t), \mathbb{E}\Lambda(t)\bar{N}^{q-1}(t), \ldots, \mathbb{E}\Lambda^q(t)\bar{N}^0(t), \mathbb{E}\Lambda^{q+1}(t)]^\top,$$

it follows that the vector $Z^{(q+2)}(t)$, $q = 0, 1, 2, \ldots,$ satisfies the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}Z^{(q+2)}(t) = A_1^{(q+2)}Z^{(q+2)}(t) + A_0^{(q+2)} \tag{3.10}$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

with

$$A_1^{(q+2)} = \begin{bmatrix} -d_0^{(q+1)} & q+1 & 0 & \cdots & 0 & 0 \\ 0 & -d_1^{(q)} & q & \cdots & 0 & 0 \\ 0 & 0 & -d_2^{(q-1)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -d_q^{(1)} & 1 \\ 0 & 0 & 0 & \cdots & 0 & -d_{q+1}^{(0)} \end{bmatrix}, \qquad A_0^{(q+2)} = \begin{bmatrix} b_0^{(q)} \\ b_1^{(q-1)} \\ b_2^{(q-2)} \\ \vdots \\ b_q^{(0)} \\ b_{q+1}^{(-1)} \end{bmatrix},$$

where, for $k \in \{0, \dots, q+1\}$,

$$\begin{aligned} d_k^{(q+1-k)} &:= (q+1-k)\mu + k(r - \mathbb{E}B), \\ b_k^{(q-k)} &:= \mathbf{1}_{\{k\geq 1\}}\, k\lambda_\infty r \mathbb{E}\Lambda^{k-1}(t)\bar{N}^{q-k}(t) + (q-k+1)k\mathbb{E}B\mathbb{E}\Lambda^k(t)\bar{N}^{q-k-1}(t) \\ &\quad + \mathbf{1}_{\{k\geq 2\}} \sum_{j=0}^{k-2} \binom{k}{j} \mathbb{E}B^{k-j}[(q-k+1)\mathbb{E}\Lambda^{j+1}(t)\bar{N}^{q-k-1}(t) \\ &\qquad\qquad + \mathbb{E}\Lambda^{j+1}(t)\bar{N}^{q-k}(t)]. \end{aligned}$$

Note that $d_0^{(q+1)}, \dots, d_{q+1}^{(0)}$ are the (distinct) eigenvalues of the matrix $A_1^{(q+2)}$.

Proposition 3.1. *The solution of ODE (3.10) is*

$$Z^{(q+2)}(t) = \mathrm{e}^{A_1^{(q+2)}t} Z^{(q+2)}(0) + \int_0^t \mathrm{e}^{A_1^{(q+2)}(t-s)} A_0^{(q+2)}\, \mathrm{d}s, \tag{3.11}$$

where, with I a $(q+2)\times(q+2)$ identity matrix,

$$\mathrm{e}^{A_1^{(q+2)}t} = \sum_{i=1}^{q+2} \mathrm{e}^{(d_{i-1}^{(q-i+2)})t} \prod_{j=1,\, j\neq i}^{q+2} \frac{A_1^{(q+2)} - d_{j-1}^{(q-j+2)} I}{d_{i-1}^{(q-i+2)} - d_{j-1}^{(q-j+2)}}. \tag{3.12}$$

Proof. Equation (3.11) follows from (3.10) in a straightforward way. The exponential of the matrix is computed using the interpolation-based formula from [11, p. 101, Equation (4.18)], completing the proof. □

Corollary 3.1. *If $r \neq b_1 := \mathbb{E}B$ and with $r_0 := r - b_1$, it holds that*

$$\mathbb{E}\Lambda(t) = \frac{\lambda_\infty r}{r_0} - \frac{\lambda_\infty b_1}{r_0}\mathrm{e}^{-r_0 t}, \qquad \mathbb{E}N(t) = \frac{\lambda_\infty r}{\mu r_0} - \frac{\lambda_\infty b_1}{r_0(\mu - r_0)}\mathrm{e}^{-r_0 t} - \frac{\lambda_\infty (r-\mu)}{\mu(\mu - r_0)}\mathrm{e}^{-\mu t}$$

with an obvious adaptation when $\mu = r_0$.

Proof. To obtain the first moments of the model, we can use Proposition 3.1 for the $q = 0$ case, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} Z^{(2)}(t) = A_1^{(2)} Z^{(2)}(t) + A_0^{(2)} \tag{3.13}$$

with

$$Z^{(2)}(t) = \begin{bmatrix} \mathbb{E}N(t) \\ \mathbb{E}\Lambda(t) \end{bmatrix}, \qquad A_1^{(2)} = \begin{bmatrix} -\mu & 1 \\ 0 & -r_0 \end{bmatrix}, \qquad A_0^{(2)} = \begin{bmatrix} 0 \\ \lambda_\infty r \end{bmatrix}.$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

We know the solution of the general ODE (3.10) from (3.11). Therefore, for the $q = 0$ case, we obtain the solution to ODE (3.13), i.e.

$$\begin{bmatrix} \mathbb{E}N(t) \\ \mathbb{E}\Lambda(t) \end{bmatrix} = \mathrm{e}^{A_1^{(2)}t} \begin{bmatrix} \mathbb{E}N(0) \\ \mathbb{E}\Lambda(0) \end{bmatrix} + \int_0^t \mathrm{e}^{A_1^{(2)}(t-s)} A_0^{(2)} \, \mathrm{d}s, \tag{3.14}$$

where $(\mathbb{E}N(0), \mathbb{E}\Lambda(0)) = (0, \lambda_\infty)$. It now remains to evaluate $\mathrm{e}^{A_1^{(2)}t}$. By virtue of (3.12), we find that

$$\mathrm{e}^{A_1^{(2)}t} = \begin{bmatrix} \mathrm{e}^{-\mu t} & \dfrac{\mathrm{e}^{-\mu t} - \mathrm{e}^{-r_0 t}}{-\mu + r_0} \\ 0 & \mathrm{e}^{-r_0 t} \end{bmatrix}.$$

Substituting this exponential matrix into (3.14) yields the desired results. □

We now explicitly state the second transient moments. Define $b_2 \coloneqq \mathbb{E}B^2$.

Corollary 3.2. *If $r_0 \neq 0$ then, for all $t \geq 0$,*

$$\begin{aligned}
\mathbb{E}\Lambda^2(t) &= \frac{\lambda_\infty r(b_2 + 2\lambda_\infty r)}{2r_0^2} - \frac{\lambda_\infty b_1(b_2 + 2\lambda_\infty r)}{r_0^2}\mathrm{e}^{-r_0 t} + D_1 \mathrm{e}^{-2r_0 t}, \\
\mathbb{E}\Lambda(t)N(t) &= \frac{\lambda_\infty r}{r_0(\mu + r_0)}\left(\frac{b_2 + 2\lambda_\infty r}{2r_0} + \frac{\lambda_\infty r + \mu b_1}{\mu}\right) - \frac{\lambda_\infty^2 r(r-\mu)}{\mu r_0(\mu - r_0)}\mathrm{e}^{-\mu t} \\
&\quad - \frac{\lambda_\infty b_1}{\mu r_0}\left(b_1 + \frac{b_2 + 2\lambda_\infty r}{r_0} + \frac{\lambda_\infty r}{\mu - r_0}\right)\mathrm{e}^{-r_0 t} + \frac{D_1}{\mu - r_0}\mathrm{e}^{-2r_0 t} + D_2 \mathrm{e}^{-(\mu + r_0)t}, \\
\mathbb{E}N^2(t) &= \frac{\lambda_\infty r}{\mu r_0(\mu + r_0)}\left(\frac{b_2 + 2\lambda_\infty r}{2r_0} + \frac{\mu(\mu + r) + \lambda_\infty r}{\mu}\right) \\
&\quad - \frac{\lambda_\infty(r-\mu)(\mu r_0 + 2\lambda_\infty r)}{\mu^2 r_0(\mu - r_0)}\mathrm{e}^{-\mu t} \\
&\quad - \frac{2\lambda_\infty b_1}{\mu r_0(2\mu - r_0)}\left(b_1 + \frac{b_2 + 2\lambda_\infty r}{r_0} + \frac{\lambda_\infty r}{\mu - r_0} + \frac{\mu}{2}\frac{2\mu - r_0}{\mu - r_0}\right)\mathrm{e}^{-r_0 t} \\
&\quad + \frac{D_1}{(\mu - r_0)^2}\mathrm{e}^{-2r_0 t} + \frac{2D_2}{\mu - r_0}\mathrm{e}^{-(\mu + r_0)t} + D_3 \mathrm{e}^{-2\mu t},
\end{aligned}$$

where the constants D_1, D_2, and D_3 follow from the initial conditions, i.e. the requirements that $\mathbb{E}N^2(0) = 0$, $\mathbb{E}\Lambda(0)N(0) = 0$, and $\mathbb{E}\Lambda^2(0) = \lambda_\infty^2$.

Proof. To obtain all second moments, we again use Proposition 3.1 for $q = 1$ (with $\bar{N}^0(t) = N(t)$). This yields

$$\frac{\mathrm{d}}{\mathrm{d}t} Z^{(3)}(t) = A_1^{(3)} Z^{(3)}(t) + A_0^{(3)}(t) \tag{3.15}$$

with

$$Z^{(3)}(t) = \begin{bmatrix} \mathbb{E}\bar{N}^1(t) \\ \mathbb{E}\Lambda(t)N(t) \\ \mathbb{E}\Lambda^2(t) \end{bmatrix}, \qquad A_1^{(3)} = \begin{bmatrix} -2\mu & 2 & 0 \\ 0 & -\mu - r_0 & 1 \\ 0 & 0 & -2r_0 \end{bmatrix},$$

$$A_0^{(3)}(t) = \begin{bmatrix} 0 \\ b_1 \mathbb{E}\Lambda(t) + \lambda_\infty r \mathbb{E}N(t) \\ (b_2 + 2\lambda_\infty r)\mathbb{E}\Lambda(t) \end{bmatrix}.$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Again, using (3.11) for the $q = 1$ case yields the solution to ODE (3.15), i.e.

$$\begin{bmatrix} \mathbb{E}\bar{N}^1(t) \\ \mathbb{E}\Lambda(t)N(t) \\ \mathbb{E}\Lambda^2(t) \end{bmatrix} = \mathrm{e}^{A_1^{(3)}t} \begin{bmatrix} \mathbb{E}\bar{N}^1(0) \\ \mathbb{E}\Lambda(0)N(0) \\ \mathbb{E}\Lambda^2(0) \end{bmatrix} + \int_0^t \mathrm{e}^{A_1^{(3)}(t-s)} A_0^{(3)}(s)\,\mathrm{d}s,$$

where

$$\begin{bmatrix} \mathbb{E}\bar{N}^1(0) \\ \mathbb{E}\Lambda(0)N(0) \\ \mathbb{E}\Lambda^2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \lambda_\infty^2 \end{bmatrix}.$$

Now we only need to calculate matrix $\mathrm{e}^{A_1^{(3)}t}$ to obtain the complete solution of (3.15). Using (3.12), we have

$$\mathrm{e}^{A_1^{(3)}t} = \begin{bmatrix} \mathrm{e}^{-2\mu t} & \dfrac{2\mathrm{e}^{-\mu t}(\mathrm{e}^{-r_0 t} - \mathrm{e}^{-\mu t})}{\mu - r_0} & \dfrac{(\mathrm{e}^{-r_0 t} - \mathrm{e}^{-\mu t})^2}{(\mu - r_0)^2} \\ 0 & \mathrm{e}^{-(\mu + r_0)t} & \dfrac{\mathrm{e}^{-r_0 t}(\mathrm{e}^{-r_0 t} - \mathrm{e}^{-\mu t})}{\mu - r_0} \\ 0 & 0 & \mathrm{e}^{-2r_0 t} \end{bmatrix}.$$

For $\mu = r_0$, this expression has to be adapted using the L'Hôpital rule. Since we now know the first factorial moment $\mathbb{E}\bar{N}^1(t)$ and the first moment $\mathbb{E}N(t)$, the relation $\mathbb{E}N^2(t) = \mathbb{E}\bar{N}^1(t) + \mathbb{E}N(t)$ provides the second transient moment. We do not write the constants D_1, D_2, and D_3 explicitly, since their expressions are rather lengthy. □

Remark 3.2. Note that in Corollaries 3.1 and 3.2, the speed of convergence to the stationary moments of $N(\infty)$, and of $\mathrm{cov}(\Lambda(\infty), N(\infty))$, is determined by the minimum of r_0 and μ.

Remark 3.3. The results in Corollaries 3.1 and 3.2 are stated under the condition that $r \neq b_1$. The case in which $r = b_1$ (i.e. $r_0 = 0$) can be derived from those results, if one takes the limit $r \to b_1$. From the expressions in Corollary 3.2, note that if $r - b_1 = r_0 < 0$ then $\mathbb{E}N(t) \to \infty$ as $t \to \infty$. It turns out that the condition $r - b_1 = r_0 > 0$ leads to finite moments, and can be interpreted as the model's stability condition. This stability condition could also have been derived *a priori* by considering the Hawkes process as a branching process (as in Definition 2.2). From branching process theory it is known that a 'population' (or in our case, a cluster) exterminates with probability 1 if and only if the average number of children per parent is less than unity; see [12]. In this case, this translates to the condition

$$b_1 \int_0^\infty h(s)\,\mathrm{d}s = b_1 \int_0^\infty \mathrm{e}^{-rs}\,\mathrm{d}s = \frac{b_1}{r} < 1.$$

If this does not hold then in the long run there will be more and more clusters, which implies that the intensity and, hence, the number of customers will grow indefinitely on average. Observe that the value of the service rate μ does not affect stability as long as $\mu > 0$.

3.3. Stationary moments

In this subsection we consider stationary moments. To this end, we assume that the queueing process is stable, i.e. $b_1 < r$; see Remark 3.3. Stationary moments can be easily derived from the previous section by taking the limit $t \to \infty$. Define by Λ and N the stationary versions of $\Lambda(t)$ and $N(t)$, respectively.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Corollary 3.3. *If $\mathbb{E}B^g < \infty$ then the gth moment of the stationary Hawkes intensity is*

$$\mathbb{E}\Lambda^g = \frac{1}{gr_0}\left[\mathbf{1}_{\{g\geq 1\}}\, g\lambda_\infty r\mathbb{E}\Lambda^{g-1} + \mathbf{1}_{\{g\geq 2\}}\sum_{j=0}^{g-2}\binom{g}{j}\mathbb{E}B^{g-j}\mathbb{E}\Lambda^{j+1}\right], \qquad g = 1, 2, 3, \ldots.$$

Proof. Considering (3.8) in the steady-state case and then taking $z = 1$ yields the desired result. □

Corollary 3.4. *Let b_1 and, where needed, b_2 be finite. The stationary means and (co-)variances of Λ and N are*

$$\mathbb{E}\Lambda = \frac{\lambda_\infty r}{r_0}, \qquad \mathbb{E}N = \frac{\lambda_\infty r}{\mu r_0}, \qquad \operatorname{var}(\Lambda) = \frac{\lambda_\infty r b_2}{2r_0^2},$$
$$\operatorname{var}(N) = \frac{\lambda_\infty r}{2r_0^2}\frac{b_2 + 2(\mu + r)r_0}{\mu(\mu + r_0)}, \qquad \operatorname{cov}(N, \Lambda) = \frac{\lambda_\infty r}{2r_0^2}\frac{b_2 + 2r_0 b_1}{\mu + r_0}.$$

In particular, the correlation coefficient of N and Λ is

$$\rho(N, \Lambda) = \frac{b_2 + 2r_0 b_1}{\sqrt{b_2(b_2 + 2r_0(\mu + r))}}\sqrt{\frac{\mu}{\mu + r_0}}.$$

Proof. This follows directly from taking $t \to \infty$ in Corollaries 3.1 and 3.2. □

4. A non-Markovian Hawkes-fed infinite-server queue

In this section we allow the excitation function h to be a general nonnegative function, and we allow J to be any nonnegative random variable. We derive a fixed-point equation for the z-transform of $N(t)$ in Theorem 4.1, which can be iterated so as to obtain an (increasingly accurate) approximation; in Proposition 4.1 we derive error bounds of the resulting approximation scheme.

In this section we rely heavily on the representation of a Hawkes process as a branching process (recall Definition 2.2). Immigrants arrive according to a Poisson process with rate λ_∞. Each of these immigrants increases the future arrival rate. The arrivals that occur due to this increase are called the *children* of the immigrant. In turn, those children are potentially the parents of a next generation, and so forth. Since the customers enter an infinite-server queue, each of them is served independently of the rest with i.i.d. service requirements (distributed as a nonnegative random variable J).

Note that, by definition, each parent produces children independently from all other parents at a rate $Bh(u)$ at time u after its own birth, conditional on its own arrival rate increment B. Let $S(u)$ denote the number of children of a parent, u time units after its own birth, including the parent itself if it is still being served. Let its probability generating function be denoted by $\eta(u, z) := \mathbb{E}z^{S(u)}$ for $0 \leq u \leq t$. Then the number of jobs in the system at time t, $N(t)$ with $N(0) = 0$, satisfies

$$\mathbb{E}z^{N(t)} = \sum_{n=0}^{\infty}\frac{(\lambda_\infty t)^n}{n!}\mathrm{e}^{-\lambda_\infty t}\left(\frac{1}{t}\int_0^t \eta(u, z)\,\mathrm{d}u\right)^n = \exp\left(\lambda_\infty\int_0^t(\eta(u, z) - 1)\,\mathrm{d}u\right), \quad (4.1)$$

which follows by conditioning on the number of immigrants (i.e. the number of clusters). The next step is to identify $\eta(u, z)$ by studying each cluster separately. First, consider the

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

distributional equality, for $0 \le u \le t$,

$$S(u) \stackrel{\mathrm{D}}{=} \mathbf{1}_{\{J>u\}} + \sum_{i=1}^{K(u)} S^{(i)}(u - t_i), \tag{4.2}$$

where $S(u) \stackrel{\mathrm{D}}{=} S^{(i)}(u)$ for all i, $K(\cdot)$ is an inhomogeneous Poisson counting process with rate $B_i h(\cdot)$ (conditional on B_i) that counts the number of children, and $t_1, t_2, \ldots$ are the birth times of the corresponding children. Note that $S^{(i)}(u)$ can be interpreted as the number of children of child i (including itself if it is still in the system, u time units after its birth). Denote by $P_t(s)$ the probability that, conditional on the fact that a child was born before time t, it was already born before time s. Then it holds that

$$\begin{aligned} P_t(s) &= \frac{\mathbb{P}(K(s) = 1, K(t) - K(s) = 0)}{\mathbb{P}(K(t) = 1)} \\ &= \frac{\exp(-\int_0^s Bh(u)\,\mathrm{d}u) \int_0^s Bh(u)\,\mathrm{d}u \exp(-\int_s^t Bh(u)\,\mathrm{d}u)}{\exp(-\int_0^t Bh(u)\,\mathrm{d}u) \int_0^t Bh(u)\,\mathrm{d}u} \\ &= \frac{\int_0^s h(u)\,\mathrm{d}u}{\int_0^t h(u)\,\mathrm{d}u}. \end{aligned}$$

Defining $H(u) := \int_0^u h(v)\,\mathrm{d}v$, the corresponding probability density function is thus given by

$$p_u(s) := P_u'(s) = \frac{h(s)}{\int_0^u h(v)\,\mathrm{d}v} = \frac{h(s)}{H(u)}.$$

Define $\mathcal{J}(u) := \mathbb{P}(J > u)$ and $\bar{\mathcal{J}}(u) := 1 - \mathcal{J}(u)$. Then, it follows that

$$\begin{aligned} \eta(u, z) &= \mathbb{E}_B\left[\sum_{n=0}^{\infty} \mathbb{E}[z^{S(u)} \mid K(u) = n, B]\mathbb{P}(K(u) = n \mid B)\right] \\ &= (\bar{\mathcal{J}}(u) + \mathcal{J}(u)z)\mathbb{E}_B\left[\sum_{n=0}^{\infty} \mathbb{E}[z^{S^{(1)}(u)}]^n \mathrm{e}^{-BH(u)} \frac{B^n H(u)^n}{n!}\right] \\ &= (\bar{\mathcal{J}}(u) + \mathcal{J}(u)z)\mathbb{E}_B\left[\sum_{n=0}^{\infty} \left(\int_0^u p_u(s)\eta(u - s, z)\,\mathrm{d}s\right)^n \mathrm{e}^{-BH(u)} \frac{B^n H(u)^n}{n!}\right] \\ &= (\bar{\mathcal{J}}(u) + \mathcal{J}(u)z)\mathbb{E}_B \exp\left(B \int_0^u h(s)(\eta(u - s, z) - 1)\,\mathrm{d}s\right). \end{aligned}$$

Recognizing the Laplace–Stieltjes transform (LST) of B, we can write

$$\eta(u, z) = (\bar{\mathcal{J}}(u) + \mathcal{J}(u)z)\beta\left(\int_0^u h(s)(1 - \eta(u - s, z))\,\mathrm{d}s\right). \tag{4.3}$$

Let $\mathcal{G}$ be the class of all time-dependent z-transforms $f(u, z) := \mathbb{E}z^{X(u)}$, where $X(u)$ for $u \in [0, t]$ is a nonnegative discrete random variable residing on the integers with a possibly defective distribution (i.e. the probabilities may sum to strictly less than 1). For $f \in \mathcal{G}$, define the functional ϕ by

$$\phi(f)(u, z) = (\bar{\mathcal{J}}(u) + \mathcal{J}(u)z)\beta\left(\int_0^u h(s)(1 - f(u - s, z))\,\mathrm{d}s\right). \tag{4.4}$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

With this definition, (4.3) can be summarized as $\eta = \phi(\eta)$, i.e. η can be seen as a fixed point of ϕ. We stress that the following result involves *complex* z. This is important, because for the numerical analysis in Section 5 we need to iterate ϕ a finite number of times for a particular set of complex-valued arguments. Furthermore, we will use the following notation. For a generic $f_0 \in \mathcal{G}$, we define $f_{n+1} := \phi(f_n)$ for $n = 0, 1, \ldots$.

Theorem 4.1. *Suppose that* $\mathbb{E}B\int_0^\infty h(s)\,\mathrm{d}s < \infty$. *Let* $z \in \mathbb{C}$, *such that* $|z| < 1$, *be fixed. Then the following hold:*

(i) *the* z*-transform of* $N(t)$ *is given by (4.1);*

(ii) *the function* η, *defined by* $\eta(u, z) = \mathbb{E}z^{S(u)}$ *for* $u \in [0, t]$ *is the unique fixed point of* ϕ, *i.e.* $\phi(\eta) = \eta$, *with* ϕ *defined in (4.4);*

(iii) *the sequence* $(f_n)_n$ *has the property that* $f_n \in \mathcal{G}$ *for every* n *if* $f_0 \in \mathcal{G}$;

(iv) f_n *converges pointwise to* $\eta \in \mathcal{G}$, *regardless of the initial approximation* $f_0 \in \mathcal{G}$.

Proof. First fix $-1 < z < 1$, where z is a real number. Note that (i) and $\phi(\eta) = \eta$ (see (4.3)) have already been proven. To prove (ii), it remains to show uniqueness, which will be shown after the proof of (iii) and (iv). It holds that $f_0 \in \mathcal{G}$ implies $f_1 = \phi(f_0) \in \mathcal{G}$ following a similar probabilistic reasoning as in, e.g. [2, Theorem 1]. Indeed, consider the related operator $\tilde{\phi}$, which maps a possibly defective cumulative distribution function (CDF) F to

$$\tilde{\phi}(F)(u, k) = \mathbb{P}\left(\mathbf{1}_{\{J>u\}} + \sum_{i=1}^{K(u)} S^{(i)}(u - t_i) \le k\right), \qquad 0 \le u \le t; \tag{4.5}$$

see also (4.2), where the $S^{(i)}$ are i.i.d. with CDF F. Let $f_0 \in \mathcal{G}$ be the z-transform of a possibly defective CDF F. Note that $\phi(f_0)$ is then the z-transform of the CDF $\tilde{\phi}(F)$ by the construction of ϕ and $\tilde{\phi}$. Statement (iii) now follows by induction.

Now we show that, for arbitrary $f_0, g_0 \in \mathcal{G}$, f_n and g_n have a limit, which is, in fact, the same for both sequences. To do this, we show by induction that, for every $n \in \mathbb{N}$,

$$|f_n(u, z) - g_n(u, z)| \le \frac{1}{n!}(Cu)^n, \qquad 0 \le u \le t. \tag{4.6}$$

We first prove the base case $n = 1$, i.e. for every $0 \le u \le t$,

$$\begin{aligned}
&|f_1(u, z) - g_1(u, z)| \\
&\quad \le \mathbb{E}\left|\exp\left(-B\int_0^u h(s)(1 - f_0(u - s, z))\,\mathrm{d}s\right) - \exp\left(-B\int_0^u h(s)(1 - g_0(u - s, z))\,\mathrm{d}s\right)\right| \\
&\quad \le \mathbb{E}B\int_0^u h(u - s)|f_0(s, z) - g_0(s, z)|\,\mathrm{d}s \\
&\quad \le \mathbb{E}B\int_0^u h(s)\,\mathrm{d}s\int_0^u |f_0(s, z) - g_0(s, z)|\,\mathrm{d}s \\
&\quad \le Cu,
\end{aligned} \tag{4.7}$$

with $C = 2\mathbb{E}B\int_0^\infty h(s)\,\mathrm{d}s$, which is finite by assumption. The first inequality of (4.7) follows from (4.4) and $|\bar{\mathcal{J}}(u) + \mathcal{J}(u)z| \le 1$ for $|z| < 1$. The second inequality follows by the mean-value theorem, which implies that, for $x, y \in \mathbb{R}$, $\mathrm{e}^x - \mathrm{e}^y = (x - y)\mathrm{e}^{z(x,y)}$, where $z(x, y)$ takes a value between x and y. In our application, the exponent on the right-hand side can be bounded from above by 1, since $f_0, g_0 \in \mathcal{G}$ imply that $f_0, g_0 \le 1$, and, hence, the

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

exponential is bounded by $e^0 = 1$. Finally, the third inequality follows from Young's inequality for convolutions (see [6, Theorem 3.9.4]), and the fourth inequality from $|f_0 - g_0| \le 2$ for $f_0, g_0 \in \mathcal{G}$, and $0 \le \int_0^u h(s)\,\mathrm{d}s \le \int_0^\infty h(s)\,\mathrm{d}s$ since $h \ge 0$. We now consider the inductive step. If (4.6) holds for some $n \in \mathbb{N}$ then by the same reasoning as in (4.7), for $0 \le u \le t$,

$$\begin{aligned}|f_{n+1}(u,z) - g_{n+1}(u,z)| &\le \mathbb{E}B \int_0^u h(s)\,\mathrm{d}s \int_0^u |f_n(s,z) - g_n(s,z)|\,\mathrm{d}s \\ &\le C\frac{1}{n!}\int_0^u (Cs)^n\,\mathrm{d}s \\ &= \frac{1}{(n+1)!}(Cu)^{n+1}.\end{aligned}$$

It follows that (4.6) holds for every $n \in \mathbb{N}$. As $n \to \infty$, we can conclude that f_n and g_n have the same limit f, regardless of the initial approximations f_0 and g_0. Moreover, it turns out that the limit f is an element of $\mathcal{G}$, which follows from $f_n \in \mathcal{G}$ for every n in combination with Lévy's convergence theorem; see [31, Chapter XVIII]. From the definition of ϕ, it is easy to see that ϕ is a continuous operator, and, hence,

$$f = \lim_{n\to\infty} f_{n+1} = \lim_{n\to\infty} \phi(f_n) = \phi\Big(\lim_{n\to\infty} f_n\Big) = \phi(f);$$

thus, the limit $f \in \mathcal{G}$ is a fixed point of ϕ, and, hence, (iv) is proved. The above now also immediately yields the uniqueness of the fixed point of ϕ, which completes the proof of (ii). Indeed, suppose that $\eta_0, \tilde{\eta}_0 \in \mathcal{G}$ are both fixed points of ϕ, then

$$\eta_0 = \lim_{n\to\infty} \eta_n = \lim_{n\to\infty} \tilde{\eta}_n = \tilde{\eta}_0.$$

Finally, we need to show that the above not only holds for real z but also complex z such that $|z| < 1$. The conditions of [2, Theorem 5] hold, so that we can conclude that $f_n \to \eta$ for all *complex* z with $|z| < 1$, regardless of the initial approximation $f_0 \in \mathcal{G}$. (Technically the conditions of [2, Theorem 5] concern LSTs, but they can be adapted to a result for z-transforms by applying a variable transform.) □

The iterates of z-transforms can be used to construct lower and upper bounds on the CDF of $S(u)$. We write $[n]$ in the superscript of a mapping to denote an n-fold application of the mapping. Define $F^* \equiv 1$, $F_* \equiv 0$, $F^{n*} = \tilde{\phi}^{[n]}(F^*)$, and $F_{n*} = \tilde{\phi}^{[n]}(F_*)$.

Proposition 4.1. *For every $0 \le u \le t$ and for all $n, k \in \mathbb{N}$,*

$$\begin{aligned}1 &= F^*(u,k) \\ &\ge F^{n*}(u,k) \\ &\ge F^{(n+1)*}(u,k) \\ &\ge F(u,k) \\ &\ge F_{(n+1)*}(u,k) \\ &\ge F_{n*}(u,k) \\ &\ge F_*(u,k) \\ &= 0,\end{aligned}$$

where F_{n*} and F^{n*} are the (defective) CDFs associated with $\phi^{[n]}(z \mapsto 1)$ and $\phi^{[n]}(z \mapsto 0)$, respectively.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Proof. Owing to [2, Theorem 8], it is sufficient to show that $\tilde{\phi}$, as defined in (4.5), regarded as an operator mapping possible defective CDFs into possibly defective CDFs, is monotone in the stochastic ordering. In other words, if we define $F_1 \leq_{\mathrm{st}} F_2$ to mean $F_1(k) \geq F_2(k)$ for all $k \in \mathbb{N}$, then we need to show that $F_1 \leq_{\mathrm{st}} F_2$ implies $\tilde{\phi}(F_1) \leq_{\mathrm{st}} \tilde{\phi}(F_2)$. If $F_1 \leq_{\mathrm{st}} F_2$ then we can construct i.i.d. random variables $\{S_j^{(i)} : i \geq 1\}$ for $j = 1, 2$ such that $S_j^{(1)}$ has CDF F_j for $j = 1, 2$, and $S_1^{(i)} \leq S_2^{(i)}$ for each i. Hence, for every $0 \leq u \leq t$,

$$\mathbf{1}_{\{J>u\}} + \sum_{i=1}^{K(u)} S_1^{(i)}(u - t_i) \leq \mathbf{1}_{\{J>u\}} + \sum_{i=1}^{K(u)} S_2^{(i)}(u - t_i)$$

with probability 1, which implies that, for every $0 \leq u \leq t$,

$$\begin{aligned}\tilde{\phi}(F_1)(u,k) &= \mathbb{P}\left(\mathbf{1}_{\{J>u\}} + \sum_{i=1}^{K(u)} S_1^{(i)}(u - t_i) \leq k\right) \\ &\geq \mathbb{P}\left(\mathbf{1}_{\{J>u\}} + \sum_{i=1}^{K(u)} S_2^{(i)}(u - t_i) \leq k\right) \\ &= \tilde{\phi}(F_2)(u,k) \quad \text{for all } k.\end{aligned}$$

The result follows. □

Remark 4.1. We start in the extremal functions 0 and 1 in $\mathcal{G}$ to ensure that the associated (defective) CDFs provide an upper and lower bound to the CDF of the solution. However, this yields the worst possible bounds and leaves some room for improvement. To save computation time, another practical issue is to find a good initial approximation $f_0 \in \mathcal{G}$. One may, for instance, use the approximation $e^x \approx 1+x$ in (4.4). For example, in the case of exponential jobs, the resulting functional equation is solvable using techniques from [29, Chapter I]. The solution will provide a better initial approximation than simply the function identical to 1 or 0. In this paper, however, we do not pursue these issues any further.

5. Numerics and simulations

The purpose of this section is to show the practical applicability of the tools developed in the preceding sections. At the same time, this exercise verifies the main distributional results of this paper. To be more precise, we apply the Markovian method (as developed in Section 3) and the cluster-based approach (as developed in Section 4). In order to be able to compare the results of both methods, we work with the more restrictive assumptions mentioned at the beginning of Section 3.

Some comments about both approaches are in order. First, in order for us to use Theorem 3.1 to obtain numerical results, we need to numerically invert the transform in (3.1). We do this by using the UBASIC fast Fourier inversion algorithm POISSON; see [1]. This algorithm requires the evaluation of (3.1) for complex values. In each instance, this means solving an ODE with a complex boundary, which can be accomplished by standard ODE solvers. Second, with regards to the functional equation (4.4), in the $J = \infty$ case (i.e. for the usual Hawkes process without departures), Hawkes and Oakes [15] commented that such relations are 'rather intractable'. Also, in [10] it was mentioned that the problem of finding the PMF of the number of customers is generally 'numerically challenging'. We will now show, however, that (4.4) is in fact tractable.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

TABLE 1: Parameter choice, where B is deterministic.

t	λ_∞	r	μ	B
10	1.45	2.15	1.25	0.98

Indeed, we again use the POISSON algorithm for the branching process approach, for which we need to evaluate (4.1) for complex values. Since η is only implicitly known, the idea is now that for each complex argument required by POISSON, we iterate (4.4) a number of times until it is 'close enough' to its fixed point. We have verified that convergence occurs for complex z in Section 4; see Theorem 4.1. Furthermore, we draw upon Proposition 4.1 to derive numerical upper and lower bounds. The upper and lower bounds provide a maximum error that we obtain by using a finite number of iterations of the functional equation in Theorem 4.1.

Proposition 4.1 provides upper and lower bounds on the CDF, while we are actually interested in the PMFs. Clearly, upper and lower bounds on the CDF can be used to derive upper and lower bounds on the PMF. For example, for $k \geq 1$,

$$\left.\begin{aligned} a_0 \leq \mathbb{P}(X \leq k) \leq b_1 \\ a_1 \leq \mathbb{P}(X \leq k+1) \leq b_2 \end{aligned}\right\} \quad \Longrightarrow \quad a_1 - b_1 \leq \mathbb{P}(X = k+1) \leq b_2 - a_0.$$

As a third way to validate the results, we perform simulations. The simulation algorithm is based directly on Definition 2.2, in combination with the thinning procedure of [21] for inhomogeneous Poisson processes. For closely related simulation algorithms, see [24]. After simulating the Hawkes arrival process, we simply flip a coin with success probability $\mathbb{P}(J > t - t_i)$ for arrivals entering at time t_i to determine if they are still in the system at time t.

Example 5.1. In the fast Fourier inversion algorithm, in this example, we use an accuracy of $\gamma = 4$ with γ defined in [1]. Suppose that we have an exponential excitation function with rate r, and exponential job sizes with mean $1/\mu$. Then we are in the Markovian setting and we can use the PDE method from Section 3, as well as the cluster approach from Section 4. In this example, we take the shot sizes B to be deterministic. We choose the parameters as presented in Table 1. The parameters are arbitrarily chosen; however, one should realize that increasing t leads to a longer computation time for both methods. Increasing B or decreasing r leads to a longer computation time only for the method of Section 4, given a fixed accuracy. The results are presented in Table 2. The accuracy of the results is determined by the discretization step in solving the ODE, which is set at 10^{-4}, and by the number of iterations in the functional approach, which is taken to be 10, and the discretization size of the numerical integrals, which we take to be 2^{-12}. These parameters are chosen in such a way that the computation time is reasonable (in the order of a couple of minutes), and such that finer discretization does not improve the results substantially.

We perform 100 batches of 100 000 simulations (totalling 10 million runs), and calculate the standard deviations over these 100 batches (runtime approximately one hour). Note that the estimated values of all methods agree quite well for small values of $N(10)$ in Table 2. For larger values of $N(10)$, the simulation estimates become more uncertain, as well as the spread between the upper and lower bounds of the cluster method. Greater accuracy in the tail can be achieved by performing more than the current number of iterations.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

TABLE 2: The probabilities of there being 0, 1, ..., 13 customers in the system, as determined by several approaches. The standard deviation of the simulations is listed under 'Simulation SD'. The columns headed 'Cluster upper' and 'Cluster lower' refer to the upper and lower bounds obtained using Proposition 4.1. The column 'Cluster' uses the initial approximation $f_0 \equiv 1$. The column 'ODE' is the solution that relies on Theorem 3.1.

Number of customers	Probability					
	Cluster upper	Cluster lower	Cluster	ODE	Simulation	Simulation SD
0	1.83×10^{-1}	1.83×10^{-1}	1.83×10^{-1}	1.83×10^{-1}	1.83×10^{-1}	0.01×10^{-1}
1	2.54×10^{-1}	2.53×10^{-1}	2.54×10^{-1}	2.54×10^{-1}	2.54×10^{-1}	0.01×10^{-1}
2	2.19×10^{-1}	2.17×10^{-1}	2.18×10^{-1}	2.18×10^{-1}	2.18×10^{-1}	0.01×10^{-1}
3	1.51×10^{-1}	1.48×10^{-1}	1.50×10^{-1}	1.50×10^{-1}	1.50×10^{-1}	0.01×10^{-1}
4	9.22×10^{-2}	8.91×10^{-2}	9.08×10^{-2}	9.09×10^{-2}	9.10×10^{-2}	0.09×10^{-2}
5	5.24×10^{-2}	4.89×10^{-2}	5.07×10^{-2}	5.09×10^{-2}	5.09×10^{-2}	0.07×10^{-2}
6	2.87×10^{-2}	2.49×10^{-2}	2.68×10^{-2}	2.70×10^{-2}	2.70×10^{-2}	0.05×10^{-2}
7	1.56×10^{-2}	1.17×10^{-2}	1.36×10^{-2}	1.38×10^{-2}	1.37×10^{-2}	0.04×10^{-2}
8	8.69×10^{-3}	4.74×10^{-3}	6.73×10^{-3}	6.81×10^{-3}	6.80×10^{-3}	0.30×10^{-3}
9	5.23×10^{-3}	1.23×10^{-3}	3.24×10^{-3}	3.29×10^{-3}	3.30×10^{-3}	0.20×10^{-3}
10	3.53×10^{-3}	0.00	1.53×10^{-3}	1.56×10^{-3}	1.60×10^{-3}	0.10×10^{-3}
11	2.73×10^{-3}	0.00	7.12×10^{-4}	7.20×10^{-4}	7.40×10^{-4}	0.80×10^{-4}
12	2.34×10^{-3}	0.00	3.27×10^{-4}	3.34×10^{-4}	3.40×10^{-4}	0.60×10^{-4}
13	2.16×10^{-3}	0.00	1.48×10^{-4}	1.53×10^{-4}	1.60×10^{-4}	0.40×10^{-4}

6. Asymptotic results

In this section we use the findings from the previous section to derive asymptotic results. In Subsection 6.1 we assume that B is regularly varying of index $-\alpha$. We show that $N(t)$ is then also regularly varying and of the same index. In Subsection 6.2 we study the heavy-traffic behavior of $N(\infty)$. In both cases we rely strongly on the representation (4.1).

6.1. Heavy-tailed asymptotics

In this subsection we consider the case that the random variables B_i are heavy tailed. We use the following definition of a regularly varying random variable.

Definition 6.1. A random variable X on $[0, \infty)$ is called regularly varying of index $-\alpha$, denoted by $\mathcal{R}(-\alpha)$, with $\alpha > 0$, if

$$\mathbb{P}(X > x) = \ell(x)x^{-\alpha}, \qquad x \geq 0,$$

with $\ell(x)$ a slowly varying function at ∞, i.e. $\ell(\gamma x)/\ell(x) \to 1$ as $x \to \infty$ for all $\gamma > 1$.

We now prove that if the B_i are $\mathcal{R}(-\alpha)$ with $1 < \alpha < 2$ then so is the number of customers $N(t)$ at time t in the Hawkes/G/∞ queue. Note that this contrasts with the fact that the number of customers in the M/G/∞ queue (starting empty at time 0) is Poisson distributed at any time t.

In what follows, '$*$' denotes convolution, e.g.

$$(f * g)(u) := \int_0^u f(s)g(u-s)\,\mathrm{d}s,$$

and n-fold convolutions are iteratively defined by

$$f^{n*}(u) = (f * f^{(n-1)*})(u) \quad \text{for } n \geq 2$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

with $f^{1*} \equiv f$. In addition we define, with $1 < \alpha < 2$,

$$R_1(u) \coloneqq \sum_{n=0}^{\infty} (b_1)^n (h^{n*} * \mathcal{J})(u), \qquad u > 0, \tag{6.1}$$

$$R_\alpha(u) \coloneqq \Gamma(1-\alpha)\ell(\infty) \sum_{n=0}^{\infty} (b_1)^n (h^{n*} * (h * R_1)^\alpha)(u), \qquad u > 0. \tag{6.2}$$

Throughout we impose the stability condition, which now reads as (see Remark 3.3) $\rho \coloneqq b_1 \int_0^\infty h(s)\,\mathrm{d}s < 1$.

Theorem 6.1. *If B is $\mathcal{R}(-\alpha)$ with $\alpha \in (1,2)$ then so is $N(t)$, i.e. as $z \uparrow 1$,*

$$\mathbb{E}z^{N(t)} - 1 + \lambda_\infty (1-z) \int_0^t R_1(u)\,\mathrm{d}u \sim -\lambda_\infty (1-z)^\alpha \int_0^t R_\alpha(u)\,\mathrm{d}u.$$

Proof. We use Theorem 4.1 to characterize the z-transform of the distribution of $N(t)$. From [4, Theorem 8.1.6], which relates the behavior of a regularly varying function at ∞ and the behavior of its LST near 0, it follows that $\beta(s) - 1 + sb_1 \sim -\Gamma(1-\alpha)s^\alpha \ell(s^{-1})$ as $s \downarrow 0$ for $\alpha \in (1,2)$. Thus, as $z \uparrow 1$,

$$\begin{aligned}
&\beta\left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right) - 1 + \left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right) b_1 \\
&\qquad \sim -\Gamma(1-\alpha)\left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha \ell\left(\frac{1}{\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s}\right).
\end{aligned}$$

Substituting this into (4.3), we obtain, as $z \uparrow 1$,

$$\begin{aligned}
1-\eta(u,z) \sim 1 - (\bar{\mathcal{J}}(u) + z\mathcal{J}(u))\Bigg[1 &- \left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right) b_1 \\
&- \Gamma(1-\alpha)\left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha \\
&\times \ell\left(\frac{1}{\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s}\right)\Bigg].
\end{aligned} \tag{6.3}$$

Rearranging (6.3) and simplifying yields, as $z \uparrow 1$, up to $O((z-1)^2)$ terms,

$$\begin{aligned}
1-\eta(u,z) &= 1 - (1 - \mathcal{J}(u)(1-z))\Bigg[1 - \left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right) b_1 \\
&\qquad - \Gamma(1-\alpha)\left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha \\
&\qquad \times \ell\left(\frac{1}{\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s}\right)\Bigg] \\
&= \mathcal{J}(u)(1-z) + \left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right) b_1 \\
&\quad + \Gamma(1-\alpha)\left(\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s\right)^\alpha \ell\left(\frac{1}{\int_0^u h(s)(1-\eta(u-s,z))\,\mathrm{d}s}\right).
\end{aligned} \tag{6.4}$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Observe that $1-\eta(u,1)=0$, and if $\mathbb{E}S(u)$ is finite then $1-\eta(u,z)=1-\mathbb{E}z^{S(u)}=\mathbb{E}S(u)(1-z)+o(1-z)$. Now we claim that the leading term of $1-\eta(u,z)$ in (6.4) should be $\mathbb{E}S(u)(1-z)$ with $\mathbb{E}S(u)=R_1(u)$ for $z\uparrow 1$, as given in (6.1). To this end, observe that it cannot be $R_1(u)(1-z)^{1+\varepsilon}$ for some $\varepsilon>0$ due to the term $\mathcal{J}(u)(1-z)$ on the right-hand side of (6.4)—realize that if that would be the case then dividing by $(1-z)$ and letting $z\uparrow 1$ yields a contradiction. So the dominant term of $1-\eta(u,z)$ for $z\uparrow 1$ is $R_1(u)(1-z)$, where

$$R_1(u)=\mathcal{J}(u)+b_1\int_0^u h(s)R_1(u-s)\,\mathrm{d}s. \tag{6.5}$$

The above equation can be recognized as a Volterra integral equation of the second kind; see [29, Chapter I]. By using Picard iteration, we obtain for $u\geq 0$,

$$\begin{aligned}R_1(u)&=\mathcal{J}(u)+b_1\int_0^u h(s)\mathcal{J}(u-s)\,\mathrm{d}s\\&\quad+b_1\int_0^u h(s)\left[b_1\int_0^{u-s}h(v)\mathcal{J}(u-s-v)\,\mathrm{d}v\right]\mathrm{d}s+\cdots\\&=\mathcal{J}(u)+b_1(h*\mathcal{J})(u)+(b_1)^2(h^{2*}*\mathcal{J})(u)+\cdots\\&=\sum_{n=0}^{\infty}(b_1)^n(h^{n*}*\mathcal{J})(u),\qquad u>0.\end{aligned}$$

We now claim that when $z\uparrow 1$ (using the same reasoning as for the first term) the next term of $1-\eta(u,z)$ is $R_\alpha(u)(1-z)^\alpha$ with $R_\alpha(u)$ given in (6.2). Substituting this into (6.4) and calculating the coefficient of $(1-z)^\alpha$, we find that, in the limit $z\uparrow 1$,

$$R_\alpha(u)=b_1\int_0^u h(s)R_\alpha(u-s)\,\mathrm{d}s+\Gamma(1-\alpha)\ell(\infty)\left(\int_0^u h(s)R_1(u-s)\,\mathrm{d}s\right)^\alpha. \tag{6.6}$$

This equation is also a Volterra integral equation of the second kind (see [29, Chapter I]) and using Picard iteration, we obtain

$$R_\alpha(u)=\Gamma(1-\alpha)\ell(\infty)\sum_{n=0}^{\infty}(b_1)^n(h^{n*}*(h*R_1)^\alpha)(u),\qquad u>0.$$

We now use Theorem 4.1, in combination with the asymptotics just established, for the first and second terms. As $z\uparrow 1$,

$$\begin{aligned}\mathbb{E}z^{N(t)}&\sim\exp\left(-\lambda_\infty\int_0^t(R_1(u)(1-z)+R_\alpha(u)(1-z)^\alpha)\,\mathrm{d}u\right)\\&=\exp\left(-\lambda_\infty(1-z)\int_0^t R_1(u)\,\mathrm{d}u\right)\exp\left(-\lambda_\infty(1-z)^\alpha\int_0^t R_\alpha(u)\,\mathrm{d}u\right).\end{aligned}$$

Using the exponential function expansion in the above equation, we obtain

$$\begin{aligned}\mathbb{E}z^{N(t)}&\sim\left(1-\lambda_\infty(1-z)\int_0^t R_1(u)\,\mathrm{d}u+O((1-z)^2)\right)\\&\quad\times\left(1-\lambda_\infty(1-z)^\alpha\int_0^t R_\alpha(u)\,\mathrm{d}u+O((1-z)^{2\alpha})\right)\\&=1-\lambda_\infty(1-z)\int_0^t(R_1(u)+(1-z)^{\alpha-1}R_\alpha(u))\,\mathrm{d}u+O((1-z)^2).\end{aligned}$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Using [4, Theorem 8.1.6] in the reverse way, we conclude that, indeed, $N(t) \in \mathcal{R}(-\alpha)$, with $\alpha \in (1, 2)$. □

Remark 6.1. For special choices of $h(\cdot)$ and $\mathcal{J}(u)$, one can obtain explicit expressions for $\int_0^t R_1(u)\,\mathrm{d}u$ and $\int_0^t R_\alpha(u)\,\mathrm{d}u$. Below we consider the case $h(u) = \mathrm{e}^{-ru}$ and $\mathcal{J}(u) = \mathrm{e}^{-\mu u}$. Substituting these into (6.5), we obtain

$$R_1(u) = \mathrm{e}^{-\mu u} + b_1 \int_0^u \mathrm{e}^{-r(u-s)} R_1(s)\,\mathrm{d}s. \tag{6.7}$$

The above equation is a Volterra integral equation of the second kind with kernel $k(u, s) = \mathrm{e}^{-r(u-s)}$. Solving it using the resolvent kernel method given in [29, Chapter I], we obtain

$$R_1(u) = \frac{1}{r_0 - \mu}[(r - \mu)\mathrm{e}^{-\mu u} - b_1 \mathrm{e}^{-r_0 u}].$$

From this, or directly from (6.7) after integrating both sides from 0 to ∞, we obtain

$$\int_0^\infty R_1(u)\,\mathrm{d}u = \frac{1/\mu}{1 - \rho}.$$

We next turn to $R_\alpha(u)$. We substitute $h(\cdot)$ and $R_1(\cdot)$ into (6.6), which leads to a Volterra integral equation of the second kind, i.e.

$$R_\alpha(u) = b_1 \int_0^u \mathrm{e}^{-r(u-s)} R_\alpha(s)\,\mathrm{d}s + \Gamma(1 - \alpha)\ell(\infty)\left(\frac{\mathrm{e}^{-\mu u} - \mathrm{e}^{-r_0 u}}{r_0 - \mu}\right)^\alpha.$$

Substituting $r_0 = r(1 - \rho)$ and then integrating both sides from 0 to ∞ yields

$$\int_0^\infty R_\alpha(u)\,\mathrm{d}u = \frac{b_1}{r}\int_0^\infty R_\alpha(s)\,\mathrm{d}s + \Gamma(1 - \alpha)\ell(\infty)\int_0^\infty \left(\frac{\mathrm{e}^{-r(1-\rho)s} - \mathrm{e}^{-\mu s}}{\mu - r(1 - \rho)}\right)^\alpha \mathrm{d}u,$$

and, hence,

$$\int_0^\infty R_\alpha(u)\,\mathrm{d}u = \frac{\Gamma(1 - \alpha)\ell(\infty)}{1 - \rho}\int_0^\infty \mathrm{e}^{-\alpha r(1-\rho)s}\left(\frac{1 - \mathrm{e}^{r(1-\rho)s - \mu s}}{\mu - r(1 - \rho)}\right)^\alpha \mathrm{d}s.$$

Substituting $\mathrm{e}^{r(1-\rho)s - \mu s} \coloneqq v$ into the above equation yields

$$\begin{aligned}\int_0^\infty R_\alpha(u)\,\mathrm{d}u &= \frac{\Gamma(1 - \alpha)\ell(\infty)}{(1 - \rho)(\mu - r(1 - \rho))^{\alpha+1}}\int_0^1 v^{\alpha r(1-\rho)/(\mu - r(1-\rho)) - 1}(1 - v)^\alpha\,\mathrm{d}v \\ &= \frac{\Gamma(1 - \alpha)\ell(\infty)}{(1 - \rho)(\mu - r(1 - \rho))^{\alpha+1}}\,\mathrm{beta}\left(\frac{\alpha r(1 - \rho)}{\mu - r(1 - \rho)}, \alpha + 1\right),\end{aligned}$$

where $\mathrm{beta}(\cdot, \cdot)$ is the beta function. Using a well-known property of the beta function, i.e. $p\,\mathrm{beta}(p, q) = (p + q)\,\mathrm{beta}(p + 1, q)$, we obtain

$$\begin{aligned}\int_0^\infty R_\alpha(u)\,\mathrm{d}u = \frac{\Gamma(1 - \alpha)\ell(\infty)}{(1 - \rho)(\mu - r(1 - \rho))^{\alpha+1}}&\frac{\mu(\alpha + 1) - r(1 - \rho)}{\alpha r(1 - \rho)} \\ &\times \mathrm{beta}\left(\frac{\alpha r(1 - \rho)}{\mu - r(1 - \rho)} + 1, \alpha + 1\right).\end{aligned}$$

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

6.2. Heavy-traffic asymptotics

In this subsection we discuss the heavy-traffic behavior of $\Lambda \equiv \Lambda(\infty)$ and $N \equiv N(\infty)$, i.e. we consider the stationary Hawkes intensity process and the number of customers in the regime where we let the load generated by the Hawkes process approach its instability boundary. In other words, we study the system's behavior when $\rho = b_1 \int_0^\infty h(u)\,\mathrm{d}u \uparrow 1$. Let $\Gamma(\alpha_1, \alpha_2)$ denote a gamma distribution with shape parameter α_1 and rate parameter α_2. We assume in this subsection that $h(u) = \mathrm{e}^{-ru}$, so that $\rho = b_1/r$.

Theorem 6.2. *Consider a Hawkes process for which the first two moments of B are finite, and where $h(u) = \mathrm{e}^{-ru}$, $u \geq 0$. As $\rho \uparrow 1$,*

$$(1-\rho)\Lambda \xrightarrow{\mathrm{D}} \Gamma\left(\frac{2r\lambda_\infty}{b_2}, \frac{2r}{b_2}\right).$$

Proof. We start with the steady-state version of (3.7). Take $z = 1$, so that we focus on Λ as $\mathbb{E}\mathrm{e}^{-s\Lambda}$ is the LST of Λ. Then we obtain the ODE

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\mathrm{e}^{-s\Lambda} = -\frac{rs\lambda_\infty}{rs + \beta(s) - 1}\mathbb{E}\mathrm{e}^{-s\Lambda}.$$

The solution to this equation is

$$\mathbb{E}\mathrm{e}^{-s\Lambda} = \exp\left[-\lambda_\infty r \int_0^s \frac{u}{ru + \beta(u) - 1}\,\mathrm{d}u\right]. \tag{6.8}$$

Since we assume that the first two moments of B are finite, we can write $\beta(s) = 1 - sb_1 + \frac{1}{2}s^2 b_2 + o(s^2)$ as $s \downarrow 0$. Now consider $\mathbb{E}\mathrm{e}^{-s(1-\rho)\Lambda}$ with $\rho = b_1/r$. Substituting $u = v(1-\rho)$ and the above $\beta(s)$ expansion into (6.8), we obtain for $\rho \uparrow 1$,

$$\mathbb{E}\mathrm{e}^{-s(1-\rho)\Lambda} = \exp\left[-\lambda_\infty \int_0^s \frac{1}{1 + vb_2/2r + o(1-\rho)}\,\mathrm{d}v\right].$$

For $\rho \uparrow 1$, using the above equation with $\kappa = 2r/b_2$, we obtain

$$\lim_{\rho\uparrow 1} \mathbb{E}\mathrm{e}^{-s(1-\rho)\Lambda} = \left(\frac{\kappa}{\kappa + s}\right)^{\kappa\lambda_\infty}.$$

By virtue of Lévy's convergence theorem [31, Chapter XVIII], the result follows. □

We now prove that a very similar result holds for the steady-state number of customers N in the Hawkes/M/∞ queue. We accomplish this by first observing that the first two moments of $(1-\rho)\Lambda$ and $(1-\rho)\mu N$ have the same limit for $\rho \uparrow 1$ (in fact, this holds for the first g moments, if $\mathbb{E}B^g < \infty$, as can easily be verified from the steady-state version of (3.9)). We subsequently apply the following lemma. Denote $\sigma_{X_n}^2 := \operatorname{var} X_n$, $\sigma_{Y_n}^2 := \operatorname{var} Y_n$, and $C_n := \operatorname{cov}(X_n, Y_n)/\sqrt{\operatorname{var} X_n \operatorname{var} Y_n}$.

Lemma 6.1. *Suppose that the following conditions hold:*

(i) $\mathbb{E}Y_n/\mathbb{E}X_n \to \beta$ *and* $\sigma_{Y_n}/\sigma_{X_n} \to \beta$ *as* $n \to \infty$;

(ii) $C_n \to 1$ *as* $n \to \infty$;

(iii) $X_n \xrightarrow{\mathrm{D}} X$;

(iv) *there is a finite* M *such that, for all* n, $0 \leq \mathbb{E}Y_n < M$ *and* $0 \leq \sigma_{Y_n}^2 < M$.

Then $Y_n \xrightarrow{\mathrm{D}} \beta X$.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

Proof. Define $\alpha_n := C_n\sigma_{Y_n}/\sigma_{X_n}$. Then by using (i) and (ii), $\alpha_n \to \beta$. In addition,

$$\begin{aligned}\operatorname{var}(\alpha_n X_n - Y_n) &= \alpha_n^2\sigma_{X_n}^2 - 2\alpha_n C_n\sigma_{X_n}\sigma_{Y_n} + \sigma_{Y_n}^2\\ &= C_n^2\sigma_{Y_n}^2 - 2C_n^2\sigma_{Y_n}^2 + \sigma_{Y_n}^2\\ &= (1-C_n^2)\sigma_{Y_n}^2\\ &\to 0 \end{aligned} \tag{6.9}$$

by virtue of (ii) and (iv). Then note that

$$\mathbb{E}(\alpha_n X_n - Y_n)^2 = \operatorname{var}(\alpha_n X_n - Y_n) + (\alpha_n\mathbb{E}X_n - \mathbb{E}Y_n)^2 \to 0;$$

the first term goes to 0 due to (6.9), and the second term due to (i), (iv), and $\alpha_n \to \beta$ as $n \to \infty$. Hence, $\alpha_n X_n - Y_n \to 0$ in L^2; using Chebyshev's inequality, it immediately follows that $\varepsilon_n := \alpha_n X_n - Y_n \xrightarrow{\mathbb{P}} 0$.

Now consider $Y_n = \alpha_n X_n - \varepsilon_n$. Using

- $\alpha_n \to \beta$ as $n \to \infty$;
- $X_n \xrightarrow{\mathrm{D}} X$ due to (iii);
- $\varepsilon_n \xrightarrow{\mathbb{P}} 0$;
- Slutsky's lemma [30, Lemma 2.8], i.e. $A_n \xrightarrow{\mathrm{D}} A$ and $B_n \xrightarrow{\mathbb{P}} 0$ implies that $A_n + B_n \xrightarrow{\mathrm{D}} A$ (irrespective of A_n and B_n being dependent);

the claim $Y_n \xrightarrow{\mathrm{D}} \beta X$ follows. □

This lemma, in combination with Theorem 6.2, can be used to obtain the following heavy-traffic result for N, the number of customers in the Hawkes/M/∞ queue. For this purpose, we take $X_n = (1-\rho_n)\Lambda_n$ and $Y_n = (1-\rho_n)N_n$, where ρ_n, $n = 1, 2, \ldots$, is a sequence of parameters converging to 1, and where Λ_n and N_n correspond to the quantities Λ and N in our Hawkes/M/∞ queue when $\rho = \rho_n$.

Theorem 6.3. *Consider the Hawkes/M/∞ queue with $\mathcal{J}(u) = \mathrm{e}^{-\mu u}$. Assume that the first two moments of B are finite, and that $h(u) = \mathrm{e}^{-ru}$, $u \geq 0$. As $\rho \uparrow 1$,*

$$(1-\rho)N \xrightarrow{\mathrm{D}} \Gamma\left(\frac{2r\lambda_\infty}{b_2}, \frac{2r\mu}{b_2}\right).$$

7. Discussion and concluding remarks

In this paper we have analyzed an infinite-server queue fed by a Hawkes arrival process. Under Markovian assumptions, a fairly explicit analysis was possible, leading to, e.g. explicit expressions for the (transient and stationary) moments of the number of customers in the system. Lifting the Markovian assumptions, the analysis became less explicit: results were derived in terms of a fixed-point equation describing the z-transform of the number of customers. We then used this fixed-point equation to derive asymptotic results.

Several branches of further research offer themselves.

In the first place, one could consider *single-server* queues with Hawkes input. In our analysis in the infinite-server setting, we repeatedly used the fact that the customers are served independently of each other, a property that we do not have in a single-server context. This may entail that exact analysis is prohibitively difficult, but analysis in a heavy-traffic setting might be possible.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

In [19] the authors considered networks of infinite-server queues with shot-noise-driven input. There it turned out that the network setting could be analyzed by essentially the same techniques as the single-queue setting. This raises the question as to whether the results of the present paper also naturally extend to that of a *network* of infinite-server queues with Hawkes input. We would have to appeal to the multivariate counterpart of self-exciting processes, which are called *mutually-exciting* arrival processes; see [14]. As the name suggests, in this case arrivals to a particular queue are able to excite arrivals in other queues.

There is a vast body of literature that deals with statistical inference for Hawkes processes, which implies that Hawkes processes can be fitted to real data. Most extensive practical applications in the literature are related to limit order book data; see, e.g. [3], [5], and [28]. In the case of Markovian Hawkes processes, the likelihood function has an explicit expression (see [25]), and the resulting maximum likelihood estimators are known to be consistent, asymptotically normal, and efficient [23]. There are also several more recent papers in which nonparametrics estimation of non-Markovian Hawkes processes are considered; see, e.g. [18]. For a concise survey on statistical literature on Hawkes processes, see [3, Appendix C]. Blanchet *et al.* [5] empirically support the use of infinite-server queues in limit order book models. An interesting line of further research would be to verify if the predictions derived from the model in this paper also correspond to empirical order book data.

One could pursue improving the bounds of Proposition 4.1 in order to make them more useful for larger values of n. Also, better initial approximations can be determined (see Remark 4.1).

In the heavy-traffic setting, there are several interesting directions still to be explored. For example, one could attempt to generalize Theorem 6.3 to the case of generally distributed J; it would also be interesting to study the heavy-traffic behavior of Λ and N in the heavy-tailed setting of Section 6.1.

Acknowledgements

Following the submission of the ArXiv version of this paper (available at https://arxiv.org/abs/1707.02196), J. Pender (Cornell University) kindly informed us that A. Daw and he were also completing a paper [9] on infinite-server queues with Hawkes arrival processes. We thank P. Spreij (University of Amsterdam) for useful discussions.

The research for this paper was partly funded by the NWO Gravitation Project NETWORKS (grant number 024.002.003–Boxma, Koops, and Mandjes) and an NWO Top Grant (grant number 613.001.352–Boxma, Mandjes, and Saxena). The research of O. Boxma was also partly funded by the Belgian Government via the IAP Bestcom Project.

References

[1] Abate, J. and Whitt, W. (1992). Numerical inversion of probability generating functions. *Operat. Res. Lett.* **12,** 245–251.

[2] Abate, J. and Whitt, W. (1992). Solving probability transform functional equations for numerical inversion. *Operat. Res. Lett.* **12,** 275–281.

[3] Bacry, E., Mastromatteo, I. and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure Liquidity* **1,** 1550005.

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

[5] Blanchet, J., Chen, X. and Pei, Y. (2017). Unraveling limit order books using just bid/ask prices. Preprint. Available at http://www.columbia.edu/∼jb2814/papers/LOB_v1.pdf.

[6] Bogachev, V. I. (2007). *Measure Theory*, I. Springer, Berlin.

[7] Cont, R. and De Larrard, A. (2012). Order book dynamics in liquid markets: limit theorems and diffusion approximations. Preprint. Available at https://arxiv.org/abs/1202.6412.

[8] Dassios, A. and Zhao, H. (2011). A dynamic contagion process. *Adv. Appl. Prob.* **43,** 814–846.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press

[9] Daw, A. and Pender, J. (2018). Queues driven by Hawkes processes. *Stoch. Systems* **8,** 192–229.
[10] Gao, X. and Zhu, L. (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems* **90,** 161–206.
[11] Ghufran, S. M. (2010). The computation of matrix functions in particular, the matrix exponential. Masters Thesis, University of Birmingham.
[12] Harris, T. E. (1963). *The Theory of Branching Processes*. Springer, Berlin.
[13] Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *J. R. Statist. Soc. B* **33,** 438–443.
[14] Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58,** 83–90.
[15] Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Prob.* **11,** 493–503.
[16] Heemskerk, M., van Leeuwaarden, J. and Mandjes, M. (2017). Scaling limits for infinite-server systems in a random environment. *Stoch. Systems* **7,** 1–31.
[17] Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centres using Poisson mixtures. *Appl. Stoch. Models Business Industry* **17,** 307–318.
[18] Kirchner, M. (2017). An estimation procedure for the Hawkes process. *Quant. Finance* **17,** 571–595.
[19] Koops, D. T., Boxma, O. J. and Mandjes, M. R. H. (2017). Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems* **86,** 301–325.
[20] Laub, P. J., Taimre, T. and Pollett, P. K. (2015). Hawkes processes. Preprint. Available at https://arxiv.org/abs/1507.02822.
[21] Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of nonhomogenous Poisson processes by thinning. *Naval Res. Logistics Quart.* **26,** 403–413.
[22] Mathijsen, B. W. J., Janssen, A. J. E. M., van Leeuwaarden, J. S. H. and Zwart, B. (2018). Robust heavy-traffic approximations for service systems facing overdispersed demand. *Queueing Systems* **90,** 257–289.
[23] Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.* **30,** 243–261.
[24] Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Trans. Inf. Theory* **27,** 23–31.
[25] Ogata, Y. and Akaike, H. (1982). On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes. *J. R. Statist. Soc. B* **44,** 102–107.
[26] Rizoiu, M.-A. and Xie, L. (2017). Online popularity under promotion: viral potential, forecasting, and the economics of time. Preprint. Available at https://arxiv.org/abs/1703.01012.
[27] Rizoiu, M.-A. *et al.* (2017). Expecting to be HIP: Hawkes intensity processes for social media popularity. Preprint. Available at https://arxiv.org/abs/1602.06033.
[28] Toke, I. M. and Pomponio, F. (2012). Modelling trades-through in a limit order book using Hawkes processes. *Economics Open-Access Open-Assessment E-Journal* **6,** 2012–22.
[29] Tricomi, F. G. (1957). *Integral Equations*. Interscience, New York.
[30] Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
[31] Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

https://doi.org/10.1017/jpr.2018.58 Published online by Cambridge University Press