Adv. Appl. Probab. **54,** 340–403 (2022)
*doi:*10.1017/apr.2021.35

AN EPHEMERALLY SELF-EXCITING POINT PROCESS

ANDREW DAW,[*] *University of Southern California*
JAMOL PENDER,[**] *Cornell University*

Abstract

Across a wide variety of applications, the self-exciting Hawkes process has been used to model phenomena in which the history of events influences future occurrences. However, there may be many situations in which the past events only influence the future as long as they remain active. For example, a person spreads a contagious disease only as long as they are contagious. In this paper, we define a novel generalization of the Hawkes process that we call the *ephemerally self-exciting process*. In this new stochastic process, the excitement from one arrival lasts for a randomly drawn activity duration, hence the ephemerality. Our study includes exploration of the process itself as well as connections to well-known stochastic models such as branching processes, random walks, epidemics, preferential attachment, and Bayesian mixture models. Furthermore, we prove a batch scaling construction of general, marked Hawkes processes from a general ephemerally self-exciting model, and this novel limit theorem both provides insight into the Hawkes process and motivates the model contained herein as an attractive self-exciting process in its own right.

Keywords: Self-excitement; Hawkes process; point process; contagion; interaction models

2020 Mathematics Subject Classification: Primary 60G55
Secondary 60K25; 60J75

1. Introduction

What's past is prologue—unavoidably, the present is shaped by what has already occurred. The current state of the world is indebted to our history. Our actions, behaviors, and decisions are both precursory and prescriptive to those that follow, and this can be observed across a variety of different scenarios. For example, the spread of an infectious disease is accelerated as more people become sick and dampened as they recover. In finance, a flurry of recent transactions can prompt new buyers or sellers to enter a market. On social media platforms, as more and more users interact with a post it can become trending or viral and thus be broadcast to an even larger audience.

Self-exciting processes are an intriguing family of stochastic models in which the history of events influences the future. The paper [36] introduced the concept of self-excitement—defining what is now known as the Hawkes process, a model in which 'the

Received 21 August 2019; revision received 29 June 2021.

[*] Postal address: Marshall School of Business, University of Southern California, 401B Bridge Hall, Los Angeles, CA 90089. Email address: andrew.daw@usc.edu
[**] Postal address: School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, Ithaca, NY 14853.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

current intensity of events is determined by events in the past'. That is, the Hawkes process is a stochastic-intensity point process that depends on the history of the point process itself. The rate of new event occurrences increases as each event occurs. As time passes between occurrences, the intensity is governed by a deterministic excitement kernel. Most often, this kernel is specified so that the intensity jumps upward at event epochs and strictly decreases in the interim. In this way, occurrences beget occurrences; hence the term 'self-exciting'. Unlike in the Poisson process, disjoint increments are not independent in sample paths of the Hawkes process. Instead, they are positively correlated and, by definition, the events of the former influence the events of the latter. Furthermore, the Hawkes process is known to be over-dispersed—meaning that its variance is larger than its mean—which is commonly found in real-world data, whereas the Poisson process has equal mean and variance.

Because of the practical relevance of these model features, self-exciting processes have been used in a wide variety of applications, many of which are quite recent additions to the literature. Seismology was among the first domains to incorporate these models, for example in [51], as the occurrence of an earthquake increases the risk of subsequent seismic activity in the form of aftershocks. Finance has since followed as a popular application and is now perhaps the most active area of work. In these studies, self-excitement is used to capture the often contagious nature of financial activity; see e.g. [2, 4, 5, 7, 15, 24, 28, 53, 62]. Similarly, there have been many recent internet and social media scenarios that have been modeled using self-exciting processes, drawing upon the virality of modern web traffic. For example, see [26, 54, 55]. Notably, this also includes use of Hawkes processes for constructing data-driven methods in the artificial intelligence and machine learning literatures, e.g. [22, 50, 63]. In an intriguing area of work, self-excitement has also been used to model interpersonal communication, for example in application to conversation audio recordings in [49] or in studying email correspondence in [34, 47]. Hawkes processes have also recently been used to represent arrivals to service systems in queueing models, e.g. in [20, 29, 30, 41]. This is of course not an exhaustive list of works in these areas, nor is it a complete account of all the modern applications of self-excitement. Examples of other notable uses include neuroscience [42, 57], environmental management [32], public health [68], movie trailer generation [64], energy conservation [44], and industrial preventative maintenance [65].

As the variety of uses for self-excitement has continued to grow, the number of Hawkes process generalizations has kept pace. By modifying the definition of the Hawkes process in some way, the works in this generalized self-exciting process literature provide new perspectives on these concepts while also empowering and enriching applications. For example, [13] introduces a nonlinear Hawkes process that adapts the definition of the process intensity to feature a general, nonnegative function of the integration over the process history, as opposed to the linear form given originally. Similarly, the quadratic Hawkes process model given by [11] allows for excitation kernels that have quadratic dependence on the process history, rather than simply linear. This is also an example of a generalization motivated by application, as the authors seek to capture time-reversal asymmetry observed in financial data. As another finance-motivated generalization, [17] proposes the dynamic contagion process. This model can be thought of as a hybrid between a Hawkes process and a shot-noise process, as the stochastic intensity of the model features both self-excited and externally excited jumps. The authors take motivation from an application in credit risk, in which the dynamics are shaped both by the process history and by exogenous shocks. The affine point processes studied in e.g. [24, 66, 67] are also motivated by credit risk applications. The models in these works combine the self-exciting dynamics of Hawkes process with those of an affine jump-diffusion process,

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

imbedding modeling concepts of feedback and dependency into the process intensity. An exact simulation procedure for the Hawkes process with Cox–Ingersoll–Ross (CIR) intensity, a generalization of the Hawkes process that is a special case of the affine point process, is shown in [18]. In that case, the authors discuss an application to portfolio loss processes.

There have also been several Hawkes process generalizations proposed in social media and data analytics contexts. For example, [54] introduces a finite-population Hawkes process that couples self-excitement dynamics with those of the susceptible–infected–recovered (SIR) process. Drawing upon the use of the SIR process for the spread of both disease and ideas, the authors propose this SIR–Hawkes process as a method of studying information cascades. Similarly, [50] introduces the neural Hawkes process as a new point process model in the machine learning literature. As the name suggests, this model combines self-excitement with concepts from neural networks. Specifically, a recurrent neural network effectively replaces the excitation kernel, governing the effect of the past events on the rate of future occurrences. In the literature for Bayesian nonparametric models, [22] presents the Dirichlet–Hawkes process for topic clustering in document streams. In this case, the authors combine a Hawkes process and a Dirichlet process, so that the intensity of the stream of new documents is self-exciting while the type of each new document is determined by the Dirichlet process, leading to a preferential attachment structure among the document types.

In this paper, we propose the *ephemerally self-exciting process* (ESEP), a novel generalization of the Hawkes process. Rather than regulating the excitement through the gradual, deterministic decay provided by an excitation kernel function, we instead incorporate randomly timed down-jumps. We will refer to this random length of time as the arrival's activity duration. The down-jumps are equal in size to the up-jumps, and between events the arrival rate does not change. Thus, this process increases in arrival rate upon each occurrence, and these increases are then mirrored some time later by decreases in the arrival rate once the activity duration expires. In this way, the self-excitement is ephemeral: it is only in effect as long as the excitement is active. Much of the body of this work will discuss how this ephemeral, piecewise constant self-excitement compares to the eternal but ever-decaying notion from Hawkes's original definition. As we will see in our analysis, this new process is both a promising model of self-excitement and an explanation of its origins in natural phenomena.

1.1. Practical relevance

While this paper will not be focused on any one application, in this subsection we summarize several domain areas in which the models in this work can be applied. A natural example is in public health and the management of epidemics. For example, consider influenza. When a person becomes sick with the flu, she increases the rate of spread of the virus through her contact with others. This creates a self-exciting dynamic in the spread of the virus. However, a person only spreads a disease as long as she is contagious; once she has recovered she no longer has a direct effect on the rate of new infections. From a system-level perspective, the ESEP can thus be thought of as modeling the arrivals of new infections, capturing the self-exciting and ephemeral nature of sick patients. This motivates the use of this model as an arrival process to queueing models for healthcare, as the rate of *arrivals to* clinics serving patients with infectious diseases should depend on the number of people currently infected. The healthcare service can also be separately modeled, as an infinite-server queue may be a fitting representation for the number of infected individuals but the clinic itself likely has limited capacity. This concept of course extends to the modeling and management of any other viral disease, including the novel coronavirus that has caused the COVID-19 pandemic.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Of course, epidemic models need not be applied exclusively to disease spread. These same ideas can be used for information spread and product adoption, such as in the aforementioned Hawkes-infused models in [54, 68]. In these contexts, one can think of the duration in system as being the time a person actively promotes a concept or product. A single person only affects the self-excitement of the idea or product spread as long as she is in the system, which distinguishes this model from those in the aforementioned works. Epidemic models have also been used to study social issues, such as the contagious nature of imprisonment demonstrated by [46]. We discuss the relevance of ephemeral self-excitement for epidemics in detail in Subsection 3.3 by relating this model to the susceptible–infected–susceptible (SIS) process through a convergence in distribution. In fact, throughout Section 3 we establish connections from this process to other relevant stochastic models. This includes classical processes such as branching processes and random walks, as well as models popular both in Bayesian nonparametrics and in preferential attachment settings, such as the Dirichlet process and the Chinese restaurant process (CRP).

In the context of service systems, self-excitement can be motivated by the same rationale that inspires restaurants to seat customers at the tables by the windows. Potential new customers could choose to dine at the establishment because they can see others already eating there, taking an implicit recommendation from those already being served. This same example also motivates the ephemerality. After a customer seated by the window finishes her dinner and departs, any passing potential patron only sees an empty table; the implicit recommendation vanishes with the departing customer. A similar dynamic can be observed in online streaming platforms. For example, on popular music streaming services like Spotify and Apple Music, users can see what songs and albums have been recently played by their friends. If a user sees that many of her friends have listened to the same album recently, she may be more inclined to listen to it as well. However, this applies only as long as the word ‘recently’ does. If her friends don’t play the album within a certain amount of time, the platform will no longer promote the album to her in that fashion. Again, this displays the ephemerality of the underlying self-excitement: the album grows more attractive as more users listen to it, but only as long as those listens are ‘recent’ enough.

In finance, limit order books (LOBs) are among the many concepts that have been modeled using Hawkes process, for example in [6, 53]. LOBs have also been studied through queueing models, where one can model the state of the LOB (or, more specifically, the number of unresolved bids and asks) as the length of a queueing process. Moreover, there has been recent work that models this process as not just a queue, but a queue with Hawkes process arrivals; for example see [29, 31]. Conceptually, the self-excitement may arise from traders reacting to the activity of other traders, creating runs of transactions. However, the desire not to act on stale information may mean that this excitement lasts only as long as trades are actively being conducted. In fact, the idea of the self-excitement in LOB models being ‘queue-reactive’ has just very recently been considered by [62], a work related to this one.

One can also consider failures in a mechanical system as an application of this model. For example, consider a network of water pipes. When one pipe breaks or bursts, it can place stress on the pipes connected to it. This stress may then cause further failures within the pipe network. However, once the pipe is properly repaired, it should no longer place strain on the surrounding components. Thus, the increase in pipe failure rate caused by a failure is only in effect until the repair occurs, inducing ephemeral self-excitement. The self-excitement (albeit without the ephemerality) arising in this scenario was modeled using Hawkes processes in [65], which includes an empirical study. A similar problem for electrical systems is

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

considered in [25]. The reactive point process considered in that work is perhaps the model most similar to the ones studied herein, as the rate of new power failures both increases at the prior failure times and decreases upon inspection or repair. However, a key difference is that in [25], the authors treat the inspection times as controlled by management, whereas in this paper the model is fully stochastic and thus the repair durations are random. Regardless, that work is an excellent example of how generalized self-exciting processes can be used to shape practical policy. Because power outages have significant and wide-reaching consequences, it is critical to understand the interdependency between electrical grid failures and to study the ESEP that arises from them.

1.2. Organization and contributions of paper

Let us now detail the remainder of this paper's organization, as well as the contributions therein.

- In Section 2, we define the ephemerally self-exciting process (ESEP), a Markovian model for self-excitement that lasts for only a finite amount of time. After defining the model, we develop fundamental distributional quantities and compare the ESEP to the Hawkes process.
- In Section 3, we relate the ESEP to many other important and well-known stochastic processes. This includes branching processes, which gives us further comparisons between the Hawkes process and the ESEP, models for preferential attachment and Bayesian statistics, and epidemic models. The lattermost of these motivates the ESEP as a representation for the times of infection within an epidemic, and this also provides a formal link between the conceptually similar concepts of epidemics and self-excitement.
- In Section 4, we broaden our exploration of ephemeral self-excitement to non-Markovian models with general activity durations and batches of arrivals. In this general setting, we establish a limit theorem providing an alternate construction of general Hawkes processes. This batch scaling limit thus yields intuition for the observed occurrence of self-excitement in natural phenomena and stands as a fundamental cornerstone for studying such processes.

In addition to these main avenues of study, we also have extensive auxiliary analysis housed in this paper's appendix. Appendix A contains lemmas and side results that support our analysis but are outside the main narrative. In Appendix B, we explore a model that is a hybrid between the ESEP and the Hawkes process, in that it regulates the excitement with both down-jumps and decay. Appendix C is devoted to a finite-capacity version of the ESEP, in which arrivals that would put the active number in system above the capacity are blocked from occurring. Finally, Appendix D contains an algebraically cumbersome proof of a result from Section 2.

2. Modeling ephemeral self-excitement

We begin this paper by defining our ephemerally self-exciting model and conducting an initial analysis of some fundamental quantities, including the transient moment generating function and the steady-state distribution. Before doing so, though, let us first review the Hawkes process, which is the original self-exciting probability model.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

2.1. Preliminary models and concepts

Introduced and pioneered through the series of papers [35–37], the Hawkes process is a stochastic-intensity point process in which the current rate of arrivals is dependent on the history of the arrival process itself. Formally, this is defined as follows: let $(\lambda_t, N_{t,\lambda})$ be an intensity and counting process pair such that

$$\begin{aligned}&\mathbb{P}\left(N_{t+\Delta,\lambda} - N_{t,\lambda} = 1 \mid \mathcal{F}_t^N\right) = \lambda_t \Delta + o(\Delta),\\&\mathbb{P}\left(N_{t+\Delta,\lambda} - N_{t,\lambda} > 1 \mid \mathcal{F}_t^N\right) = o(\Delta),\\&\mathbb{P}\left(N_{t+\Delta,\lambda} - N_{t,\lambda} = 0 \mid \mathcal{F}_t^N\right) = 1 - \lambda_t \Delta + o(\Delta),\end{aligned}$$

where $\mathcal{F}_t^N$ is the filtration of $N_{t,\lambda}$ up to time t and λ_t is given by

$$\lambda_t = \lambda^* + \int_{-\infty}^{t} g(t-u)\mathrm{d}N_{u,\lambda},$$

where $\lambda^* > 0$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$ is such that $\int_0^\infty g(x)\mathrm{d}x < 1$. Through this definition, the intensity λ_t captures the history of the arrival process up to time t. Thus, λ_t encapsulates the sequence of past events and uses it to determine the rate of future occurrences. We refer to λ^* as the baseline intensity and $g(\cdot)$ as the excitation kernel. The baseline intensity represents an underlying stationary arrival rate, and the excitation kernel governs the effect that the history of the process has on the current intensity. A common modeling choice is to set $g(x) = \alpha e^{-\beta x}$, where $\beta > \alpha > 0$. This is often referred to as the 'exponential' kernel, and it is perhaps the most widely used form of the Hawkes process. In this case, $(\lambda_t, N_{t,\lambda})$ is a Markov process obeying the stochastic differential equation

$$\mathrm{d}\lambda_t = \beta(\lambda^* - \lambda_t)\mathrm{d}t + \alpha \mathrm{d}N_{t,\lambda}.$$

That is, at arrival epochs λ_t jumps upward by the amount α and the $N_{t,\lambda}$ increases by 1; between arrivals λ_t decays exponentially at rate β towards the baseline intensity λ^*. Thus, each arrival increases the likelihood of additional arrivals occurring soon afterwards—hence, it self-excites. This form of the Hawkes process is also often alternatively stated with an initial value for λ_t, say $\lambda_0 \geq \lambda^*$. In this case, the intensity can be expressed as

$$\lambda_t = \lambda^* + (\lambda_0 - \lambda^*)e^{-\beta t} + \alpha \int_0^t e^{-\beta(t-u)}\mathrm{d}N_{u,\lambda}.$$

Additional overview of the Hawkes process with the exponential kernel can be found in Section 2 of [20]. Another common choice for excitation kernel is the 'power-law' kernel $g(x) = \frac{k}{(c+x)^p}$, where $k > 0$, $c > 0$, and $p > 0$. This kernel was originally popularized in seismology [51].

2.2. Defining the ephemerally self-exciting process

As we have discussed in the introduction, a plethora of natural phenomena exhibit self-exciting features but only for a finite amount of time. This prompts the notion of ephemeral self-excitement. By comparison to the traditional Hawkes process we have reviewed in Subsection 2.1, we seek a model in which a new occurrence increases the arrival rate only so long as the newly entered entity remains active in the system. Thus, we now define the *ephemerally self-exciting process* (ESEP), which trades the Hawkes process's eternal decay for randomly drawn expiration times. Moreover, in the following Markovian model, exponential decay is replaced with exponentially distributed durations. In Section 4, we extend these

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

concepts to generally distributed service. As another generalization, in Appendix B we consider a Markovian model with both decay and down-jumps. For now, we explore the effects of ephemerality through the ESEP model in Definition 1.

Definition 1. *(Ephemerally self-exciting process.)* For times $t \geq 0$, a baseline intensity $\eta^* > 0$, intensity jump size $\alpha > 0$, and expiration rate $\beta > 0$, let N_t be a counting process with stochastic intensity η_t such that

$$\eta_t = \eta^* + \alpha Q_t, \tag{1}$$

where Q_t is incremented with N_t and then is depleted at unit down-jumps according to the rate βQ_t. We say that (η_t, N_t) *is an ephemerally self-exciting process* (ESEP).

We will assume that η_0 and Q_0 are known initial values such that $\eta_0 = \eta^* + \alpha Q_0$. In addition to this definition, one could also describe the ESEP through its dynamics. In particular, the behavior of this process can be summarily cast through the life cycle of its arrivals:

i. At each arrival, the arrival rate η_t increases by α.

ii. Each arrival remains active for an activity duration drawn from an independent and identically distributed (i.i.d.) sequence of exponential random variables with rate β.

iii. At the expiration of a given activity duration, η_t decreases by α.

The ephemerality of the ESEP is embodied by this cycle. Because arrivals only contribute to the intensity for the length of their activity duration, their effect on the process's excitation vanishes when this clock expires. Furthermore, there is an affine relationship between the number of active 'exciters'—meaning unexpired arrivals still causing excitation—and the intensity, i.e. $\eta_t = \eta^* + \alpha Q_t$. Thus, we could also track the arrival rate through Q_t in place of η_t and still have full understanding of this process. This also means that results are readily transferrable between these two processes; we will often make use of this fact.

Because the ESEP is quite parsimonious, there are many alternative perspectives we could take to gain additional understanding of it. For example, one could consider Q_t a Markovian queueing system with infinitely many servers and a state-dependent arrival rate. Equivalently, one could also describe the ESEP as a Markov chain on the nonnegative integers where transitions at state i are to $i + 1$ at rate $\eta^* + \alpha i$ and to $i - 1$ at rate μi, with the counting process then defined as the epochs of the upward jumps in this chain. This Markov chain perspective certifies the existence and uniqueness of Definition 1. A visualization of this linear birth–death–immigration process is given in Figure 1. Stability for this chain occurs when $\beta > \alpha$; we will assume this henceforth, although of course it is not necessary for transient results. One could also view the ESEP as a generalization of Hawkes's original definition where the excitation kernel function $g(\cdot)$ is replaced with a randomly drawn indicator function that is different for each arrival. Each indicator function compares time to an independently drawn exponential random variable, and this perspective is closely aligned with our analysis in Section 3.

In the remainder of this subsection, let us now develop a few fundamental quantities for this stochastic process, particularly its intensity and active number in system, as these capture the self-exciting behavior of the process. First, in Proposition 1 we compute the transient moment generating function for the intensity η_t. As we have noted, this can also be used to immediately derive the same transform for Q_t, and the proof makes use of this fact.

Proposition 1. *Let $\eta_t = \eta^* + \alpha Q_t$ be the intensity of an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then the moment generating function for η_t is given by*

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

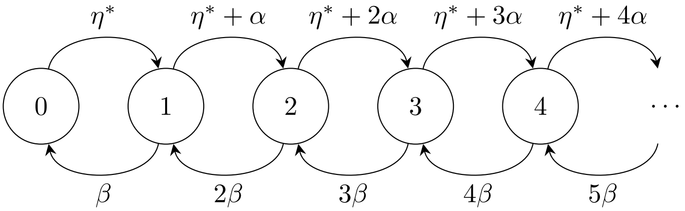


FIGURE 1. The transition diagram of the Markov chain for Q_t.

$$\mathbb{E}\left[e^{\theta\eta_t}\right] = \left(\frac{\beta - \alpha e^{\alpha\theta} - \beta\left(1 - e^{\alpha\theta}\right)e^{-(\beta-\alpha)t}}{\beta - \alpha e^{\alpha\theta} - \alpha\left(1 - e^{\alpha\theta}\right)e^{-(\beta-\alpha)t}}\right)^{\frac{\eta_0 - \eta^*}{\alpha}}$$
$$\cdot\left(\frac{\beta e^{\alpha\theta}}{\beta - \alpha e^{\alpha\theta}} - \frac{\alpha e^{\alpha\theta}}{\beta - \alpha e^{\alpha\theta}}\left(\frac{\beta - \alpha e^{\alpha\theta} - \beta\left(1 - e^{\alpha\theta}\right)e^{-(\beta-\alpha)t}}{\beta - \alpha e^{\alpha\theta} - \alpha\left(1 - e^{\alpha\theta}\right)e^{-(\beta-\alpha)t}}\right)\right)^{\frac{\eta^*}{\alpha}},$$

for all $t \geq 0$ *and* $\theta < \frac{1}{\alpha}\log\left(\frac{\beta}{\alpha}\right)$.

Proof. We will approach this through the perspective of the active number in system, Q_t. Using Lemma 2, we have that the probability generating function for Q_t, say $\mathcal{P}(z, t) = \mathbb{E}\left[z^{Q_t}\right]$ for $z \in [0, 1]$, is given by the solution to the following partial differential equation (PDE):

$$\frac{\partial}{\partial t}\mathbb{E}\left[z^{Q_t}\right] = \mathbb{E}\left[\left(\eta^* + \alpha Q_t\right)\left(z^2 - z\right)z^{Q_t - 1} + \beta Q_t\left(1 - z\right)z^{Q_t - 1}\right],$$

which is equivalently expressed by

$$\frac{\partial}{\partial t}\mathcal{P}(z, t) = \eta^*\left(z - 1\right)\mathcal{P}(z, t) + \left(\alpha\left(z^2 - z\right) + \beta(1 - z)\right)\frac{\partial}{\partial z}\mathcal{P}(z, t),$$

with initial condition $\mathcal{P}(z, 0) = z^{Q_0}$. The solution to this initial value problem is given by

$$\mathcal{P}(z, t) = \left(\frac{\beta - \alpha z - \beta(1 - z)e^{-(\beta-\alpha)t}}{\beta - \alpha z - \alpha(1 - z)e^{-(\beta-\alpha)t}}\right)^{Q_0}$$
$$\cdot\left(\frac{\beta}{\beta - \alpha z} - \frac{\alpha}{\beta - \alpha z}\left(\frac{\beta - \alpha z - \beta(1 - z)e^{-(\beta-\alpha)t}}{\beta - \alpha z - \alpha(1 - z)e^{-(\beta-\alpha)t}}\right)\right)^{\frac{\eta^*}{\alpha}},$$

yielding the probability generating function for Q_t. By setting $z = e^{\theta}$ we obtain the moment generating function. Finally, using the affine relationship $\eta_t = \eta^* + \alpha Q_t$, we have that

$$\mathbb{E}\left[e^{\theta\eta_t}\right] = \mathbb{E}\left[e^{\theta\left(\eta^* + \alpha Q_t\right)}\right] = e^{\theta\eta^*}\mathbb{E}\left[e^{\alpha\theta Q_t}\right],$$

with $\eta_0 = \eta^* + \alpha Q_0$. □

As we have mentioned, this Markov chain can be shown to be stable for $\beta > \alpha$ through standard techniques. Thus, using the moment generating function from Proposition 1, we can find the steady-state distributions of the intensity and the active number in system by taking the limit of t. We can quickly observe that this leads to a negative binomial distribution, as we state in Theorem 1. Because of the varying definitions of the negative binomial distribution, we state the probability mass function explicitly.

Theorem 1. *Let $\eta_t = \eta^* + \alpha Q_t$ be an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then the active number in system in steady state follows a negative binomial distribution with parameters $\frac{\alpha}{\beta}$ and $\frac{\eta^*}{\alpha}$, which is to say that the steady-state probability mass function is*

$$\mathbb{P}\left(Q_\infty = k\right) = \frac{\Gamma\left(k + \frac{\eta^*}{\alpha}\right)}{\Gamma\left(\frac{\eta^*}{\alpha}\right) k!} \left(\frac{\beta - \alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}} \left(\frac{\alpha}{\beta}\right)^k . \tag{2}$$

Consequently, the steady-state distribution of the intensity is given by a shifted and scaled negative binomial with parameters $\frac{\alpha}{\beta}$ and $\frac{\eta^*}{\alpha}$, shifted by η^* and scaled by α.

Proof. Using Proposition 1, we can see that the steady-state moment generating function of Q_t is given by

$$\lim_{t\to\infty} \mathbb{E}\left[e^{\theta Q_t}\right] = \left(\frac{\beta - \alpha}{\beta - \alpha e^{\theta}}\right)^{\frac{\eta^*}{\alpha}} .$$

We can observe that this steady-state moment generating function is equivalent to that of a negative binomial. By the affine transformation $\eta_t = \eta^* + \alpha Q_t$, we find the steady-state distribution for the intensity. □

The negative binomial distribution presented in Theorem 1 allows for $\frac{\eta^*}{\alpha}$ to take on any positive real value; integrality is not required. If $\frac{\eta^*}{\alpha}$ is in fact an integer, then the gamma functions will match the corresponding factorial functions and this ratio of factorials will simplify to a binomial coefficient, reproducing the most familiar form of the negative binomial distribution. In this case, $\frac{\eta^*}{\alpha}$ can be interpreted as the number of failures upon which an experiment is stopped, where the success probability is $\frac{\alpha}{\beta}$ and $k + \frac{\eta^*}{\alpha}$ is the total number of trials.

Let us pause to note that this explicit characterization of the steady-state intensity is already an advantage of the ESEP over the traditional Markovian Hawkes process, for which there is not a closed-form intensity stationary distribution available. As a consequence of Theorem 1, we can observe that the steady-state mean of the intensity is $\eta_\infty := \frac{\beta\eta^*}{\beta-\alpha}$. Interestingly, this would also be the steady-state mean of the Hawkes process when given the same baseline intensity, the same intensity jump size, and an exponential decay rate equal to the rate of expiration. This leads us to ponder how the processes would otherwise compare when given equivalent parameters. In Proposition 2 we find that although this equivalence of means extends to transient settings, for all higher moments the ESEP dominates the Hawkes process in terms of both the intensity and the counting process.

Proposition 2. *Let $(\eta_t, N_{t,\eta})$ be an ESEP intensity and counting process pair with jump size $\alpha > 0$, expiration rate $\beta > \alpha$, and baseline intensity $\eta^* > 0$. Similarly, let $(\lambda_t, N_{t,\lambda})$ be a Hawkes process intensity and counting process pair with jump size $\alpha > 0$, decay rate $\beta > 0$, and baseline intensity $\eta^* > 0$. Then, if the two processes have equal initial values, their means will satisfy*

$$\mathbb{E}[\lambda_t] = \mathbb{E}[\eta_t], \qquad \mathbb{E}\left[N_{t,\lambda}\right] = \mathbb{E}\left[N_{t,\eta}\right], \tag{3}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

and for $m \geq 2$ their mth moments are ordered so that

$$\mathbb{E}\left[\lambda_t^m\right] \leq \mathbb{E}\left[\eta_t^m\right], \qquad \mathbb{E}\left[N_{t,\lambda}^m\right] \leq \mathbb{E}\left[N_{t,\eta}^m\right], \tag{4}$$

for all time $t \geq 0$.

Proof. Let us start with the means. For the intensities, we can note that these are given by the solutions to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda_t] = \alpha\mathbb{E}[\lambda_t] - \beta(\mathbb{E}[\lambda_t] - \eta^*) \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\eta_t] = \alpha\mathbb{E}[\eta_t] - \beta\alpha\left(\frac{\mathbb{E}[\eta_t] - \eta^*}{\alpha}\right),$$

and through simplification one can quickly observe that these two ordinary differential equations (ODEs) are equivalent. Thus, because we have assumed that the processes have the same initial values, we find that $\mathbb{E}[\lambda_t] = \mathbb{E}[\eta_t]$. Since

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[N_{t,\lambda}\right] = \mathbb{E}[\lambda_t] \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[N_{t,\eta}\right] = \mathbb{E}[\eta_t],$$

this equality immediately extends to the means of the counting processes as well. We will now use these equations for the means as the base cases for inductive arguments, beginning again with the intensities. For the inductive step, we will assume that the intensity moment ordering holds for moments 1 to $m-1$. The mth moment of the Hawkes process intensity is thus given by the solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[{\lambda_t}^m\right] = \sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[{\lambda_t}^{k+1}\right]\alpha^{m-k} - m\beta\mathbb{E}\left[{\lambda_t}^m\right] + m\beta\eta^*\mathbb{E}\left[{\lambda_t}^{m-1}\right] := f_\lambda\left(t, \mathbb{E}\left[{\lambda_t}^m\right]\right),$$

where $f_\lambda\left(t, \mathbb{E}\left[\lambda_t^m\right]\right)$ is meant to capture that this ODE depends on the value of the mth moment and of the lower moments, which by the inductive hypothesis we take as known functions of the time t. Then, the mth moment of the ESEP intensity will satisfy

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[{\eta_t}^m\right] = &\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[{\eta_t}^{k+1}\right]\alpha^{m-k} + \frac{\beta}{\alpha}\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[{\eta_t}^{k+1}\right](-\alpha)^{m-k} \\ &- \frac{\beta\eta^*}{\alpha}\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[{\eta_t}^k\right](-\alpha)^{m-k}.\end{aligned}$$

By pulling the $k = m-1$ terms off the top of each summation, we can re-express this ODE as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\eta_t^m\right] = &\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[\eta_t^{k+1}\right]\alpha^{m-k} - m\beta\mathbb{E}\left[\eta_t^m\right] + m\beta\eta^*\mathbb{E}\left[\eta_t^{m-1}\right] \\ &+ \frac{\beta}{\alpha}\sum_{k=0}^{m-2}\binom{m}{k}(-\alpha)^{m-k}\left(\mathbb{E}\left[\eta_t^{k+1}\right] - \eta^*\mathbb{E}\left[\eta_t^k\right]\right),\end{aligned}$$

and through the definition of $f_\lambda(\cdot)$ and the inductive hypothesis, we can find the following lower bound:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\eta_t^m\right] \geq f_\lambda\left(t, \mathbb{E}\left[\eta_t^m\right]\right) + \frac{\beta}{\alpha}\sum_{k=0}^{m-2}\binom{m}{k}(-\alpha)^{m-k}\left(\mathbb{E}\left[\eta_t^{k+1}\right] - \eta^*\mathbb{E}\left[\eta_t^k\right]\right).$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

This rightmost term can then be expressed as

$$\begin{aligned}\frac{\beta}{\alpha}\sum_{k=0}^{m-2}\binom{m}{k}(-\alpha)^{m-k}\left(\mathbb{E}\left[\eta_t^{k+1}\right]-\eta^*\mathbb{E}\left[\eta_t^k\right]\right)\\ =\frac{\beta}{\alpha}\mathbb{E}\left[(\eta_t-\eta^*)\left((\eta_t-\alpha)^m-\eta_t^m+m\alpha\eta_t^{m-1}\right)\right],\end{aligned}$$

and we can now reason about the quantity inside the expectation. By definition, we have that $\eta_t \geq \eta^*$ with probability one, and furthermore we can observe that if $\eta_t-\eta^*>0$, then $\eta_t \geq \eta^*+\alpha>\alpha$. Thus, let us consider $(\eta_t-\alpha)^m-\eta_t^m+m\alpha\eta_t^{m-1}$ assuming $\eta_t>\alpha$. Dividing through by η_t^m, we have the expression

$$\left(1-\frac{\alpha}{\eta_t}\right)^m-1+\frac{m\alpha}{\eta_t}. \tag{5}$$

Since $(1-x)^m-1+mx$ is equal to 0 at $x=0$ and is non-decreasing on $x\in[0,\,1)$ via a first-derivative check, we can note that (5) is nonnegative for all $\eta_t>\eta^*$. Thus, we have that

$$\mathbb{E}\left[(\eta_t-\eta^*)\left((\eta_t-\alpha)^m-\eta_t^m+m\alpha\eta_t^{m-1}\right)\right]\geq 0,$$

and by consequence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\eta_t^m\right]\geq f_\lambda\left(t,\,\mathbb{E}\left[\eta_t^m\right]\right),$$

completing the proof of the intensity moment ordering via Lemma 3. For the counting processes, let us again assume as an inductive hypothesis that the moment ordering holds for moments 1 through $m-1$, with the mean equality serving as the base case. Then, the mth moment of the ESEP counting process will satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[N_{t,\eta}^m\right]=\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[\eta_t N_{t,\eta}^k\right],$$

and the ODE for the mth Hawkes counting process moment is analogous. By the Fortuin–Kasteleyn–Ginibre inequality, we can observe that

$$\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}\left[\eta_t N_{t,\eta}^k\right]\geq\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}[\eta_t]\,\mathbb{E}\left[N_{t,\eta}^k\right].$$

By the inductive hypothesis, we have that $\mathbb{E}\left[N_{t,\eta}^k\right]\geq\mathbb{E}\left[N_{t,\lambda}^k\right]$ for each $k\leq m-1$, and thus we can observe that

$$\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}[\eta_t]\,\mathbb{E}\left[N_{t,\eta}^k\right]\geq\sum_{k=0}^{m-1}\binom{m}{k}\mathbb{E}[\lambda_t]\,\mathbb{E}\left[N_{t,\lambda}^k\right].$$

Finally, by another application of Lemma 3, we have $\mathbb{E}\left[N_{t,\lambda}^m\right]\leq\mathbb{E}\left[N_{t,\eta}^m\right]$. □

The fact that the ESEP variance dominates the Hawkes variance should not be surprising, since the presence of both up- and down-jumps means that the ESEP sample paths should be subject to more abrupt changes. Nevertheless, this also shows that the ESEP is more over-dispersed than the Hawkes process is. This may be an attractive feature for data modeling.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

It is worth noting that matrix computations are available for all moments of these intensities via [21], through which one could use the method of moments to fit the processes to data.

2.3. The ephemerally self-exciting counting process

Thus far we have studied the intensity of the ESEP, as this process is by definition tracking the self-excitement. However, this excitation is manifested in the actual arrivals from the process, which are counted in N_t. We now turn our attention to developing fundamental quantities for this counting process. To begin, we give the probability generating function of the counting process in closed form below in Proposition 3. One can note that by comparison, the generating functions of the Hawkes process are instead only expressible as functions of ODEs with no known closed-form solutions; see for example Subsection 3.5 of [20].

Proposition 3. *Let N_t be the number of arrivals by time $t \geq 0$ in an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then, for $z \in [0, 1]$, the probability generating function of N_t is given by*

$$\mathbb{E}\left[z^{N_t}\right] = e^{\frac{\eta^*(\beta-\alpha)}{2\alpha}t} \left(\frac{2e^{\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right)^{\frac{\eta^*}{\alpha}}$$
$$\cdot \left(\frac{\beta+\alpha}{2\alpha} + \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} \left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right) \right)^{Q_0}, \tag{6}$$

where Q_0 is the active number in system at time 0.

Proof. Because of the cumbersome length of some equations, the proof is given in Appendix D. □

In addition to calculating the probability generating function, we can also find a matrix calculation for the transient probability mass function of the counting process. To do so, we recognize that the time until the next arrival occurs can be treated as the time to absorption in a continuous-time Markov chain. By building from this idea to construct a transition matrix for several successive arrivals, we find the form for the distribution given in Proposition 4.

Proposition 4. *Let N_t be the number of arrivals by time t in an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Further, let $Q_0 = k$ be the initial active number in system. Then for $i \in \mathbb{N}$, define the matrices $\mathbf{D}_i \in \mathbb{R}^{k+i+1 \times k+i+1}$ and $\mathbf{S}_i \in \mathbb{R}^{k+i+1 \times k+i+2}$ as*

$$\mathbf{D}_i = \begin{bmatrix} -(\eta^*+(k+i)(\alpha+\beta)) & (k+i)\beta & & & \\ & -(\eta^*+(k+i-1)(\alpha+\beta)) & & & \\ & & \ddots & & \\ & & & -(\eta^*+\alpha+\beta) & \beta \\ & & & & -\eta^* \end{bmatrix}$$

and

$$\mathbf{S}_i = \begin{bmatrix} \eta^* + \alpha(k+i) & & & & 0 \\ & \eta^* + \alpha(k+i-1) & & & 0 \\ & & \ddots & & \vdots \\ & & & \eta^* + \alpha & 0 \\ & & & & \eta^* \quad 0 \end{bmatrix}.$$

Further, let $\mathbf{Z}_n \in \mathbb{R}^{\hat{d}_n \times \hat{d}_n}$ *for* $\hat{d}_n = \frac{n(n+1)}{2} + (n+1)(k+1)$ *be a matrix such that*

$$\mathbf{Z}_n = \begin{bmatrix} \mathbf{D}_0 & \mathbf{S}_0 & & & & \\ & \mathbf{D}_1 & \mathbf{S}_1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \mathbf{S}_{n-2} & \\ & & & & \mathbf{D}_{n-1} & \mathbf{S}_{n-1} \\ & & & & & \mathbf{D}_n \end{bmatrix}.$$

Then the probability that $N_t = n$ *is given by*

$$\mathbb{P}\left(N_t = n\right) = \mathbf{v}_1{}^{\mathrm{T}} e^{\mathbf{Z}_n t} \mathbf{v}_{:} \tag{7}$$

where $\mathbf{v}_j \in \mathbb{R}^{\hat{d}_n}$ *is the unit column vector for the jth coordinate, and* $\mathbf{v}_{:} = \sum_{j=0}^{k+n} \mathbf{v}_{\hat{d}_n - j}$.

Proof. This follows directly from viewing $\mathbf{Z}_n$ as a sub-matrix of the generator matrix of a continuous-time Markov chain, much as one can do to calculate probabilities of phase-type distributions. Specifically, the sub-generator matrix is defined on the state space $\mathcal{S} = \bigcup_{i=0}^{n} \{(0, i), (1, i), \ldots, (k+i-1, i), (k+i, i)\}$. In this scenario, the state (s_1, s_2) represents having s_1 entities in system and having seen s_2 arrivals since time 0. Then, $\mathbf{D}_i$ is the sub-generator matrix for transitions among the sub-state space $\{(k+i, i), (k+i-1, i), \ldots, (1, i), (0, i)\}$ to itself (where the states are ordered in that fashion). Similarly, $\mathbf{S}_i$ is for transitions from states in $\{(k+i, i), (k+i-1, i), \ldots, (1, i), (0, i)\}$ to states in $\{(k+i+1, i+1), (k+i, i+1), \ldots, (1, i+1), (0, i+1)\}$. One can then consider this from an absorbing continuous-time Markov chain perspective, since if $n+1$ arrivals occur it is not possible to transition back to any state in which n arrivals had occurred. Hence, we only need to use the matrix $\mathbf{Z}_n$ to consider up to n arrivals. Then, $e^{\mathbf{Z}_n t}$ is the sub-matrix for probabilities of transitions among states in $\mathcal{S}$, where the rows will sum to less than 1 as it is possible that the chain has experienced more than n arrivals by time t. Finally, because $Q_0 = k$ we know that the chain starts in state $(k,0)$; further, because we are seeking the probability that there have been exactly n arrivals by time t, we want the probability of transitions from $(k, 0)$ to any of the states in $\{(k+n, n), (k+n-1, n), \ldots, (1, n), (0, n)\}$. □

With these fundamental quantities in hand, let us now turn to explore more nuanced connections between the ESEP and other stochastic processes in the following section. Doing so will provide further comparison between the Hawkes process and the ESEP, and moreover will formally connect the notion of self-excitement to similar concepts such as contagion, virality, and rich-get-richer effects.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

3. Relating ephemeral self-excitement to branching processes, random walks, and epidemics

Aside from the original definition, the most frequently utilized result for Hawkes processes is perhaps the immigration–birth representation first shown in [37]. By viewing a portion of arrivals as immigrants—externally driven and stemming from a homogeneous Poisson process—and then viewing the remaining portion as offspring—excitation-driven descendants of the immigrants and the prior offspring—one can take new perspectives on self-exciting processes. From this position, if an arrival is a descendant then it has a unique parent, the excitement of which spurred this arrival into existence. Every entity has the potential to generate offspring. This viewpoint takes on added meaning in the context of ephemeral self-excitement, as an entity only has the opportunity to generate descendants so long as it remains in the system. In this section, we will use this idea to connect self-exciting processes to well-known stochastic models that have applications ranging from public health to Bayesian statistics. Furthermore, these connections will also help us form comparisons between the Hawkes process and the model we have introduced, the ESEP. The different dynamics are at the forefront of this process comparison, as the branching structure is dictated by the self-excitation caused by a single arrival. For the Hawkes process, this increase in the arrival rate is eternal but ever-diminishing, whereas in the ESEP the jump is ephemeral but constant when it does exist.

3.1. Discrete-time perspectives through branching processes

Let us first view these processes through a discrete-time lens as branching processes. In this subsection we will interpret classical branching process results in application to these self-exciting processes. Taking the immigration–birth representation as inspiration, we start by considering the distribution of the total number of offspring of a single arrival. That is, we want to calculate the probability mass function for the number of arrivals that are generated directly from the excitement caused by the initial arrival. To constitute the *total* number of offspring, we will consider all the children of this initial entity across all time. For the ESEP, this equals the number of arrivals generated by the entity throughout its duration in the system; in the Hawkes process this counts the number of arrivals spurred by the entity as time goes to infinity. Given that the stability conditions are satisfied throughout, in Proposition 5 we calculate these distributions by way of inhomogeneous Poisson processes, yielding a Poisson mixture form for each.

Proposition 5. *Let $\beta > \alpha > 0$. Let X^η be the number of new arrivals generated by the excitement caused by an arbitrary initial arrival throughout its duration in the system in an ESEP with jump size α and expiration rate β. Then this offspring distribution is geometrically distributed with probability mass function*

$$\mathbb{P}\left(X^\eta = k\right) = \left(\frac{\beta}{\alpha+\beta}\right)\left(\frac{\alpha}{\alpha+\beta}\right)^k . \tag{8}$$

Similarly, let X^λ be the number of new arrivals generated by the excitement caused by an arbitrary initial arrival in a Hawkes process with jump size α and decay rate β. This offspring distribution is then Poisson distributed with probability mass function

$$\mathbb{P}\left(X^\lambda = k\right) = \frac{e^{-\frac{\alpha}{\beta}}}{k!}\left(\frac{\alpha}{\beta}\right)^k , \tag{9}$$

where all $k \in \mathbb{N}$.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Proof. Without loss of generality, we assume that the initial arrival in each process occurred at time 0. Then, at time $t \geq 0$, the excitement generated by these initial arrivals has intensities given by $\alpha e^{-\beta t}$ and $\alpha \mathbf{1}\{t < S\}$ for the Hawkes and ESEP processes, respectively, where $S \sim \text{Exp}(\beta)$. Using [16], one can note that the offspring distributions across all time can then be expressed as

$$X^{\lambda} \sim \text{Pois}\left(\alpha \int_0^{\infty} e^{-\beta t} \text{d}t\right) \qquad \text{and} \qquad X^{\eta} \sim \text{Pois}\left(\alpha \int_0^{\infty} \mathbf{1}\{t < S\} \text{d}t\right),$$

which are equivalently stated as $X^{\lambda} \sim \text{Pois}\left(\alpha/\beta\right)$ and $X^{\eta} \sim \text{Pois}\left(\alpha S\right)$. This now immediately yields the stated distributions for X^{λ} and X^{η}, as the Poisson–exponential mixture is known to yield a geometric distribution; see for example the overview of Poisson mixtures in [39]. □

We now move towards considering the total progeny of an initial arrival, meaning the total number of arrivals generated by the excitement of an initial arrival *and* the excitement of its offspring, and of their offspring, and so on across all time. It is important to note that in contrast to the number of offspring, the progeny includes the initial arrival itself. As we will see, the stability of the self-exciting processes implies that this total number of descendants is almost surely finite. This demonstrates the necessity of immigration for these processes to survive. From the offspring distributions in Proposition 5, the Hawkes descendant process is a Poisson branching process, and similarly the ESEP is a geometric branching process. These are well-studied models in branching processes, so we have many results available to us. In fact, we now use a result for random walks with potentially multiple simultaneous steps forward to derive the progeny distributions for these two processes. This is through the well-known hitting time theorem, stated below in Lemma 1.

Lemma 1. (Hitting time theorem.) *The total progeny Z of a branching process with descendant distribution equivalent to X_1 is*

$$\mathbb{P}\left(Z = k\right) = \frac{1}{k} \mathbb{P}\left(X_1 + X_2 + \cdots + X_k = k - 1\right),$$

where $X_1, \ldots, X_k$ are i.i.d. for all $k \in \mathbb{Z}^+$.

Proof. See [52] for the original statement and proof in terms of random walks; a review and elementary proof are given in the brief note [58]. □

We now use the hitting time theorem to give the progeny distributions for the Hawkes process and the ESEP in Proposition 6. This is a common technique for branching processes, and it now yields valuable insight into these two self-exciting models.

Proposition 6. *Let $\beta > \alpha > 0$. Let Z^{η} be a random variable for the total progeny of an arbitrary arrival in an ESEP with intensity jump α and expiration rate β. Likewise, let Z^{λ} be a random variable for the total progeny of an arbitrary arrival in a Hawkes process with intensity jump α and decay rate β. Then the probability mass functions for Z^{η} and Z^{λ} are given by*

$$\mathbb{P}\left(Z^{\eta} = k\right) = \frac{1}{k}\binom{2k-2}{k-1}\left(\frac{\beta}{\beta+\alpha}\right)^{k}\left(\frac{\alpha}{\beta+\alpha}\right)^{k-1} \text{ and } \mathbb{P}\left(Z^{\lambda} = k\right) = \frac{e^{-\frac{\alpha}{\beta}k}}{k!}\left(\frac{\alpha k}{\beta}\right)^{k-1}, \tag{10}$$

where $k \in \mathbb{Z}^+$.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

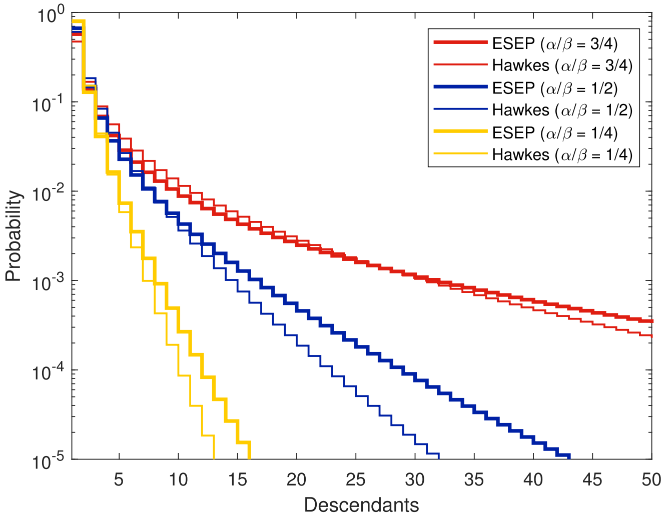


FIGURE 2. Progeny distributions for the ESEP and the Hawkes process with matching parameters.

Proof. This follows from applying Lemma 1 to Proposition 5. Because the sum of independent Poisson random variables is Poisson distributed with the sum of the rates, we have that

$$\frac{1}{k}\mathbb{P}\left(X_1^\lambda + X_2^\lambda + \dots X_k^\lambda = k-1\right) = \frac{1}{k}\mathbb{P}\left(K_1 = k-1\right),$$

where $K_1 \sim \text{Pois}\left(\frac{\alpha k}{\beta}\right)$. This now yields the expression for the probability mass function for Z^λ. Similarly for Z_η we note that the sum of independent geometric random variables has a negative binomial distribution, which implies that

$$\frac{1}{k}\mathbb{P}\left(X_1^\eta + X_2^\eta + \dots X_k^\eta = k-1\right) = \frac{1}{k}\mathbb{P}\left(K_2 = k-1\right),$$

where $K_2 \sim \text{NegBin}\left(k, \frac{\alpha}{\beta+\alpha}\right)$, and this completes the proof. □

For a visual comparison of the descendants in the ESEP and the Hawkes process, we plot these two progeny distributions for equivalent parameters in Figure 2. As suggested by the variance ordering in Proposition 2, the tail of the ESEP progeny distribution is heavier than that of the Hawkes process.

We can note that while one can calculate the mean of each progeny distribution via the probability mass functions in Proposition 6, it can also easily be found using Wald's identity. Through standard infinitesimal generator approaches, one can calculate that the expected number of arrivals (including by immigration) in the ESEP is

$$\mathbb{E}[N_t] = \frac{\beta\eta^* t}{\beta-\alpha} + \frac{\eta_0 - \eta_\infty}{\beta-\alpha}(1 - e^{-(\beta-\alpha)t}).$$

However, using these branching process representations, we can also express this as

$$\mathbb{E}[N_t] = \mathbb{E}\left[\sum_{i=1}^{M_t} Z_i(t)\right],$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

where M_t is a Poisson process with rate η^* and $Z_i(t)$ are the total progeny up to time $t \geq 0$ that are descended from the ith immigrant arrival. Now, by applying Wald's identity to the limit of $\frac{1}{t}\mathbb{E}[N_t]$ as $t \to \infty$, we see that

$$\frac{\beta\eta^*}{\beta - \alpha} = \lim_{t\to\infty} \frac{\mathbb{E}[N_t]}{t} = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{M_t} Z_i(t)\right] = \eta^*\mathbb{E}\big[Z^\eta\big],$$

and so $\mathbb{E}[Z^\eta] = \frac{\beta}{\beta-\alpha}$. By analogous arguments for the Hawkes process, we see that $\mathbb{E}\big[Z^\lambda\big] = \frac{\beta}{\beta-\alpha}$.

As a final branching process comparison between these two processes, we calculate the distribution of the total number of generations of descendants of an initial arrival in the ESEP and the Hawkes process. That is, let the first entity be the first generation, its offspring the second generation, their offspring the third, and so on. In Proposition 7 we find the probability mass function for the ESEP in closed form and a recurrence relation for the cumulative distribution function for the Hawkes process.

Proposition 7. *Let $\beta > \alpha > 0$. Let $\mathcal{G}^\eta$ be the number of distinct arrival generations across the full progeny of an initial arrival in an ESEP with intensity jump α and service rate β. Then the probability mass function for $\mathcal{G}^\eta$ is given by*

$$\mathbb{P}\left(\mathcal{G}^\eta = k\right) = \frac{\alpha^{k-1}(\beta-\alpha)}{\beta^k - \alpha^k} - \frac{\alpha^k(\beta-\alpha)}{\beta^{k+1} - \alpha^{k+1}}. \tag{11}$$

Likewise, let $\mathcal{G}^\lambda$ be the number of distinct arrival generations in the full progeny of an initial arrival for a Hawkes process with intensity jump $\alpha > 0$ and decay rate β. Then $\mathcal{G}^\lambda$ has cumulative distribution function $F_{\mathcal{G}^\lambda}(k) = \mathbb{P}\left(\mathcal{G}^\lambda \leq k\right)$ satisfying the recursion

$$F_{\mathcal{G}^\lambda}(k) = e^{-\frac{\alpha}{\beta}\left(1 - F_{\mathcal{G}^\lambda}(k-1)\right)}, \tag{12}$$

where $F_{\mathcal{G}^\lambda}(0) = 0$ and all $k \in \mathbb{Z}^+$.

Proof. Let Y_k^λ and Y_k^η be Galton–Watson branching processes defined as

$$Y_k^\lambda = \sum_{i=1}^{Y_{k-1}^\lambda} X_{\lambda,i}^{(k)}, \qquad Y_k^\eta = \sum_{i=1}^{Y_{k-1}^\eta} X_{\eta,i}^{(k)}, \tag{13}$$

with $X_{\lambda,i}^{(k)} \overset{i.i.d.}{\sim} \text{Pois}\left(\frac{\alpha}{\beta}\right)$, $X_{\eta,i}^{(k)} \overset{i.i.d.}{\sim} \text{Geo}\left(\frac{\alpha}{\alpha+\beta}\right)$, and $Y_0^\lambda = Y_0^\eta = 1$. These processes then have probability generating functions

$$\mathcal{P}_k^\lambda(z) = \sum_{j=0}^\infty z^j\, \mathbb{P}\left(Y_k^\lambda = j\right) \quad \text{and} \quad \mathcal{P}_k^\eta(z) = \sum_{j=0}^\infty z^j\, \mathbb{P}\left(Y_k^\eta = j\right)$$

that are given by the recursions $\mathcal{P}_{k+1}^\lambda(z) = \mathcal{P}_{X^\lambda}\left(\mathcal{P}_k^\lambda(z)\right)$ and $\mathcal{P}_{k+1}^\lambda(z) = \mathcal{P}_{X^\eta}\left(\mathcal{P}_k^\eta(z)\right)$ with $\mathcal{P}_1^\lambda(z) = \mathcal{P}_{X^\lambda}(z)$ and $\mathcal{P}_1^\eta(z) = \mathcal{P}_{X^\eta}(z)$, where $\mathcal{P}_{X^\lambda}(z)$ and $\mathcal{P}_{X^\eta}(z)$ are the probability generating functions of $X_{\lambda,1}^{(1)}$ and $X_{\eta,1}^{(1)}$, respectively; see e.g. Section XII.5 of [27]. One can then use induction to observe that

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

$$\mathcal{P}_k^{\eta}(z) = 1 - \frac{\alpha^k(1-z)}{\beta^k + \sum_{j=1}^{k} \alpha^j \beta^{k-j}(1-z)},$$

whereas $\mathcal{P}_k^{\lambda}(z) = e^{-\frac{\alpha}{\beta}\left(1-\mathcal{P}_{k-1}^{\lambda}(z)\right)}$, with $\mathcal{P}_1^{\lambda}(z) = e^{-\frac{\alpha}{\beta}(1-z)}$. Because of their shared offspring distribution constructions, the number of the progeny in the kth arrival generations of the Hawkes process and the ESEP are equivalent in distribution to Y_k^{λ} and Y_k^{η}, respectively. In this way, we can express $\mathcal{G}^{\lambda}$ and $\mathcal{G}^{\eta}$ as

$$\mathcal{G}^{\lambda} = \inf\{k \in \mathbb{Z}^+ \mid Y_k^{\lambda} = 0\} \quad \text{and} \quad \mathcal{G}^{\eta} = \inf\{k \in \mathbb{Z}^+ \mid Y_k^{\eta} = 0\}.$$

This leads us to observe that the events $\{\mathcal{G}^{\lambda} = j\}$ and $\{Y_j^{\lambda} = 0,\, Y_{j-1}^{\lambda} > 0\}$ are equivalent, as are $\{\mathcal{G}^{\eta} = j\}$ and $\{Y_j^{\eta} = 0,\, Y_{j-1}^{\eta} > 0\}$. Focusing for now on $\mathcal{G}^{\lambda}$, we have that

$$\begin{aligned}\mathbb{P}\left(Y_j^{\lambda} = 0,\, Y_{j-1}^{\lambda} > 0\right) &= \sum_{i=1}^{\infty} \mathbb{P}\left(X_{\lambda,1}^{(1)} = 0\right)^i \mathbb{P}\left(Y_{j-1}^{\lambda} = i\right) \\ &= \mathcal{P}_{j-1}^{\lambda}\left(\mathbb{P}\left(X_{\lambda,1}^{(1)} = 0\right)\right) - \mathbb{P}\left(Y_{j-1}^{\lambda} = 0\right),\end{aligned}$$

and since $\mathbb{P}(K = 0) = \mathcal{P}(0)$ for any nonnegative discrete random variable K with probability generating function $\mathcal{P}(z)$, this yields

$$\mathbb{P}\left(\mathcal{G}^{\lambda} = j\right) = \mathcal{P}_j^{\lambda}(0) - \mathcal{P}_{j-1}^{\lambda}(0).$$

Using $\mathcal{P}_0^{\lambda}(0) = 0$, this telescoping sum now produces the stated form of the cumulative distribution function for $\mathcal{G}^{\lambda}$. By analogous arguments for $\mathcal{G}^{\eta}$, we complete the proof. □

In the following subsection we focus on the ESEP, using the insight we have now gained from branching processes to connect this process to stochastic models for preferential attachment that are popular in the Bayesian nonparametric and machine learning literatures.

3.2. Similarities with preferential attachment and Bayesian statistics models

In the branching process perspective of the ESEP, consider the total number of active families at one point in time. That is, across all the entities present in the system at a given time, we are interested in the number of distinct families to which these entities belong. As each arrival occurs, the new entity either belongs to one of the existing families, meaning that the entity is a descendant, or it forms a new family, which is to say that it is an immigrant. If the entity is joining an existing family, it is more likely to join families that have more presently active family members.

We can note that these dynamics are quite similar to the definition of the Chinese restaurant process (CRP); see Chapter 11 in [3]. The CRP models the successive arrivals of customers to a restaurant with infinitely many tables that each have infinitely many seats. Each arriving customer chooses which table to join based on the decisions of those before. Specifically, the nth customer to arrive joins table i with probability $\frac{c_i}{n-1+\lambda}$ or otherwise starts a new table with probability $\frac{\lambda}{n-1+\lambda}$, where c_i is the number of customers at table i and $\lambda > 0$. As the number seated at table i grows larger, it is increasingly likely that the next customer will choose to sit at table i. In the ESEP, a new arrival at time $t \geq 0$ is generated as part of active excitement family i with probability $\frac{\alpha Q_{t,i}}{\alpha Q_t + \eta^*}$ and otherwise is an externally generated arrival with probability $\frac{\eta^*}{\alpha Q_t + \eta^*}$, where $Q_{t,i}$ is the number of active exciters in the system at time t in the ith excitement

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

family with $Q_t = \sum_i Q_{t,i}$. By normalizing the numerator and denominator of these probabilities by $\frac{1}{\alpha}$, we see that these dynamics match the CRP almost exactly. The difference is hardly a novel idea for restaurants—in the ESEP diners eventually leave. This departure then decreases the number of customers at the table, making it less attractive to the next person to arrive.

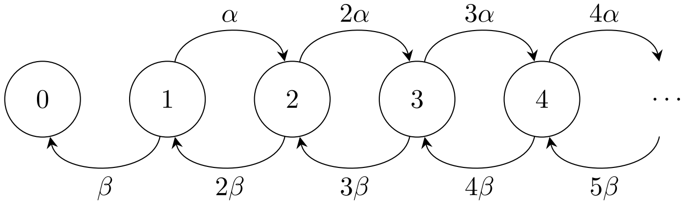
In addition to being an intriguing stochastic model, the CRP is also of interest for Bayesian statistics and machine learning through its connection to Bayesian nonparametric mixture models, specifically Dirichlet process mixtures. By consequence, the CRP then also has commonality with urn models and models for preferential attachment; see e.g. [10]. The CRP is also established enough to have its own generalizations, such as the distance-dependent CRP in [12], in which the probability that a customer joins a table is dependent on a distance metric, and the recurrent CRP in [1], in which the restaurant closes at the end of each day, forcing all of that day's customers to simultaneously depart. Drawing inspiration from the CRP and from the branching process perspectives of the ESEP, we investigate the distribution of the number of active families in the ESEP. This is equivalently stated as the number of active tables in a continuous-time CRP in which customers leave after their exponentially distributed meal durations. To begin, we first find the expected amount of time until a newly formed table becomes empty.

Proposition 8. *Suppose that an ESEP receives an initial arrival at time 0. Let X_t be the number of entities in the system at time $t \geq 0$ that are progeny of the initial arrival, and let τ be a stopping time such that $\tau = \inf\{t \geq 0 \mid X_t = 0\}$. Then the expected value of τ is*

$$\mathbb{E}[\tau] = \frac{1}{\alpha} \log\left(\frac{\beta}{\beta - \alpha}\right), \tag{14}$$

where $\alpha > 0$ is the intensity jump size and $\beta > \alpha$ is the expiration rate.

Proof. To observe this, we note that X_t can be viewed as the state of an absorbing continuous-time Markov chain on the nonnegative integers. State 0 is the single absorbing state; in any other state j, the two possible transitions are to $j+1$ at rate αj and to $j-1$ at rate μj, as visualized below.



Then, τ is the time of absorption into state 0 when starting in state 1, and so $\mathbb{E}[\tau]$ can be calculated by standard first-step analysis approaches, yielding

$$\mathbb{E}[\tau] = \sum_{i=1}^{\infty} \frac{1}{\alpha i} \prod_{j=1}^{i} \frac{\alpha j}{\beta j} = \frac{1}{\alpha} \sum_{i=1}^{\infty} \frac{1}{i} \left(\frac{\alpha}{\beta}\right)^i = \frac{1}{\alpha} \log\left(\frac{1}{1 - \frac{\alpha}{\beta}}\right),$$

and this simplifies to the stated result. □

Proposition 8 gives the expectation of the total time an excitement family is active in the system. Using this, in Proposition 9 we now employ a classical queueing theory result to find

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

the exact distribution of the number of active families simultaneously in the system in steady state.

Proposition 9. *Let B be the number of distinct excitement families that have progeny active in the system in steady state of an ESEP with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, and expiration rate $\beta > \alpha$. Then $B \sim \text{Pois}\left(\frac{\eta^*}{\alpha}\log\left(\frac{\beta}{\beta-\alpha}\right)\right)$.*

Proof. We first note that new excitement families are started when a baseline-generated arrival occurs, which follows a Poisson process with rate ν^*. The duration of the excitement family's time in system then has mean given by Proposition 8. Because there is no limitation on the number of possible families in the system at once, this is equivalent to an infinite-server queue with Poisson process arrivals and generally distributed service, an M/G/∞ queue in Kendall notation. This process is known to have Poisson distributed steady-state distribution (see e.g. [23]), with mean given by the product of the arrival rate and the mean service duration, which yields the stated form for B. □

An interesting consequence of the number of active families being Poisson distributed and the total number in system being negative-binomially distributed is that it suggests that the number of simultaneously active family members is logarithmically distributed. We observe this via the known compound Poisson representation of the negative binomial distribution [60]. For $B \sim \text{Pois}\left(\frac{\eta^*}{\alpha}\log\left(\frac{\beta}{\beta-\alpha}\right)\right)$, $Q \sim \text{NegBin}\left(\frac{\alpha}{\beta}, \frac{\eta^*}{\alpha}\right)$, and $L_i \overset{\text{i.i.d.}}{\sim} \text{Log}\left(\frac{\alpha}{\beta}\right)$, one can observe that

$$Q \overset{D}{=} \sum_{i=1}^{B} L_i,$$

where $\mathbb{P}(L_1 = k) = \left(\frac{\alpha}{\beta}\right)^k \left(k \log\left(\frac{\beta}{\beta-\alpha}\right)\right)^{-1}$ for all $k \in \mathbb{Z}^+$. Thus, the idea that the number of active members of each family is logarithmically distributed follows from the fact that this is a sum of positive integer-valued random variables, of which there are as many as there are active families, and this sum is equal to the total number in system.

3.3. Connections to epidemic models

As a final observation regarding the ESEP and its connections to other stochastic models, consider disease spread. As we discussed in the introduction to this paper, when a person becomes sick with a contagious disease she increases the rate of new infections through her contact with others. Furthermore, when a person recovers from a disease such as the flu, she is no longer contagious and thus no longer contributes to the rate of disease spread. While we have discussed in the introduction that this scenario has the hallmarks of self-excitement qualitatively, a classic model for studying this phenomenon is the susceptible–infected–susceptible (SIS) process.

In the SIS model there is a finite population of $N \in \mathbb{Z}^+$ individuals. Each individual takes on one of two states, either infected or susceptible. Let I_t be the number infected at time $t \geq 0$, and let S_t be the number susceptible. In the continuous-time stochastic SIS model, each infected individual recovers after an exponentially distributed duration of the illness. Once a person recovers from the disease, she becomes susceptible again. Because there is a finite population, the rate of new infections depends on both the number infected and the number susceptible; a new person falls ill at a rate proportional to $I_t \cdot \frac{S_t}{N}$. Because this continuous-time Markov

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

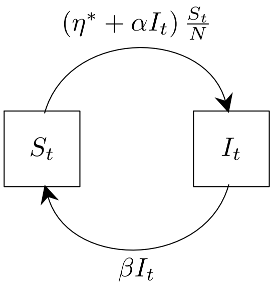


FIGURE 3. Stochastic SIS model with exogenous infections.

chain would be absorbed into state $I_t = 0$, it is common to include an exogenous infection rate proportional to just $\frac{S_t}{N}$. We will refer to this model as the stochastic SIS with exogenous infections; Figure 3 shows the rate diagram for the transitions from infected to susceptible and from susceptible to infected. For the sake of comparison, we set the exogenous infection rate as η^*, the epidemic infection rate as α, and the recovery rate as β.

One can note that there are immediate similarities between this process and the ESEP. That is, new infections increase the infection rate while recoveries decrease it, and infections can be the result of either external or internal stimuli. However, the primary difference between these two models is that the SIS process has a finite population, whereas the ESEP does not. In Proposition 10 we find that as this population size grows large the difference between these models fades, yielding that the distribution of the number infected in the exogenously driven SIS model converges to the distribution of the queue length in the ESEP.

Proposition 10. *Let I_t be the number of infected individuals at time $t \geq 0$ in an exogenously driven stochastic SIS model with population size $N \in \mathbb{Z}^+$, exogenous infection rate $\eta^* > 0$, epidemic infection rate $\alpha > 0$, and recovery rate $\beta > 0$. Then, as $N \to \infty$,*

$$I_t \overset{D}{\Longrightarrow} Q_t,$$

where Q_t is the active number in system at time t for an ESEP with baseline intensity η^*, intensity jump α, and expiration rate β.

Proof. Because the SIS model is a Markov process, one can use the infinitesimal generator approach to find a time derivative for the moment generating function of the number of infected individuals at time $t \geq 0$. Thus, by noting that $S_t = N - I_t$ we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[e^{\theta I_t}\big] &= \mathbb{E}\bigg[\frac{\alpha I_t S_t}{N}\left(e^\theta - 1\right)e^{\theta I_t} + \beta I\left(e^{-\theta} - 1\right)e^{\theta I_t} + \frac{\eta^* S_t}{N}\left(e^\theta - 1\right)e^{\theta I_t}\bigg]\\ &= \mathbb{E}\bigg[\frac{\alpha I_t(N - I_t)}{N}\left(e^\theta - 1\right)e^{\theta I_t}\bigg] + \mathbb{E}\big[\beta I_t\left(e^{-\theta} - 1\right)e^{\theta I_t}\big]\\ &\quad + \mathbb{E}\bigg[\frac{\eta^*(N - I_t)}{N}\left(e^\theta - 1\right)e^{\theta I_t}\bigg],\end{aligned}$$

which we can re-express in PDE form as

$$\begin{aligned}\frac{\partial \mathbb{E}\big[e^{\theta I_t}\big]}{\partial t} &= \left(\alpha\left(e^\theta - 1\right) + \beta\left(e^{-\theta} - 1\right) - \frac{\eta^*}{N}\left(e^\theta - 1\right)\right)\frac{\partial \mathbb{E}\big[e^{\theta I_t}\big]}{\partial \theta} - \frac{\alpha}{N}\left(e^\theta - 1\right)\frac{\partial^2 \mathbb{E}\big[e^{\theta I_t}\big]}{\partial \theta^2}\\ &\quad + \eta^*\left(e^\theta - 1\right)\mathbb{E}\big[e^{\theta I_t}\big].\end{aligned}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

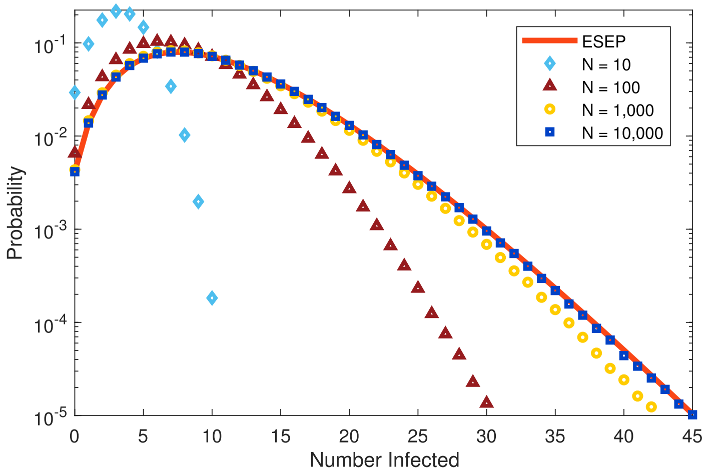


FIGURE 4. Steady-state distribution of the number infected in the exogenously driven SIS model for increasing population size N, where $\eta^* = 10$, $\alpha = 2$, and $\beta = 3$.

Now as the population size $N \to \infty$, this converges to

$$\frac{\partial \mathbb{E}\left[e^{\theta I_t}\right]}{\partial t} = \left(\alpha\left(e^{\theta} - 1\right) + \beta\left(e^{-\theta} - 1\right)\right)\frac{\partial \mathbb{E}\left[e^{\theta I_t}\right]}{\partial \theta} + \eta^*\left(e^{\theta} - 1\right)\mathbb{E}\left[e^{\theta I_t}\right],$$

which we can recognize as the PDE for the moment generating function of the ESEP through its own infinitesimal generator. □

As a demonstration of this convergence, we plot the empirical steady-state distribution of the SIS process for increasing population size below in Figure 4. Note that in this example the distributions appear fairly close for populations of size $N = 1,000$. On the scale of the populations of cities or even some larger high schools, this is quite small. At a medium university size of $N = 10,000$, the distributions are essentially indistinguishable.

We would be remiss if we did not note that connections from epidemic models to birth–death processes are not new. For example, [8] demonstrated that epidemic models converge to birth–death processes, and [56] even noted that the exogenously driven susceptible–infected–recovered (SIR) model—in which people cannot become reinfected—converges to a linear birth–death–immigration process; however, these works did not outright form connections to self-exciting processes. In [54], the similarities between the Hawkes process and the SIR process are shown and formal connections are made, although this is through a generalization of the Hawkes process defined on a finite population rather than through increasing the epidemic model population size. Regardless, these prior works serve to expand the practical relevance of the ESEP, as they show that such epidemic models are also of use outside of public health. For example, these models have been used to study the phenomenon of contagion in areas such as product adoption, idea spread, and social influence. These all also naturally relate to the concept of self-excitement, and in Proposition 10 we observe that this connection can be formalized.

These connections take on added importance in the contemporary context of the COVID-19 public health crisis. It is worth noting that the convergence of distributions in Proposition 10 does not require our commonly assumed stability condition $\beta > \alpha$. Thus, this convergence also covers potential pandemic scenarios in which $\alpha \geq \beta$. In such settings, one can quickly observe that the mean number infected by time t is such that $\mathbb{E}[N_t] \in O(e^{(\alpha-\beta)t})$, reproducing

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

the exponential growth exhibited by the novel coronavirus in these early stages. The ESEP arrival process then captures the times of new infections, some portion of which may then be used as the arrival process to a queueing model for the cases that require hospitalization.

4. Constructing eternal self-excitement from ephemeral self-excitement

Thus far we have exclusively considered a Markovian model for ephemeral self-excitement. However, just as the Hawkes process need not be Markovian, we do not have to restrict ourselves to settings in which the Markov property holds. Recall from Subsection 2.1 that the general definition from [36] described the intensity as

$$\lambda_t = \lambda^* + \int_{-\infty}^{t} g(t-u)\mathrm{d}N_{u,\lambda} = \lambda^* + \sum_{i=1}^{N_t} g(t - A_i),$$

where $g : \mathbb{R}^+ \to \mathbb{R}^+$ and $\int_0^\infty g(x)\mathrm{d}x < 1$ for stability. One could also consider a marked Hawkes process that draws jump sizes from a sequence of i.i.d. positive random variables $\{M_i \mid i \in \mathbb{Z}^+\}$, in which case the summation form of intensity would instead be expressed as

$$\lambda_t = \lambda^* + \sum_{i=1}^{N_t} M_i g(t - A_i).$$

The ESEP model has provided us a natural comparison for the popular Markovian case where $g(x) = \alpha e^{-\beta x}$ (with no marks), but let us now consider more general excitation kernels and jump sizes. To do so, we will make two main generalizations while preserving the other key elements of the ESEP, such as the affine relationship between the intensity and the number of active exciters. First, we will change the activity duration to be a general distribution, mimicking the general excitation kernel. Second, we also change from solitary arrivals to batch arrivals, meaning groups of events that occur simultaneously at each arrival epoch. The sizes of these batches may be drawn from an i.i.d. sequence of positive integer random variables, so we will use the parameter n to represent the relative size of this batch through the mean of the batch size distribution. This leads us to the definition of the *nth general ephemerally self-exciting process* (n-GESEP).

Definition 2. *(The nth general ephemerally self-exciting process.)* For times $t \geq 0$, a baseline intensity $\eta^* > 0$, a cumulative distribution function $G : \mathbb{R}^+ \to [0, 1]$, an i.i.d. sequence of positive random variables $\{B_i \mid i \in \mathbb{Z}^+\}$, and $n \in \mathbb{Z}^+$ such that $\mathbb{E}[B_i] \in O(n)$, let $N_t(n)$ be a counting process for the arrival epochs occurring according to the stochastic intensity $\eta_t(n)$, defined to be such that

$$\eta_t(n) = \eta^* + \frac{\alpha}{n} Q_t(n), \tag{15}$$

where $Q_t(n)$ is incremented by $B_i(n)$ at the ith arrival epoch in $N_t(n)$ and then is depleted at unit down-jumps at the expiration times of the individual arrivals' activity durations, which are i.i.d. draws from the distribution $G(\cdot)$ across all batches and all epochs. We say that $(\eta_t(n), N_t(n))$ is the *nth general ephemerally self-exciting process* (n-GESEP).

It is important to note that in this definition, a batch of size n occurring at the current time would increase the present arrival rate by α. However, these n activity durations are mutually independent. Thus, despite the common arrival epoch, each expiration should cause an instantaneous decrease of just α/n if the activity duration distribution is continuous. It is also worth

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

noting that this n-GESEP model encapsulates the simple generalization of the ESEP with general service durations, as this is given by $(\eta_t(1), N_t(1))$ with $\mathbb{P}\,(B_1(1) = 1)$. Just as the ESEP could be viewed as a Markovian infinite-server queue with a state-dependent arrival rate, the n-GESEP can be seen as an $\mathrm{M}^B/\mathrm{G}/\infty$ queue with a state-dependent arrival rate. This perspective implies the existence and uniqueness of Definition 2.

Just as it is often beneficial to think of the length of an infinite-server queue as a sum over all customers that remain in service, it will be quite useful for us to think about the number active within the n-GESEP as the sum of all arrivals that have not yet expired. For A_i as the ith arrival epoch and $S_{i,j}$ as the jth activity duration within the ith batch, this can be expressed as

$$Q_t(n) = \sum_{i=1}^{N_t(n)} \sum_{j=1}^{B_i} \mathbf{1}\{t < A_i + S_{i,j}\},$$

or equivalently for the intensity,

$$\eta_t = \eta^* + \frac{\alpha}{n} \sum_{i=1}^{N_t(n)} \sum_{j=1}^{B_i} \{t < A_i + S_{i,j}\}.$$

A nice consequence of this representation is that the ephemerality is at the forefront. Because the event within the indicator is that the current time t is less than the sum of a given arrival time and activity duration, this indicator counts whether that particular excitement is currently active. Thus, this indicator will switch from 1 to 0 when the present time passes a given expiration time, causing the intensity to drop by α/n. From this point forward, the excitement brought by this particular arrival no longer has a direct effect on the system.

We will now use the n-GESEP to establish a main result of this work, in which we connect this general form of ephemeral self-excitement to general forms of the traditional, eternal notion of self-excitement. In Theorem 2, we prove a scaling limit that incorporates random batch distributions and general service to construct marked, general decay Hawkes processes. We refer to this limit as a ‘batch scaling’, as we are letting the relative batch size n grow large, which simultaneously shrinks the size of the excitement generated by each individual entity within a batch of arrivals. Thus, while the effect of one individual exciter shrinks, the collective effect of a batch arrival remains fixed. In the limit, this provides an alternate construction of the Hawkes process.

Theorem 2. *For $t \geq 0$ and $n \in \mathbb{Z}^+$, let $(\eta_t(n), N_t(n))$ be an n-GESEP with baseline intensity $\eta^* > 0$, intensity jump size $\alpha > 0$, activity duration cumulative distribution function $G(\cdot)$, and i.i.d. batch size sequence of nonnegative, discrete random variables $\{B_i \mid i \in \mathbb{Z}^+\}$. Additionally, let $\eta_0(n) = \eta^*$, i.e. $Q_0(n) = 0$. Then, for all $t \geq 0$, this n-GESEP process is such that*

$$\eta_t(n) \overset{D}{\Longrightarrow} \lambda_t \text{ and } N_t(n) \overset{D}{\Longrightarrow} N_{t,\lambda} \tag{16}$$

as $n \to \infty$, where $(\lambda_t, N_{t,\lambda})$ is the general Hawkes process intensity and counting process pair such that

$$\lambda_t = \eta^* + \sum_{i=1}^{N_{t,\lambda}} M_i \bar{G}(t - A_i), \tag{17}$$

where $\{A_i \mid i \in \mathbb{Z}^+\}$ are the Hawkes process arrival epochs, $\bar{G}(x) = 1 - G(x)$ for all $x \geq 0$, and $\{M_i \mid i \in \mathbb{Z}^+\}$ is an i.i.d. sequence of positive real random variables such that $\alpha B_1/n \overset{D}{\Longrightarrow} M_1$ and $B_1/n^2 \overset{p}{\longrightarrow} 0$.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Proof. We will organize the proof into two parts. Each part is oriented around the process arrival times, as these fully determine the sample path of the Hawkes process. We will first show through induction that the distributions of inter-arrival times converge. Then we will demonstrate that given the same arrival times, the dynamics of the processes converge.

To begin, let A_i^η for $i \in \mathbb{Z}^+$ be the time of the ith arrival in the n-GESEP, and similarly let A_i^λ be the ith arrival time for the Hawkes process. We start with the base case: for the time of the first arrival, we can note that for all n-GESEP models,

$$\mathbb{P}\left(A_1^\eta > x\right) = e^{-\eta^* x},$$

as $Q_0(n) = 0$ and thus the first arrival is driven by the external baseline rate. Likewise, for the Hawkes process, since (17) implies that $\lambda_t = \nu^*$ for $0 \le t < A_1^\lambda$, we can see that

$$\mathbb{P}\left(A_1^\lambda > x\right) = e^{-\eta^* x},$$

and thus $\mathbb{P}\left(A_1^\lambda > x\right) = \mathbb{P}\left(A_1^\eta > x\right)$. As an inductive hypothesis, we now assume that $\{A_1^\eta, \ldots, A_k^\eta\}$ converge in joint and marginal distributions to $\{A_1^\lambda, \ldots, A_k^\lambda\}$ where $k \in \mathbb{Z}^+$. Now, for the Hawkes process we can observe that

$$\mathrm{P}_k\left(A_{k+1}^\lambda - A_k^\lambda > x\right) := \mathbb{P}\left(A_{k+1}^\lambda - A_k^\lambda > x \mid \{A_1^\lambda, \ldots, A_k^\lambda\}\right) = \mathrm{E}_k\left[e^{-\int_0^x \lambda_{A_k^\lambda + t} \mathrm{d}t}\right],$$

because when conditioned on the arrival history, the Hawkes process behaves like an inhomogeneous Poisson process until the next arrival occurs. Using (17), we can express this as

$$\begin{aligned}\mathrm{E}_k\left[e^{-\int_0^x \lambda_{A_k^\lambda + t} \mathrm{d}t}\right] &= e^{-\eta^* x} \mathrm{E}_k\left[e^{-\int_0^x \sum_{i=1}^k M_i \bar{G}(A_k^\lambda - A_i^\lambda + t) \mathrm{d}t}\right] \\ &= e^{-\eta^* x} \prod_{i=1}^k \mathrm{E}_k\left[e^{-M_i \int_0^x \bar{G}(A_k^\lambda - A_i^\lambda + t) \mathrm{d}t}\right].\end{aligned}$$

Turning to the ESEP epochs, we define $N_{i,j}^\eta\left((t, t+s]\right)$ as the number of arrivals on the time interval $(t, t+s]$ that are generated by the excitement caused by the jth entity within the ith batch. Furthermore, let $N_*^\eta\left((t, t+s]\right)$ be the number of arrivals on $(t, t+s]$ that are generated by the external, baseline rate η^*. Then, using this notation, we have that

$$\begin{aligned}\mathrm{P}_k\left(A_{k+1}^\eta - A_k^\eta > x\right) &:= \mathbb{P}\left(A_{k+1}^\eta - A_k^\eta > x \mid \{A_1^\eta, \ldots, A_k^\eta\}\right) \\ &= \mathrm{P}_k\left(\bigcap_{i=1}^k \bigcap_{j=1}^{B_i} \left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\} \cap \left\{N_*^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right) \\ &= \mathrm{E}_k\left[\prod_{i=1}^k \prod_{j=1}^{B_i} \mathbf{1}\left\{N_{i,j}^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\} \mathbf{1}\left\{N_*^\eta\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right].\end{aligned}$$

From the independence of each of these arrival processes, we can move the probability for no arrivals in the external arrival process and the product over i outside of the expectation to obtain

$$\mathrm{E}_k\left[\prod_{i=1}^{k}\prod_{j=1}^{B_i}\mathbf{1}\left\{N_{i,j}^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\mathbf{1}\left\{N_*^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right]$$
$$= e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\mathbf{1}\left\{N_{i,j}^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right].$$

Consider an arbitrary entity, say the jth entity in the ith batch. Let $S_{i,j}$ be its service duration. If this entity has departed from the queue before A_k^η, then it cannot generate further arrivals, and thus

$$\mathrm{P}_k\left(N_{i,j}^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0 \mid S_{i,j} \le A_k^\eta - A_i^\eta\right) = 1.$$

Likewise, if it does not depart until after $A_k^\eta + x$, then the probability that it generates an arrival on $(A_k^\eta, A_k^\eta + x]$ is

$$\mathrm{P}_k\left(N_{i,j}^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0 \mid S_{i,j} \ge A_k^\eta - A_i^\eta + x\right) = e^{-\frac{\alpha}{n}x}.$$

Finally, if the entity departs in the interval $(A_k^\eta, A_k^\eta + x]$, the probability that it generates an arrival before departing is

$$\mathrm{P}_k\left(N_{i,j}^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0 \mid S_{i,j} = A_k^\eta - A_i^\eta + z\right) = e^{-\frac{\alpha}{n}z},$$

where $0 < z < x$. Therefore, through conditioning on each entity's service duration, we have that

$$e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\mathbf{1}\left\{N_{i,j}^{\eta}\left((A_k^\eta, A_k^\eta + x]\right) = 0\right\}\right]$$
$$= e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)\right],$$

where $g(\cdot)$ is the density corresponding to $G(\cdot)$. Since the term inside the inner product does not depend on the specific entity within a batch but only on the batch itself, we can evaluate this inside the expectation as

$$e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\prod_{j=1}^{B_i}\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)\right]$$
$$= e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)^{B_i}\right].$$

Since the base term of this exponent is deterministic, we will simplify it as follows. Using integration by parts on $\int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z$ and expanding $\bar{G}(x) = 1 - G(x)$, we obtain

$$G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta + x) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z$$
$$= e^{-\frac{\alpha}{n}x} + \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z} G(A_k^\eta - A_i^\eta + z)\mathrm{d}z.$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

If we express $e^{-\frac{\alpha}{n}x}$ in integral form via $e^{-\frac{\alpha}{n}x} = 1 - \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z}\mathrm{d}z$, we can further simplify this expression for the base to

$$e^{-\frac{\alpha}{n}x} + \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z} G(A_k^\eta - A_i^\eta + z)\mathrm{d}z = 1 - \frac{\alpha}{n}\int_0^x e^{-\frac{\alpha}{n}z}\bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z.$$

This form makes it quick to observe that this base term is at most 1. Thus we are justified in taking the expectation of this term raised to B_i, since that is equivalent to the probability generating function of the batch size, and this exists for all discrete random variables when evaluated on values less than or equal to 1 in absolute value. Returning to this expectation, we first note that for all x, rearranging the Taylor expansion of e^x produces

$$1 + x = e^x - \sum_{j=2}^{\infty}\frac{x^j}{j!} = e^x\left(1 - e^{-x}\sum_{j=2}^{\infty}\frac{x^j}{j!}\right) = e^{x+\log\left(1-e^{-x}\sum_{j=2}^{\infty}\frac{x^j}{j!}\right)}.$$

Thus we re-express the expectation in exponential function form as

$$\begin{aligned} &e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[\left(G(A_k^\eta - A_i^\eta) + e^{-\frac{\alpha}{n}x}\bar{G}(A_k^\eta - A_i^\eta) + \int_0^x e^{-\frac{\alpha}{n}z} g(A_k^\eta - A_i^\eta + z)\mathrm{d}z\right)^{B_i}\right] \\ &\quad = e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x e^{-\frac{\alpha}{n}z}\bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right]. \end{aligned}$$

Through use of a Taylor expansion on $e^{-\frac{\alpha}{n}z}$ and absorbing higher terms into the $O\left(\frac{B_i}{n^2}\right)$ notation, we can further simplify to

$$\begin{aligned} &e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x e^{-\frac{\alpha}{n}z}\bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right] \\ &\quad = e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x \bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right]. \end{aligned}$$

We can now take the limit as $n \to \infty$ and observe that

$$e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-\frac{\alpha}{n}B_i\int_0^x \bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z + O\left(\frac{B_i}{n^2}\right)}\right] \longrightarrow e^{-\eta^* x}\prod_{i=1}^{k}\mathrm{E}_k\left[e^{-M_i\int_0^x \bar{G}(A_k^\eta - A_i^\eta + z)\mathrm{d}z}\right],$$

as we have that $\frac{\alpha}{n}B_1 \overset{D}{\Longrightarrow} M_1$ and $\frac{B_i}{n^2} \overset{D}{\Longrightarrow} 0$. This is equal to the Hawkes process inter-arrival probability $\mathrm{P}_k\left(A_{k+1}^\lambda - A_k^\lambda > x\right)$. Hence by induction and total probability the arrival times converge, completing the first part of the proof.

For the second part of the proof, we show that the dynamics of the processes converge when we condition on having the same fixed arrival times, which we now denote by $\{A_i \mid i \in \mathbb{Z}^+\}$ for both processes. Since $N_t(n)$ is defined as the counting process of arrival epochs rather than total number of arrivals, $N_t(n) = N_{t,\lambda}$ for all n and all t. We now treat the intensity in two cases, the jump at arrivals and the dynamics between these times. For the first case, we take $k \in \mathbb{Z}^+$ and let $\lambda_{A_{k-}} = \inf_{A_{k-1}\le t < A_k}\lambda_t$ and $\eta_{A_{k-}}(n) = \inf_{A_{k-1}\le t<A_k}\eta_t(n)$ for all n, where $A_0 = 0$. Then the jump in the n-GESEP intensity at the kth jump is such that

$$\eta_{A_k}(n) - \eta_{A_{k-}}(n) = \frac{\alpha}{n}B_k \overset{D}{\Longrightarrow} M_k = \lambda_{A_k} - \lambda_{A_{k-}}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

as $n \to \infty$. For the behavior between arrival times we first note that for S_j independent and distributed with cumulative distribution function $G(\cdot)$ for all $j \in \mathbb{Z}^+$, the probability generating function of $\frac{1}{n}\sum_{j=1}^{B_1} \mathbf{1}\{y < S_j\}$ is

$$\mathbb{E}\left[z^{\frac{1}{n}\sum_{j=1}^{B_1}\mathbf{1}\{y<S_j\}}\right] = \mathbb{E}\left[\left(G(y) + \bar{G}(y) z^{\frac{1}{n}}\right)^{B_1}\right] = \mathbb{E}\left[\left(1 - \bar{G}(y)\left(1 - e^{\frac{1}{n}\log z}\right)\right)^{B_1}\right],$$

and by a Taylor expansion approach similar to what we used in the first part of the proof, we can see that

$$\begin{aligned}\mathbb{E}\left[\left(1 - \bar{G}(y)\left(1 - e^{\frac{1}{n}\log z}\right)\right)^{B_1}\right] &= \mathbb{E}\left[e^{-B_1\bar{G}(y)\left(1-e^{\frac{1}{n}\log z}\right) + O\left(\frac{B_1}{n^2}\right)}\right] \\ &= \mathbb{E}\left[e^{\frac{1}{n}B_1\bar{G}(y)\log(z) + O\left(\frac{B_1}{n^2}\right)}\right].\end{aligned}$$

Taking the limit as $n \to \infty$, this yields

$$\mathbb{E}\left[z^{\frac{1}{n}\sum_{j=1}^{B_1}\mathbf{1}\{y<S_j\}}\right] \longrightarrow \mathbb{E}\left[z^{\bar{G}(y)M_1}\right],$$

which is to say that

$$\frac{1}{n}\sum_{j=1}^{B_1}\mathbf{1}\{y < S_j\} \overset{D}{\Longrightarrow} \bar{G}(y)M_1.$$

Using this, we can now see that for $k \in \mathbb{Z}^+$ and $0 \le x < A_{k+1} - A_k$, the intensity of the n-GESEP satisfies

$$\begin{aligned}\eta_{A_k+x}(n) &= \eta^* + \frac{\alpha}{n}\sum_{i=1}^{k}\sum_{j=1}^{B_i}\mathbf{1}\{A_k + x < A_i + S_{i,j}\} \\ &\overset{D}{\Longrightarrow} \eta^* + \sum_{i=1}^{k} M_i\bar{G}(A_k - A_i + x) \\ &= \lambda_{A_k+x}\end{aligned}$$

as $n \to \infty$. Thus, both the jump sizes of $\eta_t(n)$ and the behavior of $\eta_t(n)$ between jumps converge to those of λ_t, completing the proof. □

For an empirical demonstration of this convergence, in Figure 5 we plot cumulative distribution functions for the intensities and counts of the n-GESEP across increasing batch sizes and multiple expiration distributions and compare them to the empirical distributions of the corresponding general Hawkes processes. Specifically, we consider three different choices of $\bar{G}(\cdot)$, all with a unit mean expiration time: a standard exponential $\bar{G}(x) = e^{-x}$, a two-phase Erlang distribution $\bar{G}(x) = (1 + 2x)e^{-2x}$, and a two-phase hyper-exponential $\bar{G}(x) = (e^{-2x} + e^{-2/3x})/2$. Each experiment is conducted for deterministic batch sizes in the n-GESEP models and deterministic marks in the Hawkes processes. Similar performance was observed for other distributions such as the geometric (or exponential in limiting form) and are thus

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

TABLE 1. Overview of convergence details in the batch scaling of the ESEP.

	$n \longrightarrow \infty$	
Batch	$\Longrightarrow$	Mark
Expire	$\Longrightarrow$	Decay
ESEP	$\Longrightarrow$	Hawkes

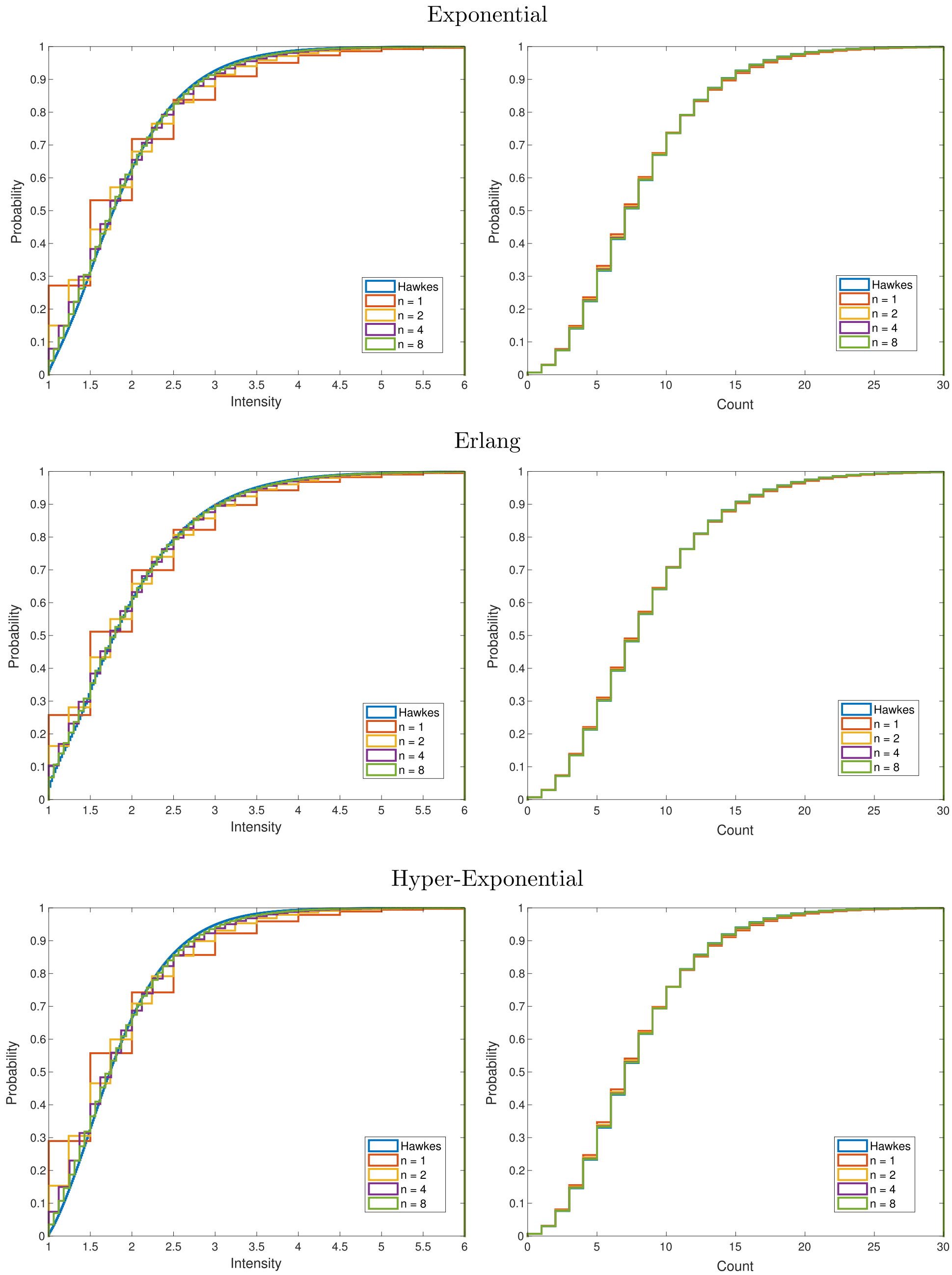
omitted for brevity. In each of the three distribution settings with batch size $n = 8$ in the n-GESEP model, the distribution of the intensity is quite close to that of the Hawkes intensity, as can be seen in the left-hand column. On the right, one can note that the count distributions are quite close in all the various settings, and again the performance is particularly strong by $n = 8$. The visible closeness of these count distributions—indeed, they are nearly indistinguishable—lends added importance to the calculations in Proposition 4. Given these observations, it is of particular interest to study the convergence rates of these models in future work.

As a reference, we list the components of the ephemerally self-exciting models and their corresponding limiting quantities in the general Hawkes process below in Table 1. We can note that because the limiting excitation kernel given in Theorem 2 is a complementary cumulative distribution function it is exclusively non-increasing, meaning that the excitement after each arrival immediately decays. It can be observed that this includes the two most popular excitation kernels, the exponential and power-law forms that we detailed in Subsection 2.1. However, it does not include kernels that have the 'hump remote from the origin' that Hawkes mentions briefly in the original paper [36]. If desired, this can be remedied through extension to multi-phase service in the n-GESEP, with the intensity defined as an affine relationship with one of the later phases.

Before concluding, let us remark that in addition to providing conceptual understanding into the Hawkes process itself, the alternate construction through the batch scaling in Theorem 2 is also of practical relevance in explaining the use of the Hawkes process in many applications. For example, in biological applications such as the environmental management problem considered in [32], one of the invasive species studied may produce multiple offspring simultaneously but only for the duration of its life cycle. That is, many species give birth in litters, creating batch arrivals, but of course reproduce only during their lifetimes, yielding ephemerality. Furthermore, the numerical experiments in Figure 5 suggest that n need not be overly large for the n-GESEP and the Hawkes process to be comparable in distribution.

As another example, consider the spread of information on communication and social media platforms. This setting has recently been a popular application of Hawkes processes; see e.g. [22, 26, 34, 54, 55]. When a user shares a post on these platforms, it is immediately and simultaneously dispatched to the real-time feeds of many other users, creating a batch increase of the response rate from the other users. The post then will typically only be seen on news feeds for a short period of time, as new content comes in to replace it. On top of this, social media administrators have been adopting a trend of intentionally introducing ephemerality into their platforms. For example, the expiration of posts and messages has been a defining feature of Snapchat since its inception. Facebook and Instagram have recently adopted the same behavior with 'stories', and Twitter has responded in kind with the appropriately named 'fleets'. Just as in the case of biological offspring processes, Theorem 2 offers an explanation of why the Hawkes process has become a popular and successful model in this space. Moreover,

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press



Exponential

Erlang

Hyper-Exponential

FIGURE 5. Empirical transient distributions of the intensities (left) and counts (right) of n-GESEP with increasing batch size and for general Hawkes process simulations with $t = 5$, $\eta^* = 1$, $\alpha = 0.5$, and varying expiration distributions. These distributions are calculated from 2^{22} replications.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

the insights from these connections can be deepened through the additional model relationships that we have discussed in Section 3.

5. Conclusion

Time is fleeting; excitement is ephemeral. In this paper we have introduced the *ephemerally self-exciting process* (ESEP), a point process driven by the pieces of its history that remain presently active. That is, each arrival excites the arrival rate until the expiration of its randomly drawn activity duration, at which point its individual influence vanishes. Throughout this work we have compared ephemeral self-excitement to eternal self-excitement through contrast with the well-known Hawkes process. These comparisons include an ordering of the moments of the two processes in Proposition 2 and a study of each process's branching structure in Subsection 3.1. We have also used the ESEP to relate ephemeral self-excitement to other well-known stochastic models, including preferential attachment, random walks, and epidemics. Finally, we have also considered a generalized model with batch arrivals and general activity duration distributions, which we refer to as the *nth general ephemerally self-exciting process* (n-GESEP). The n-GESEP model provides an alternate construction of general marked Hawkes processes through a batch scaling limit. ln Theorem 2 we prove that the n-GESEP model converges to a Hawkes process as its batch arrival size grows large, with the limiting Hawkes process having an excitation kernel matching the tail cumulative distribution function of the activity duration distribution, and having marks given by the scaled limit of the batch sizes. As we have discussed, this limit both provides intuition for the occurrence of self-excitement in natural phenomena and relates the Hawkes process to the other stochastic models we connected to ephemeral self-excitement.

This presents many directions for future research. First, we have frequently emphasized in this work that our results motivate the ESEP (and by extension the n-GESEP) as a promising model for self-excitement in its own right. This follows from its tractability for analysis and its amenability to connection with other models. Because of this promise, we believe that further exploration and application of the ESEP holds great potential. Another natural and relevant avenue would be to continue to explore the connection between ephemeral self-excitement and the other stochastic models we have discussed. For example, one could study the connection between ESEP and epidemic models on more complex networks or with more complex dynamics. Doing so would give a point process representation for the times of infection in a more realistic epidemic model, which could be quite useful in practice for resource allocation and policy design. Similar deepened connections could also be pursued for other models, such as preferential attachment. In general, we believe the concept of ephemeral self-excitement merits further theoretical exploration and detailed empirical application, both of which we look forward to pursuing. Finally, to empower this model for full practical use, it is of great interest to study the estimation of the ESEP and n-GESEP processes. For example, likelihood-based estimation could be pursued in a manner similar to that used for the traditional Hawkes process, since the processes are also conditionally non-stationary Poisson when given the full history. In fact, similar estimation techniques may be achievable even if the data contains only arrival epochs (and not expiration times), perhaps through missing data techniques such as expectation–maximization algorithms. These methods have been applied successfully both for Hawkes processes and generally for branching processes (see, e.g., [43, 59]), and thus also hold potential for the ephemerally self-exciting models we have studied here.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Appendix A. Lemmas and auxiliaries

In this section of the appendix we give technical lemmas to support our analysis and brief auxiliary results that are of interest but not within the narrative of the body of this report. We begin by giving the infinitesimal generator form for time derivatives of the expectations of functions of our process. This is a valuable tool available to us because the ESEP is Markov, and it supports much of our analysis throughout this work.

Lemma 2. *For a sufficiently regular function* $f : \left(\mathbb{R}^+ \times \mathbb{N}\right) \to \mathbb{R}$, *the generator of the ESEP is given by*

$$\mathcal{L}f(\eta_t, N_t) = \underbrace{\sum_{i=1}^{n} \eta_t \left(f(\eta_t + \alpha, N_t + 1) - f(\eta_t, N_t)\right)}_{\text{Arrivals}} \\ + \underbrace{\beta\left(\frac{\eta_t - \eta^*}{\alpha}\right) \left(f\left(\eta_t - \alpha, N_t\right) - f(\eta_t, N_t)\right)}_{\text{Expirations}} . \tag{18}$$

Then, the time derivative of the expectation of $f(\eta_t, N_t)$ *is given by*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[f(\eta_t, N_t)\big] = \mathbb{E}\big[\mathcal{L}f(\eta_t, N_t)\big] \tag{19}$$

for all $t \geq 0$.

Proof. This is a direct result of the ESEP belonging to the family of piecewise deterministic Markov processes, as defined in [19]. Moreover, the specific regularity conditions are given in Theorem 5.5 of that work. □

Note that this is also immediately applicable to the active-number-in-system process perspective on the ESEP, as $Q_t = (\eta_t - \eta^*)/\alpha$. Thus, we leverage this infinitesimal generator to study each of η_t, Q_t, and N_t throughout both the main body of the text and these appendices.

As another supporting lemma, let us summarize a result that can be used with the infinitesimal generator to relate two different Markov processes. Throughout this work we make comparisons between different processes, in particular between the ESEP and the Hawkes process. One way that we do this is by investigating the differential equations found with the use of Lemma 2. In Lemma 3 we provide the method by which we make such comparisons.

Lemma 3. (A comparison lemma.) *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be a continuous function in both variables. If we assume that the initial value problem*

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(t, x(t)), \ \ x(0) = x_0 \tag{20}$$

has a unique solution for the time interval [0, T], and

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} \leq f(t, y(t)) \quad \text{for } t \in [0, T] \text{ and } y(0) \leq x_0, \tag{21}$$

then $x(t) \geq y(t)$ *for all* $t \in [0, T]$.

Proof. The proof of this result is given in [33]. □

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

For a result that is auxiliary on the surface but also beneficial in proofs, in Proposition 11 we give the probability generating function for the number in system and the number of departures, or expirations, in the ESEP. The departure process is largely outside of the scope of this work, but this result is instrumental in the proof (found in Appendix D) of Proposition 3, which gives the probability generating function for the counting process.

Proposition 11. *Let Q_t be the active number in system at time $t \geq 0$ of an ESEP with baseline intensity $\eta^* > 0$, intensity jump size $\alpha > 0$, and expiration rate $\beta > \alpha$. Let D_t be the number of arrivals by time t that are no longer active. Then the joint probability generating function of Q_t and D_t, denoted by $G(z_1, z_2, t) \equiv \mathbb{E}\left[z_1^{Q_t} z_2^{D_t}\right]$, is given by*

$$G(z_1, z_2, t) = \left(1 - \left(\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z_2} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z_1}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right)\right)^2\right)^{\frac{\eta^*}{2\alpha}}$$
$$\cdot \left(\frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z_1}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right)\right)^{Q_0}$$
$$\cdot \left(\cosh\left(\tanh^{-1}\left(\frac{2\alpha z_1-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right)\right)^{\frac{\eta^*}{\alpha}} z_2^{D_0} e^{\frac{\eta^*(\beta-\alpha)}{2\alpha}t}, \tag{22}$$

where Q_0 and D_0 are the active number in the system and the count of departures at time 0, respectively.

Proof. We will show this through the method of characteristics. We can first observe through Lemma 2 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[z_1^{Q_t} z_2^{D_t}\right] = \mathbb{E}\left[(\eta^* + \alpha Q_t)(z_1 - 1) z_1^{Q_t} z_2^{D_t} + \beta Q_t \left(\frac{z_2}{z_1} - 1\right) z_1^{Q_t} z_2^{D_t}\right],$$

and so $G(z_1, z_2, t)$ is given by the following PDE:

$$\frac{\partial}{\partial t} G(z_1, z_2, t) + \left(\alpha(z_1 - z_1^2) + \beta(z_1 - z_2)\right) \frac{\partial}{\partial z_1} G(z_1, z_2, t) = \eta^*(z_1 - 1) G(z_1, z_2, t).$$

To simplify our analysis, we will instead solve for $\log(G(z_1, z_2, t))$, which through the chain rule will be given by the solution to the PDE expressed as

$$\frac{\partial}{\partial t}\log(G(z_1, z_2, t)) + \left(\alpha(z_1 - z_1^2) + \beta(z_1 - z_2)\right) \frac{\partial}{\partial z_1}\log(G(z_1, z_2, t)) = \eta^*(z_1 - 1),$$

with initial condition $\log(G(z_1, z_2, 0)) = \log(z_1^{Q_0} z_2^{D_0})$. This gives us the characteristic equations as follows:

$$\begin{aligned}
&\frac{\mathrm{d}z_1}{\mathrm{d}s}(r, s) = \alpha(z_1 - z_1^2) + \beta(z_1 - z_2), && z_1(r, 0) = r,\\
&\frac{\mathrm{d}t}{\mathrm{d}s}(r, s) = 1, && t(r, 0) = 0,\\
&\frac{\mathrm{d}g}{\mathrm{d}s}(r, s) = \eta^*(z_1 - 1), && g(r, 0) = \log(r^{Q_0} z_2^{D_0}).
\end{aligned}$$

Solving the first two equations, we see that

$$\begin{aligned} z_1(r,s) &= \frac{\beta+\alpha}{2\alpha} + \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}{2\alpha}\tanh\left(\frac{s}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}\right. \\ &\qquad \left. - \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha r}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right) \\ t(r,s) &= s, \end{aligned}$$

which allows us to solve for $g(r, s)$. Using the solution to $z_1(r, s)$, the ODE for $g(r, s)$ is given by

$$\begin{aligned} \frac{\mathrm{d}g}{\mathrm{d}s}(r,s) &= \frac{\eta^*\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}{2\alpha}\tanh\left(\frac{s}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}\right. \\ &\qquad \left. - \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha r}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right) + \frac{\eta^*(\beta-\alpha)}{2\alpha}, \end{aligned}$$

which yields a solution of

$$\begin{aligned} g(r,s) &= \log\left(r^{Q_0} z_2^{D_0}\right) + \frac{\eta^*(\beta-\alpha)}{2\alpha}s + \frac{\eta^*}{2\alpha}\log\left(1 - \frac{(\beta+\alpha-2\alpha r)^2}{(\beta+\alpha)^2-4\alpha\beta z_2}\right) \\ &\quad + \frac{\eta^*}{\alpha}\log\left(\cosh\left(\frac{s}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2} - \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha r}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right)\right). \end{aligned}$$

Now, from these equations we can express the characteristic variables in terms of the original arguments as $s = t$ and

$$\begin{aligned} r &= \frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}\right. \\ &\qquad \left. - \tanh^{-1}\left(\frac{2\alpha z_1-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z_2}}\right)\right). \end{aligned}$$

Then, by performing the substitution $G(z_1, z_2, t) = e^{g(r(z_1,z_2,t),s(z_1,z_2,t))}$ and simplifying, we achieve the stated result. □

As another auxiliary result, in Proposition 12 we give the steady-state moment generating function for the 2-GESEP with exponentially distributed activity durations and deterministic batch sizes, meaning pairs of arrivals.

Proposition 12. *Consider the following 2-GESEP: arrivals occur at rate* $\eta_t(2) = \eta^* + \frac{\alpha}{2}Q_t(2)$, *where* $Q_t(2)$ *receives arrivals in batches of size 2. Each activity duration is independent and exponentially distributed with rate* $\beta > \alpha > 0$. *Then the steady-state moment generating function of* $Q_t(2)$ *is given by*

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

$$
\begin{aligned}
\mathbb{E}\Big[e^{\theta Q_\infty(2)}\Big] &\equiv \lim_{t\to\infty} \mathbb{E}\Big[e^{\theta Q_t(2)}\Big] \\
&= \exp\left(\frac{2\eta^*}{\sqrt{\alpha(\alpha+8\beta)}}\left(\tanh^{-1}\left(\left(2e^{\theta}+1\right)\sqrt{\frac{\alpha}{\alpha+8\beta}}\right) - \tanh^{-1}\left(3\sqrt{\frac{\alpha}{\alpha+8\beta}}\right)\right)\right) \\
&\quad \cdot \left(\frac{2\beta-2\alpha}{2\beta-\alpha(e^{\theta}+e^{2\theta})}\right)^{\frac{\eta^*}{\alpha}}.
\end{aligned}
\tag{23}
$$

Proof. Using Lemma 2, we see that the moment generating function will be given by the solution to

$$
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[e^{\theta Q_t(2)}\Big] = \mathbb{E}\left[\left(\eta^* + \frac{\alpha Q_t(2)}{2}\right)\left(e^{\theta(Q_t(2)+2)} - e^{\theta Q_t(2)}\right) + \beta Q_t(2)\left(e^{\theta(Q_t(2)-1)} - e^{\theta Q_t(2)}\right)\right],
$$

which can be equivalently expressed in PDE form as

$$
\frac{\partial}{\partial t}\mathcal{M}_2(\theta, t) = \eta^*\left(e^{2\theta}-1\right)\mathcal{M}_2(\theta, t) + \left(\frac{\alpha}{2}\left(e^{2\theta}-1\right) + \beta\left(e^{-\theta}-1\right)\right)\frac{\partial}{\partial\theta}\mathcal{M}_2(\theta, t),
$$

where $\mathcal{M}_2(\theta, t) = \mathbb{E}\big[e^{\theta Q_t(2)}\big]$. To solve for the steady-state moment generating function we consider the ODE given by

$$
\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{M}_2(\theta, \infty) = \frac{\eta^*\left(1-e^{2\theta}\right)\mathcal{M}_2(\theta, \infty)}{\frac{\alpha}{2}\left(e^{2\theta}-1\right) + \beta\left(e^{-\theta}-1\right)},
$$

with the initial condition that $\mathcal{M}_2(0, \infty) = 1$. By taking the derivative of the expression in (23), we verify the result. □

Appendix B. Exploring a hybrid self-exciting model

In the main body of the text we have defined the ESEP, a model that features arrivals that self-excite but only for a finite period of time. In contrast to the traditional Hawkes process models for self-excitement, in the ESEP the effect from one arrival does not decay through time but is fixed at a constant value for as long as it remains active. Thus, the ESEP features ephemeral but piecewise constant self-excitement, whereas the Hawkes process has eternal but ever decreasing self-excitement. One can note, though, that ephemeral self-excitement need not be piecewise constant. A model could feature both decay and down-jumps as ways of regulating its self-excitement. In this section of the appendix, we will consider such a model, specifically a Markovian one. To begin, let us now define the *hybrid ephemerally self-exciting process* (HESEP).

Definition 3. *(Hybrid ephemerally self-exciting process.)* Let $t \geq 0$ and suppose that $\nu^* > 0$, $\alpha > 0$, $\beta \geq 0$, and $\mu \geq 0$. Define ν_t, $N_{t,\nu}$, and $Q_{t,\nu}$ such that the following hold:

i. $N_{t,\nu}$ is an arrival process driven by the intensity ν_t;

ii. $Q_{t,\nu}$ is the number of arrivals from $N_{t,\nu}$ that have not yet expired according to their i.i.d. Exp(μ) activity durations;

iii. ν_t is governed by

$$
\mathrm{d}\nu_t = \beta(\nu^* - \nu_t)\mathrm{d}t + \alpha\mathrm{d}N_{t,\nu} - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\mathrm{d}D_{t,\nu},
$$

where $D_{t,\nu} = N_{t,\nu} - Q_{t,\nu}$.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

We say that the intensity–queue–counting-process triplet $(\nu_t, Q_{t,\nu}, N_{t,\nu})$ is a *hybrid ephemerally self-exciting process* (HESEP) with baseline intensity ν^*, intensity jump size α, decay rate β, expiration rate μ, and initial values $(\nu_0, Q_0^\nu, N_0^\nu)$.

By definition, one can view the HESEP as a hybrid between the ESEP and Hawkes process models. If $\beta = 0$ then we recover the ESEP; if $\mu = 0$ then we recover the Hawkes process. Hence the dynamics are largely quite familiar: up-jumps of size α at each arrival, exponential decay between events at rate β, and down-jumps upon each activity duration expiration. Perhaps the least intuitive part of this definition is the size of the down-jump, as this depends on the current levels of the intensity and the number of active exciters in the system. This draws inspiration from Markovian infinite-server queues. In an M/M/∞ queue, all jobs currently in the system are equally likely to be the next to leave, regardless of the order in which they entered the system. Similarly, in the HESEP, each exciter in the system is equally likely to be the next to leave. Moreover, the down-jump size is the same regardless of which exciter is the next to leave. When the expiration of an activity duration means that there are no longer any presently active exciters, by definition the intensity will return to the baseline value. One can note that this down-jump size is actually always bounded below by 0, since the intensity decays down towards the baseline ν^*, and is bounded above by α, since $\nu_t - \nu^* \le \alpha Q_{t,\nu}$ because each arrival increases ν_t by α before it decays. As a quick interesting fact regarding this process, in Proposition 13 we show that the size of a down-jump, $(\nu_t - \nu^*)/Q_{t,\nu}$, does not itself have down-jumps.

Proposition 13. *Let $\phi_t = \frac{\nu_t - \nu^*}{Q_{t,\nu}}$ be the size of a down-jump occurring at time $t \ge 0$. Suppose that $b \ge a \ge 0$ is such that $Q_{t,\nu}$ is positive for all $t \in [a, b]$. Then ϕ_t has no downward jumps on $[a, b]$.*

Proof. Suppose that $[a, b]$ is such an interval; then for $t \in [a, b]$ we note that

$$\frac{\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}} - \nu^*}{Q_{t,\nu} - 1} = \frac{Q_t \nu_t - \nu_t + \nu^* - Q_{t,\nu}\nu^*}{Q_{t,\nu}(Q_{t,\nu} - 1)} = \frac{\nu_t - \nu^*}{Q_{t,\nu}},$$

and this is equal to ϕ_t. □

B.1. Sandwiching the hybrid model

As one might expect for a so-called ‘hybrid’ model, the HESEP can be connected to the ESEP and the Hawkes process in many different ways. In Proposition 14, we show that one can actually sandwich this model between its two extremes. That is, we show that the means of the processes are equal when given the same parameters, whereas the variances are ordered with Hawkes as the smallest and ESEP as the largest.

Proposition 14. *Let $\alpha > 0$, $\beta > 0$, and $\mu > 0$ be such that $\mu + \beta > \alpha > 0$. Additionally, let $\nu^* > 0$. Let ν_t be an HESEP with baseline intensity ν^*, intensity jump size α, decay rate β, and service rate μ. Similarly, let λ_t be the intensity of a Hawkes process with baseline intensity ν^*, intensity jump α, and decay rate $\mu + \beta$. Finally, let η_t be the intensity of an ESEP with baseline intensity ν^*, intensity jump α, and service rate $\mu + \beta$. Then the means of these process intensities are all equal:*

$$\mathbb{E}[\lambda_t] = \mathbb{E}[\nu_t] = \mathbb{E}[\eta_t]. \tag{24}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Furthermore, the process variances are ordered as follows:

$$\text{Var}\,(\lambda_t) \le \text{Var}\,(\nu_t) \le \text{Var}\,(\eta_t). \tag{25}$$

Additionally, let $N_{t,\nu}$, $N_{t,\lambda}$, and $N_{t,\eta}$ be the counting processes of the HESEP, Hawkes process, and ESEP, respectively. Then the means of these counting processes are equal, i.e.

$$\mathbb{E}\big[N_{t,\lambda}\big] = \mathbb{E}\big[N_{t,\lambda}\big] = \mathbb{E}\big[N_{t,\eta}\big], \tag{26}$$

and the variances of the counting processes are again ordered as follows:

$$\text{Var}\,\big(N_{t,\lambda}\big) \le \text{Var}\,\big(N_{t,\nu}\big) \le \text{Var}\,\big(N_{t,\eta}\big). \tag{27}$$

Finally, the covariances between the intensity and counting process pairs are likewise ordered as follows:

$$\text{Cov}[\lambda_t, N_{t,\lambda}] \le \text{Cov}[\nu_t, N_{t,\nu}] \le \text{Cov}[\eta_t, N_{t,\eta}], \tag{28}$$

where $t \ge 0$ and where all intensities have the same initial value.

Proof. By a quick check of the differential equations for each mean, we can directly observe that $\mathbb{E}[\nu_t] = \mathbb{E}[\eta_t] = \mathbb{E}[\lambda_t]$. To show the variance ordering we begin by considering the ODE for the second moment of ν_t:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[\nu_t^2\Big] &= 2\beta\left(\nu^*\mathbb{E}[\nu_t] - \mathbb{E}\Big[\nu_t^2\Big]\right) + \alpha^2\mathbb{E}[\nu_t] + 2\alpha\mathbb{E}\Big[\nu_t^2\Big] \\ &\quad + \mu\mathbb{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right) Q_{t,\nu}\right].\end{aligned}$$

Now, let us observe that

$$\mathbb{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right) Q_{t,\nu}\right] = 2\left(\nu^*\mathbb{E}[\nu_t] - \mathbb{E}\Big[\nu_t^2\Big]\right) + \mathbb{E}\left[\frac{(\nu_t - \nu^*)^2}{Q_{t,\nu}}\right],$$

which follows from expanding the squared term. Because $\frac{(\nu_t - \nu^*)^2}{Q_{t,\nu}} \ge 0$, this gives us that

$$\mathbb{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right) Q_{t,\nu}\right] \ge 2\left(\nu^*\mathbb{E}[\nu_t] - \mathbb{E}\Big[\nu_t^2\Big]\right). \tag{29}$$

This inequality now allows us to directly compare $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\nu_t^2\big]$ to $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\lambda_t^2\big]$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\eta_t^2\big]$. First, we can use (29) to see that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[\nu_t^2\Big] &= 2\beta\left(\nu^*\mathbb{E}[\nu_t] - \mathbb{E}\Big[\nu_t^2\Big]\right) + \alpha^2\mathbb{E}[\nu_t] + 2\alpha\mathbb{E}\Big[\nu_t^2\Big] \\ &\quad + \mu\mathbb{E}\left[\left(\left(\nu_t - \frac{\nu_t - \nu^*}{Q_{t,\nu}}\right)^2 - \nu_t^2\right) Q_{t,\nu}\right] \\ &\ge 2(\mu + \beta)\left(\nu^*\mathbb{E}[\nu_t] - \mathbb{E}\Big[\nu_t^2\Big]\right) + \alpha^2\mathbb{E}[\nu_t] + 2\alpha\mathbb{E}\Big[\nu_t^2\Big].\end{aligned}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Because we have already shown that $\mathbb{E}[\lambda_t] = \mathbb{E}[\nu_t]$, we see that $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\lambda_t^2\big] \le \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\nu_t^2\big]$ when evaluated at the same point, and thus by Lemma 3, $\mathrm{Var}\,(\lambda_t) \le \mathrm{Var}\,(\nu_t)$. By analogous arguments for η_t, we achieve the stated result. For the counting process means, we can now observe that all the differential equations are such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[N_{t,\lambda}\big] = \mathbb{E}[\lambda_t] = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[N_{t,\nu}\big] = \mathbb{E}[\nu_t] = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[N_{t,\eta}\big] = \mathbb{E}[\eta_t].$$

We assume that all counting processes start at 0, and thus we have that the counting process means are equal throughout time. This also implies that the products of means, $\mathbb{E}[\lambda_t]\mathbb{E}\big[N_{t,\lambda}\big]$, $\mathbb{E}[\nu_t]\mathbb{E}\big[N_{t,\nu}\big]$, and $\mathbb{E}[\eta_t]\mathbb{E}\big[N_{t,\eta}\big]$, are equal. Hence, to show the ordering of the covariances, we will focus solely on the expectations of the products. This differential equation is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big[\nu_t N_{t,\nu}\big] = -(\mu+\beta-\alpha)\mathbb{E}\big[\nu_t N_{t,\nu}\big] + (\mu+\beta)\nu^*\mathbb{E}\big[N_{t,\nu}\big] + \alpha\mathbb{E}[\nu_t] + \mathbb{E}\Big[\nu_t^2\Big],$$

and we can note that the coefficients are the same for each of the processes. Not including the function for which we want to solve, $\mathbb{E}\big[\nu_t N_{t,\nu}\big]$, we can also observe that every function is equivalent across the processes, apart from the second moments of the intensities. We have shown that these second moments are in fact ordered, and therefore we arrive at the stated ordering of the covariances. Finally, we observe that the differential equation for the second moment of each counting process is of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[(N_{t,\nu})^2\Big] = \mathbb{E}[\nu_t] + 2\mathbb{E}\big[\nu_t N_{t,\nu}\big].$$

From the ordering of the covariances and the equivalences of the means, we can conclude the proof. □

As a simple consequence of Proposition 14, we can note that because the Hawkes process is over-dispersed, i.e. its variance is larger than its mean, so too are the ESEP and HESEP. These bounds on the variance of the HESEP are not only useful for comparison but also practical for the study of the HESEP itself, as the system of differential equations for the variance obtained via the infinitesimal generator is not closed.

B.2. Strong convergence of the HESEP counting process

In the final three subsections on the analysis of the HESEP, we obtain a trio of limiting results. We begin with the almost sure convergence of the ratio of the HESEP counting process and time, which is an elementary renewal result in the style of [9] or [45], for example. However, in contrast to the context of such works, we can bound the mean and variance of the HESEP via the ESEP, and we are instead interested solely in establishing the convergence, as we will obtain additional results by consequence. Using these expressions for the first two moments, the proof of Theorem 3 follows standard approaches using the Borel–Cantelli lemma. In Corollary 1 we use this renewal result to find a strong law of large numbers for the dependent and non-identically distributed inter-arrival times of the HESEP by way of the continuous mapping theorem, which is another standard technique.

Theorem 3. *Let $(\nu_t, Q_{t,\nu}, N_{t,\nu})$ be an HESEP with baseline intensity ν^*, intensity jump $\alpha > 0$, intensity decay rate $\beta \ge 0$, and rate of exponentially distributed service $\mu \ge 0$, where $\mu + \beta > \alpha$. Then*

$$\frac{N_{t,\nu}}{t} \xrightarrow{\text{a.s.}} \nu_\infty \tag{30}$$

as $t \to \infty$, where $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Proof. We will show this through use of the Borel–Cantelli lemma. Let $\epsilon > 0$ be arbitrary, and define the event E_s for $s \in \mathbb{N}$ as

$$E_s = \left\{ \sup_{t \in (s^2, (s+1)^2]} \frac{\left|N_{t,\nu} - \mathbb{E}\left[N_{t,\nu}\right]\right|}{t} > \epsilon \right\}.$$

We now note that $N_{t,\nu} - \mathbb{E}\left[N_{t,\nu}\right]$ is a martingale by definition, and so $|N_{t,\nu} - \mathbb{E}\left[N_{t,\nu}\right]|$ is a sub-martingale. Additionally, we can observe that

$$\mathbb{P}\left(E_s\right) \leq \mathbb{P}\left(\sup_{t \in (s^2, (s+1)^2]} \left|N_{t,\nu} - \mathbb{E}\left[N_{t,\nu}\right]\right| > s^2 \epsilon \right),$$

because $s^2 \leq t$ for any t. By Doob's martingale inequality, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{t \in (s^2, (s+1)^2]} \left|N_{t,\nu} - \mathbb{E}\left[N_{t,\nu}\right]\right| > s^2 \epsilon \right) &\leq \frac{\mathbb{E}\left[\left|N_{(s+1)^2,\nu} - \mathbb{E}\left[N_{(s+1)^2,\nu}\right]\right|^2\right]}{s^4 \epsilon^2} \\ &= \frac{\text{Var}\left(N_{(s+1)^2,\nu}\right)}{s^4 \epsilon^2}. \end{aligned}$$

From Proposition 14, we note that the variance of an HESEP counting process with baseline intensity ν^*, intensity jump size α, decay rate β, and service rate μ is bounded above by the variance of an ESEP counting process with baseline ν^*, jump size α, and service rate $\mu + \beta$. Using the explicit form of the ESEP counting process variance as computed through Lemma 2, we have the bound

$$\begin{aligned} &\text{Var}\left(N_{(s+1)^2,\nu}\right) \leq \text{Var}\left(N_{(s+1)^2,\eta}\right) \\ &= \frac{((\mu+\beta)^2 + \alpha^2)\nu_\infty}{(\mu+\beta-\alpha)^2}(s+1)^2 \\ &\quad - \frac{2\alpha\mu(\nu_0 - \nu_\infty)}{(\mu+\beta-\alpha)^3}\left(e^{-(\mu+\beta-\alpha)(s+1)^2} + (\mu+\beta-\alpha)(s+1)^2 e^{-(\mu+\beta-\alpha)(s+1)^2}\right) \\ &\quad + \left(\frac{\nu_0 - \nu_\infty}{\mu+\beta-\alpha} - \frac{\alpha\mu\nu_\infty}{(\mu+\beta-\alpha)^3} - \frac{(\alpha^2 + \alpha(\mu+\beta))\nu_0}{(\mu+\beta-\alpha)^3}\right)\left(1 - e^{-(\mu+\beta-\alpha)(s+1)^2}\right) \\ &\quad + \left(\frac{(\alpha^2 + \alpha(\mu+\beta))\nu_0}{2(\mu+\beta-\alpha)^3} - \frac{\alpha(\mu+\beta)\nu_\infty}{2(\mu+\beta-\alpha)^3}\right)\left(1 - e^{-2(\mu+\beta-\alpha)(s+1)^2}\right). \end{aligned}$$

Together, this implies that $\mathbb{P}\left(E_s\right) \in O\left(\frac{1}{s^2}\right)$. Therefore $\sum_{s=0}^{\infty} \mathbb{P}\left(E_s\right) < \infty$, and so by the Borel–Cantelli lemma,

$$\frac{\left|N_{t,\nu} - \mathbb{E}\left[N_{t,\nu}\right]\right|}{t} \xrightarrow{\text{a.s.}} 0.$$

Since $\lim_{t \to \infty} \frac{\mathbb{E}[N_{t,\nu}]}{t} = \nu_\infty$, this completes the proof. □

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

As an immediate consequence of this, we achieve a law of large numbers for the dependent inter-arrival times.

Corollary 1. *Let $(\nu_t, Q_{t,\nu}, N_{t,\nu})$ be an HESEP counting process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, intensity decay rate $\beta \geq 0$, and rate of exponentially distributed service $\mu \geq 0$, where $\mu + \beta > \alpha$. Further, let S_k^ν denote the kth inter-arrival time for $k \in \mathbb{Z}^+$. Then*

$$\frac{1}{n}\sum_{k=1}^{n} S_k^\nu \xrightarrow{\text{a.s.}} \frac{1}{\nu_\infty} \tag{31}$$

as $n \to \infty$, where $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$.

Proof. Let A_n^ν denote the time of the nth arrival for each $n \in \mathbb{Z}^+$, which is to say that $A_n^\nu = \sum_{k=1}^n S_k^\nu$. Now, observe that the time of the most recent arrival up to time t, $A_{N_{t,\nu}}^\nu$, can be bounded as

$$t - S_{N_{t,\nu}+1}^\nu \leq A_{N_{t,\nu}}^\nu \leq t,$$

since if

$$t - S_{N_{t,\nu}+1}^\nu > A_{N_{t,\nu}}^\nu,$$

then arrival $N_{t,\nu} + 1$ would have occurred before time t. We also note that because $\nu^* > 0$, we have $N_{t,\nu} \to \infty$ as $t \to \infty$, and this implies that

$$\frac{S_{N_{t,\nu}+1}}{N_{t,\nu}} \xrightarrow{\text{a.s.}} 0$$

as $t \to \infty$. From Proposition 3 and the continuous mapping theorem, we know that

$$\frac{t}{N_{t,\nu}} \to \frac{1}{\nu_\infty}$$

and

$$\frac{t - S_{N_{t,\nu}+1}}{N_{t,\nu}} \to \frac{1}{\nu_\infty}$$

almost surely. By the sandwiching of $A_{N_{t,\nu}}$, this yields the stated result. □

Because the Hawkes process and the ESEP are special cases of this hybrid model, we can note that both the renewal result and the law of large numbers apply directly to each.

Corollary 2. *Let $(\lambda_t, N_{t,\lambda})$ be the intensity and count of a Hawkes process with baseline intensity $\lambda^* > 0$, intensity jump $\alpha > 0$, and decay rate $\beta > \alpha$. Similarly, let $(\eta_t, N_{t,\eta})$ be the intensity and counting process pair for an ESEP with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and rate of exponentially distributed service $\mu > \alpha$. Then, for S_k^λ and S_k^η as the kth inter-arrival times for the Hawkes process and the ESEP process, respectively, we have that*

$$\frac{N_{t,\lambda}}{t} \xrightarrow{\text{a.s.}} \lambda_\infty, \qquad \frac{N_{t,\eta}}{t} \xrightarrow{\text{a.s.}} \eta_\infty, \tag{32}$$

and

$$\frac{1}{n}\sum_{k=1}^{n} S_k^\lambda \xrightarrow{\text{a.s.}} \frac{1}{\lambda_\infty}, \qquad \frac{1}{n}\sum_{k=1}^{n} S_k^\eta \xrightarrow{\text{a.s.}} \frac{1}{\eta_\infty}, \tag{33}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$ and $\eta_\infty = \frac{\mu\nu^*}{\mu-\alpha}$.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

B.3. Baseline fluid limit of the HESEP

In this subsection and in the sequel, we consider a baseline scaling of the HESEP. That is, we investigate limiting properties of the process as the baseline intensity grows large and the intensity and queue length are normalized in some fashion. To begin, we take the normalization as directly proportional to the baseline scaling, which is the fluid limit. The derivation of this is empowered by the following lemma, which allows us to make use of Taylor expansions.

Lemma 4. *Suppose that for some $b > 0$, $-b \leq z_n(t) \leq 0$ for all values of n. Then there exist constants C_1 and C_2, where $C_1 \leq C_2$, which imply the following bounds for sufficiently large values of n:*

$$z_n(t) + \frac{C_1}{n} \leq n \cdot \left(e^{\frac{z_n(t)}{n}} - 1 \right) \leq z_n(t) + \frac{C_2}{n}. \tag{34}$$

Proof. The proof follows from performing a second-order Taylor expansion for the exponential function and observing that since $z_n(t)$ lies in a compact interval, we can construct uniform lower and upper bounds for the exponential function. □

With this lemma in hand, we now proceed to finding the fluid limit in Theorem 4. In this case, we scale the baseline intensity by n, whereas we scale the intensity and the queue length by $\frac{1}{n}$. As one would expect, we find that the fluid limit converges to the means of the intensity and queue.

Theorem 4. *For $n \in \mathbb{Z}$, let the nth fluid-scaled HESEP $(\nu_t(n), Q_{t,\nu}(n))$ be defined so that the baseline intensity is $n\nu^*$, the intensity jump size is $\alpha > 0$, the intensity decay rate is $\beta \geq 0$, and the rate of exponentially distributed service is $\mu > 0$, where $\mu + \beta > \alpha$. Then, for the scaled quantities*

$$\left(\frac{\nu_{t,\nu}(n)}{n}, \frac{Q_{t,\nu}(n)}{n} \right),$$

the limit of the moment generating function,

$$\tilde{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q) \equiv \lim_{n\to\infty} \mathbb{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n) + \frac{\theta_Q}{n} Q_{t,\nu}(n)} \right], \tag{35}$$

is given by

$$\tilde{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q) = e^{\theta_\nu \mathbb{E}[\nu_t] + \theta_Q \mathbb{E}[Q_{t,\nu}]} \tag{36}$$

for all $t \geq 0$.

Proof. The proof has two steps. The first step is to show that the limiting moment generating function converges to a PDE given by $\tilde{\mathcal{M}}^\infty$ using properties of the exponential function and Lemma 4. The second step is to solve this PDE using the method of characteristics. Finally, by the uniqueness of moment generating functions, we can assert that the random variables to which our limit converges are deterministic functions of time, which are also known as the

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

fluid limit. We begin with the infinitesimal generator form, which is simplified through the linearity of expectation as

$$\begin{aligned}
\frac{\partial}{\partial t}\tilde{\mathcal{M}}^n\left(t,\theta_\nu,\theta_Q\right) &\equiv \frac{\partial}{\partial t}\mathbb{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&= \mathbb{E}\left[\beta(\nu^* n-\nu_t(n))\frac{\theta_\nu}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\quad+\mathbb{E}\left[\nu_t(n)\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\quad+\mathbb{E}\left[\mu Q_{t,\nu}(n)\left(e^{-\frac{\theta_\nu(\nu_t(n)-\nu^* n)}{nQ_{t,\nu}(n)}-\frac{\theta_Q}{n}}-1\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&= \beta\nu^*\theta_\nu\mathbb{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_t(n)}\right]-\beta\theta_\nu\mathbb{E}\left[\frac{\nu_t(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\quad+n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)\mathbb{E}\left[\frac{\nu_t(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_t(n)}\right]\\
&\quad+\frac{\mu}{n}\mathbb{E}\left[Q_{t,\nu}(n)n\left(e^{-\frac{\theta_\nu(\nu_t(n)-\nu^* n)}{nQ_{t,\nu}(n)}-\frac{\theta_Q}{n}}-1\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&= \beta\nu^*\theta_\nu\tilde{\mathcal{M}}(t,\theta_\nu,\theta_Q)+\left(n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)-\beta\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}(t,\theta_\nu,\theta_Q)\\
&\quad+\frac{\mu}{n}\mathbb{E}\left[Q_{t,\nu}(n)\left(-\frac{\theta_\nu(\nu_t(n)-\nu^* n)}{Q_{t,\nu}(n)}-\theta_Q+\frac{\epsilon_n}{n}\right)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right],
\end{aligned}$$

where the last equality holds for sufficiently large n, with $\epsilon_n$in some bounded interval as according to Lemma 4. Then, by rearranging further, we can see that in the limit this becomes

$$\begin{aligned}
&\frac{\partial}{\partial t}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)\\
&= \beta\nu^*\theta_\nu M^n(t,\theta_\nu,\theta_Q)+\left(n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)-\beta\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)\\
&\quad-\mu\theta_\nu\mathbb{E}\left[\frac{\nu_t(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]+\mu\theta_\nu\nu^*\mathbb{E}\left[e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\quad-\mu\theta_Q\mathbb{E}\left[\frac{Q_{t,\nu}(n)}{n}e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]+\frac{\mu\epsilon_n}{n}\mathbb{E}\left[Q_{t,\nu}(n)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&= (\mu+\beta)\nu^*\theta_\nu\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)+\left(n\left(e^{\frac{\alpha\theta_\nu}{n}+\frac{\theta_Q}{n}}-1\right)-(\mu+\beta)\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)\\
&\quad-\mu\theta_Q\frac{\partial}{\partial\theta_Q}\tilde{\mathcal{M}}^n(t,\theta_\nu,\theta_Q)+\frac{\mu\epsilon_n}{n^2}\mathbb{E}\left[Q_{t,\nu}(n)e^{\frac{\theta_\nu}{n}\nu_t(n)+\frac{\theta_Q}{n}Q_{t,\nu}(n)}\right]\\
&\overset{n\to\infty}{\longrightarrow}(\mu+\beta)\nu^*\theta_\nu\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q)+\left(\theta_Q-(\mu+\beta-\alpha)\theta_\nu\right)\frac{\partial}{\partial\theta_\nu}\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q)\\
&\quad-\mu\theta_Q\frac{\partial}{\partial\theta_Q}\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q).
\end{aligned}$$

We now solve this PDE for $\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q)$ through the method of characteristics. For simplicity's sake we will instead use this procedure to solve for $G(t,\theta_\nu,\theta_Q)=\log\left(\tilde{\mathcal{M}}^\infty(t,\theta_\nu,\theta_Q)\right)$.

This PDE is given by

$$(\mu+\beta)\nu^*\theta_\nu = \frac{\partial}{\partial t}G(t,\theta_\nu,\theta_Q) + \mu\theta_Q\frac{\partial}{\partial\theta_Q}G(t,\theta_\nu,\theta_Q) + \big((\mu+\beta-\alpha)\theta_\nu - \theta_Q\big)\frac{\partial}{\partial\theta_\nu}G(t,\theta_\nu,\theta_Q),$$

with boundary condition $G(0,\theta_\nu,\theta_Q) = \theta_Q Q_0 + \theta_\nu \nu_0$. This corresponds to the following system of characteristic equations:

$$\begin{aligned}
&\frac{\mathrm{d}\theta_Q}{\mathrm{d}z}(x,y,z) = \mu\theta_Q, && \theta_Q(x,y,0) = x,\\
&\frac{\mathrm{d}\theta_\nu}{\mathrm{d}z}(x,y,z) = (\mu+\beta-\alpha)\theta_\nu - \theta_Q, && \theta_\nu(x,y,0) = y,\\
&\frac{\mathrm{d}t}{\mathrm{d}z}(x,y,z) = 1, && t(x,y,0) = 0,\\
&\frac{\mathrm{d}g}{\mathrm{d}z}(x,y,z) = (\mu+\beta)\nu^*\theta_\nu = (\mu+\beta-\alpha)\nu_\infty\theta_\nu, && g(x,y,0) = xQ_0 + y\nu_0.
\end{aligned}$$

If $\beta \neq \alpha$, the solutions to these initial value problems are given by the following:

$$\begin{aligned}
\theta_Q(x,y,z) &= xe^{\mu z},\\
\theta_\nu(x,y,z) &= ye^{(\mu+\beta-\alpha)z} + \frac{x}{\beta-\alpha}\left(e^{\mu z} - e^{(\mu+\beta-\alpha)z}\right)\\
&= \left(y - \frac{x}{\beta-\alpha}\right)e^{(\mu+\beta-\alpha)z} + \frac{xe^{\mu z}}{\beta-\alpha},\\
t(x,y,z) &= z,\\
g(x,y,z) &= xQ_0 + y\nu_0 + \nu_\infty\left(y - \frac{x}{\beta-\alpha}\right)\left(e^{(\mu+\beta-\alpha)z} - 1\right) + \frac{x\nu_\infty(\mu+\beta-\alpha)(e^{\mu z}-1)}{\mu(\beta-\alpha)}.
\end{aligned}$$

Now we can solve for the characteristic variables in terms of the original variables and find $x = \theta_Q e^{-\mu t}$,

$$y = \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right),$$

and $z = t$, so this gives a PDE solution of

$$\begin{aligned}
G(t,\theta_Q,\theta_\nu) &= g\left(\theta_Q e^{-\mu t}, \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right), t\right)\\
&= Q_0\theta_Q e^{-\mu t} + \nu_0\left(\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right)\right)\\
&\quad + \nu_\infty\left(\theta_\nu - \frac{\theta_Q}{\beta-\alpha}\right)\left(1 - e^{-(\mu+\beta-\alpha)t}\right) + \frac{\theta_Q\nu_\infty(\mu+\beta-\alpha)(1-e^{-\mu t})}{\mu(\beta-\alpha)}.
\end{aligned}$$

If instead $\beta = \alpha$, the solutions to the characteristic ODEs are as follows:

$$\begin{aligned}
\theta_Q(x,y,z) &= xe^{\mu z},\\
\theta_\nu(x,y,z) &= e^{\mu z}(y - xz),\\
t(x,y,z) &= z,\\
g(x,y,z) &= xQ_0 + y\nu_0 + \nu_\infty y\left(e^{\mu z} - 1\right) - \frac{x\nu_\infty}{\mu}\left(e^{\mu z}(\mu z - 1) + 1\right).
\end{aligned}$$

This makes our expressions for the characteristic variables $x = \theta_Q e^{-\mu t}$, $y = \theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t}$, and $z = t$. This leads to the PDE solution

$$\begin{aligned} G(t, \theta_Q, \theta_\nu) &= g\left(\theta_Q e^{-\mu t}, \theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t}, t\right) \\ &= Q_0 \theta_Q e^{-\mu t} + \nu_0 \theta_\nu e^{-\mu t} + \nu_0 \theta_Q t e^{-\mu t} + \nu_\infty \left(\theta_\nu + \theta_Q t\right)\left(1 - e^{-\mu t}\right) \\ &\quad - \frac{\nu_\infty \theta_Q}{\mu}\left(\mu t - 1 + e^{-\mu t}\right), \end{aligned}$$

and we can observe that each of these cases simplifies to the corresponding means of the queue and the intensity, which yields the stated result. □

B.4. Baseline diffusion limit of the HESEP

Now, to consider a diffusion limit, we still scale the baseline intensity by n, but we scale the process intensity and the queue length by $\frac{1}{\sqrt{n}}$ instead. More specifically, we scale the centered versions of the processes by $\frac{1}{\sqrt{n}}$. While we can make use of some of the techniques used for the fluid limit in Theorem 4, the diffusion scaling also involves second-order terms. It is challenging to calculate such quantities for the HESEP. Thus, we will use the same idea to bound the quantities above and below via

$$0 \le \frac{(\nu_t - \nu^*)^2}{Q_{t,\nu}} \le \alpha(\nu_t - \nu^*), \tag{37}$$

which follows from our previously discussed bounds on the down-jump size. By doing so, we create upper and lower bounds for the true diffusion limit of the HESEP. To facilitate a variety of approximations that fit within these bounds, we introduce the parameter $\gamma \in [0, 1]$, with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ to the upper.

Theorem 5. *For $n \in \mathbb{Z}$, let the nth diffusion-scaled HESEP $(\nu_t(n), Q_{t,\nu}(n))$ be defined so that the baseline intensity is $n\nu^*$, the intensity jump size is $\alpha > 0$, the intensity decay rate is $\beta \ge 0$, and the rate of exponentially distributed service is $\mu > 0$, where $\mu + \beta > \alpha$. For the scaled quantities $\left(\frac{\nu_t(n)}{\sqrt{n}}, \frac{Q_{t,\nu}(n)}{\sqrt{n}}\right)$, let $\hat{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q)$ be defined by*

$$\hat{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q) \equiv \lim_{n\to\infty} \mathbb{E}\left[e^{\frac{\theta_\nu}{\sqrt{n}}(\nu_t(n) - n\nu_\infty) + \frac{\theta_Q}{\sqrt{n}}\left(Q_{t,\nu}(n) - \frac{n\nu_\infty}{\mu}\right)}\right], \tag{38}$$

if the limit converges. Then for $\beta \neq \alpha$, this is bounded above and below by $\mathcal{B}_0(t, \theta_\nu, \theta_Q) \le \hat{\mathcal{M}}^\infty(t, \theta_\nu, \theta_Q) \le \mathcal{B}_1(t, \theta_\nu, \theta_Q)$, where $\mathcal{B}_\gamma(t, \theta_\nu, \theta_Q)$ is given by

$$\begin{aligned} \mathcal{B}_\gamma(t, \theta_\nu, \theta_Q) = {} & e^{\nu_0 \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\nu_0 \theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right) + Q_0 \theta_Q e^{-\mu t}} \\ & \cdot e^{\left(\theta_\nu - \frac{\theta_Q}{\beta-\alpha}\right)^2 \left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2} + \frac{\alpha^2 \nu_\infty}{2}\right) \frac{1 - e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}} \\ & \cdot e^{\left(\theta_\nu \theta_Q - \frac{\theta_Q^2}{\beta-\alpha}\right)\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha} + \mu\right)(\nu_\infty - \nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right) \frac{1 - e^{-(2\mu+\beta-\alpha)t}}{2\mu+\beta-\alpha}} \\ & \cdot e^{\theta_Q^2 \left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty - \nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty \beta^2}{2(\beta-\alpha)^2}\right) \frac{1 - e^{-2\mu t}}{2\mu}}, \end{aligned} \tag{39}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

whereas if $\beta = \alpha$, it is instead

$$
\begin{aligned}
\mathcal{B}_\gamma(t, \theta_\nu, \theta_Q) = {} & e^{\nu_0\theta_\nu e^{-\mu t}+\nu_0\theta_Q t e^{-\mu t}+Q_0\theta_Q e^{-\mu t}+\Bigg(\bigg(\frac{\gamma\alpha(\theta_\nu+\theta_Q t)^2}{2}+\theta_\nu\theta_Q+\theta_Q^2 t\bigg)\frac{1-e^{-2\mu t}}{2}} \\
& \cdot e^{-\Big(\gamma\alpha(\theta_\nu\theta_Q+\theta_Q^2 t)+\theta_Q^2\Big)\frac{2\mu t-1+e^{-2\mu t}}{4\mu}+\frac{\gamma\alpha\theta_Q^2}{2}\bigg(\frac{2\mu t(\mu t-1)+1-e^{-2\mu t}}{4\mu^2}\bigg)\Bigg)(\nu_\infty-\nu^*)} \\
& \cdot e^{\frac{\nu_\infty}{2}\Bigg(\Big(\theta_Q^2+(\theta_Q+\alpha\theta_\nu)^2+2\Big(\alpha^2\theta_\nu\theta_Q+\alpha\theta_Q^2\Big)t+\alpha^2\theta_Q^2 t^2\Big)\frac{1-e^{-2\mu t}}{2\mu}} \\
& \cdot e^{-2\Big(\alpha^2\theta_\nu\theta_Q+\alpha\theta_Q^2+\alpha^2\theta_Q^2 t\Big)\bigg(\frac{2\mu t-1+e^{-2\mu t}}{4\mu^2}\bigg)+\alpha^2\theta_Q^2\bigg(\frac{2\mu t(\mu t-1)+1-e^{-2\mu t}}{4\mu^3}\bigg)\Bigg)}, \qquad (40)
\end{aligned}
$$

for $\gamma \in [0, 1]$ *with* $t \geq 0$ *and* $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$.

Proof. We begin by bounding the quantity $Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2$ above and below by observing

$$
0 \leq Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2 = (\nu_t(n)-n\nu^*)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right) \leq \alpha(\nu_t(n)-n\nu^*).
$$

To consolidate the development of the two bounds into one approach, we introduce the extra parameter $\gamma \in \{0, 1\}$ and replace

$$
Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2
$$

by $\gamma\alpha(\nu_t(n)-n\nu^*)$ in the following diffusion limit derivation. In this notation, $\gamma = 0$ yields the lower bound and $\gamma = 1$ the upper. These two cases share the same start—identifying the moment generating function form of the pre-limit object. By Lemma 2, this is

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu, \theta_Q, t) &= \frac{\partial}{\partial t}\mathbb{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
&= \mathbb{E}\left[\nu_t(n)\left(e^{\frac{\alpha\theta_\nu+\theta_Q}{\sqrt{n}}}-1\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
&\quad + \mathbb{E}\left[\mu Q_{t,\nu}(n)\left(e^{-\frac{\theta_\nu}{\sqrt{n}}\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)-\frac{\theta_Q}{\sqrt{n}}}-1\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right] \\
&\quad + \mathbb{E}\left[\frac{\beta\theta_\nu}{\sqrt{n}}\left(n\nu^*-\nu_t(n)\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right].
\end{aligned}
$$

As a first step, we simplify this expression through the linearity of expectation. Moving deterministic terms outside of the expectation and rescaling, we have

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) &= \sqrt{n}\left(e^{\frac{\alpha\theta_\nu+\theta_Q}{\sqrt{n}}}-1\right)\mathbb{E}\left[\frac{\nu_t(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\mathbb{E}\left[\mu Q_{t,\nu}(n)\left(e^{-\frac{\theta_\nu}{\sqrt{n}}\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)-\frac{\theta_Q}{\sqrt{n}}}-1\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\beta\theta_\nu\nu^*\sqrt{n}\mathbb{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&-\beta\theta_\nu\mathbb{E}\left[\frac{\nu_t(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right].
\end{aligned}
$$

For the terms on the first, third, and fourth lines in the right-hand side of the above equation, we are able to re-express the expectation in terms of the moment generating function or its derivatives. For the first and second lines, we perform Taylor expansions and truncate terms from the third order and above. This yields

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) &= \left(\alpha\theta_\nu+\theta_Q+\frac{(\alpha\theta_\nu+\theta_Q)^2}{2\sqrt{n}}+O\left(\frac{1}{n}\right)\right)\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right.\\
&\left.+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right)+\mathbb{E}\left[\mu Q_{t,\nu}(n)\left(-\frac{\theta_\nu}{\sqrt{n}}\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)-\frac{\theta_Q}{\sqrt{n}}\right.\right.\\
&\left.\left.+\frac{1}{2n}\left(\theta_\nu\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)+\theta_Q\right)^2+O\left(n^{-\frac{3}{2}}\right)\right)e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\beta\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)-\beta\theta_\nu\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right).
\end{aligned}
$$

We now begin distributing and combining like terms through linearity of expectation. Moreover, we distribute within the expectation on the second line and cancel $Q_{t,\nu}(n)$ across the numerator and denominator where possible, obtaining

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t) &= \left(\alpha\theta_\nu+\theta_Q+\frac{(\alpha\theta_\nu+\theta_Q)^2}{2\sqrt{n}}-\beta\theta_\nu+O\left(\frac{1}{n}\right)\right)\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right.\\
&\left.+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t)\right)-\mu\theta_\nu\mathbb{E}\left[\frac{\nu_t(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\mu\theta_\nu\nu^*\sqrt{n}\mathbb{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&-\mu\theta_Q\mathbb{E}\left[\frac{Q_{t,\nu}(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\frac{\mu\theta_\nu^2}{2n}\mathbb{E}\left[Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]
\end{aligned}
$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

$$
\begin{aligned}
&+\frac{\mu\theta_\nu\theta_Q}{\sqrt{n}}\mathbb{E}\left[\frac{\nu_t(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&-\mu\theta_\nu\theta_Q\nu^*\mathbb{E}\left[e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\frac{\mu\theta_Q^2}{2\sqrt{n}}\mathbb{E}\left[\frac{Q_{t,\nu}(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+O\left(\frac{1}{n}\right)\mathbb{E}\left[\frac{Q_{t,\nu}(n)}{\sqrt{n}}e^{\theta_\nu\left(\frac{\nu_t(n)-n\nu_\infty}{\sqrt{n}}\right)+\theta_Q\left(\frac{Q_{t,\nu}(n)-\frac{n\nu_\infty}{\mu}}{\sqrt{n}}\right)}\right]\\
&+\beta\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n(\theta_\nu,\theta_Q,t).
\end{aligned}
$$

For all remaining components of this equation that are still expressed in terms of the expectation, we substitute equivalent forms in terms of the moment generating function or its partial derivatives. Furthermore, we will now replace

$$
Q_{t,\nu}(n)\left(\frac{\nu_t(n)-n\nu^*}{Q_{t,\nu}(n)}\right)^2
$$

by $\gamma\alpha(\nu_t(n)-n\nu^*)$ inside the expectation and re-express the expectation in terms of the moment generating function accordingly. To denote that we have now made this replacement and changed the function, we add γ as a subscript to the moment generating function, i.e. $\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)$. We now have

$$
\begin{aligned}
\frac{\partial}{\partial t}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)=&\left(\alpha\theta_\nu+\theta_Q+\frac{(\alpha\theta_\nu+\theta_Q)^2}{2\sqrt{n}}-\beta\theta_\nu+O\left(\frac{1}{n}\right)\right)\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right.\\
&\left.+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right)-\mu\theta_\nu\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right)\\
&+\mu\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)-\mu\theta_Q\left(\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)+\frac{\nu_\infty\sqrt{n}}{\mu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right)\\
&+\frac{\gamma\alpha\mu\theta_\nu^2}{2\sqrt{n}}\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right)-\frac{\gamma\alpha\mu\nu^*\theta_\nu^2}{2}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\\
&+\frac{\mu\theta_\nu\theta_Q}{\sqrt{n}}\left(\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)+\nu_\infty\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right)-\mu\theta_\nu\theta_Q\nu^*\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\\
&+\frac{\mu\theta_Q^2}{2\sqrt{n}}\left(\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)+\frac{\nu_\infty\sqrt{n}}{\mu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right)+\beta\theta_\nu\nu^*\sqrt{n}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\\
&+O\left(\frac{1}{n}\right)\left(\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)+\frac{\nu_\infty\sqrt{n}}{\mu}\hat{\mathcal{M}}^n_\gamma(\theta_\nu,\theta_Q,t)\right).
\end{aligned}
$$

Before we find the limiting object, we first combine like terms of the moment generating function, consolidating coefficients and absorbing into $O(\cdot)$ notation where possible:

$$\frac{\partial}{\partial t}\hat{\mathcal{M}}_\gamma^n(\theta_\nu, \theta_Q, t) = \left(\theta_Q - (\mu+\beta-\alpha)\theta_\nu + O\left(\frac{1}{\sqrt{n}}\right)\right)\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}_\gamma^n(\theta_\nu, \theta_Q, t)$$
$$- \left(\mu\theta_Q - O\left(\frac{1}{\sqrt{n}}\right)\right)\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}_\gamma^n(\theta_\nu, \theta_Q, t)\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q(\nu_\infty-\nu^*)\right.$$
$$\left.+\frac{\theta_Q^2\nu_\infty}{2} + \frac{(\alpha\theta_\nu+\theta_Q)^2\nu_\infty}{2} + O\left(\frac{1}{\sqrt{n}}\right)\right)\hat{\mathcal{M}}_\gamma^n(\theta_\nu, \theta_Q, t).$$

Taking the limit as $n \to \infty$, we get

$$\frac{\partial}{\partial t}\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t) = \big(\theta_Q - (\mu+\beta-\alpha)\theta_\nu\big)\frac{\partial}{\partial\theta_\nu}\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t) - \mu\theta_Q\frac{\partial}{\partial\theta_Q}\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t)$$
$$+\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q(\nu_\infty-\nu^*) + \frac{\theta_Q^2\nu_\infty}{2} + \frac{(\alpha\theta_\nu+\theta_Q)^2\nu_\infty}{2}\right)\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t).$$

We will now solve this limiting PDE through the method of characteristics. To simplify this approach, we let $G_\gamma(\theta_\nu, \theta_Q, t) = \log\left(\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t)\right)$, which is the cumulant generating function. The resulting PDE for the cumulant generating function is then

$$\frac{\partial}{\partial t}G_\gamma(\theta_\nu, \theta_Q, t) + \big((\mu+\beta-\alpha)\theta_\nu - \theta_Q\big)\frac{\partial}{\partial\theta_\nu}G_\gamma(\theta_\nu, \theta_Q, t) + \mu\theta_Q\frac{\partial}{\partial\theta_Q}G_\gamma(\theta_\nu, \theta_Q, t)$$
$$= \frac{\gamma\alpha\mu(\nu_\infty-\nu^*)\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q(\nu_\infty-\nu^*) + \frac{\theta_Q^2\nu_\infty}{2} + \frac{(\alpha\theta_\nu+\theta_Q)^2\nu_\infty}{2},$$

with initial condition $G_\gamma(\theta_\nu, \theta_Q, 0) = \theta_\nu\nu_0 + \theta_Q Q_0$. Thus, the resulting system of characteristic equations is

$$\frac{\mathrm{d}t}{\mathrm{d}z}(x, y, z) = 1,$$
$$t(x, y, 0) = 0,$$
$$\frac{\mathrm{d}\theta_\nu}{\mathrm{d}z}(x, y, z) = (\mu+\beta-\alpha)\theta_\nu - \theta_Q,$$
$$\theta_\nu(x, y, 0) = x,$$
$$\frac{\mathrm{d}\theta_Q}{\mathrm{d}z}(x, y, z) = \mu\theta_Q,$$
$$\theta_Q(x, y, 0) = y,$$
$$\frac{\mathrm{d}g}{\mathrm{d}z}(x, y, z) = \left(\frac{\gamma\alpha\mu\theta_\nu^2}{2} + \mu\theta_\nu\theta_Q\right)(\nu_\infty-\nu^*) + \left(\theta_Q^2 + (\alpha\theta_\nu+\theta_Q)^2\right)\frac{\nu_\infty}{2},$$
$$g(x, y, 0) = x\nu_0 + yQ_0.$$

Assuming $\beta \neq \alpha$, we can solve these first three ODEs to find that

$$t = z, \qquad \theta_Q = ye^{\mu z}, \qquad \theta_\nu = \left(x - \frac{y}{\beta-\alpha}\right)e^{(\mu+\beta-\alpha)z} + \frac{y}{\beta-\alpha}e^{\mu z},$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

which we now use to solve the remaining equation. Rewriting the characteristic equation for g, we have

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}z}(x, y, z) &= \frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2}\left(\left(x - \frac{y}{\beta-\alpha}\right)^2 e^{2(\mu+\beta-\alpha)z} + \frac{y^2}{(\beta-\alpha)^2}e^{2\mu z}\right.\\ &\quad \left. + \frac{2}{\beta-\alpha}\left(xy - \frac{y^2}{\beta-\alpha}\right)e^{(2\mu+\beta-\alpha)z}\right) + \mu(\nu_\infty - \nu^*)\left(\left(xy - \frac{y^2}{\beta-\alpha}\right)e^{(2\mu+\beta-\alpha)z}\right.\\ &\quad \left. + \frac{y^2}{\beta-\alpha}e^{2\mu z}\right) + \frac{\nu_\infty}{2}\left(\alpha^2\left(x - \frac{y}{\beta-\alpha}\right)^2 e^{2(\mu+\beta-\alpha)z} + \left(1 + \frac{\beta^2}{(\beta-\alpha)^2}\right)y^2 e^{2\mu z}\right.\\ &\quad \left. + \frac{2\alpha\beta}{\beta-\alpha}\left(xy - \frac{y^2}{\beta-\alpha}\right)e^{(2\mu+\beta-\alpha)z}\right),\end{aligned}$$

and so by grouping coefficients of like exponential functions and then integrating with respect to z, we obtain the solution

$$\begin{aligned}g(x, y, z) &= x\nu_0 + yQ_0 + \left(x - \frac{y}{\beta-\alpha}\right)^2\left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{e^{2(\mu+\beta-\alpha)z} - 1}{2(\mu+\beta-\alpha)}\\ &\quad + \left(xy - \frac{y^2}{\beta-\alpha}\right)\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha} + \mu\right)(\nu_\infty - \nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{e^{(2\mu+\beta-\alpha)z} - 1}{2\mu+\beta-\alpha}\\ &\quad + y^2\left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty - \nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{e^{2\mu z} - 1}{2\mu}.\end{aligned}$$

From the solutions to the characteristic equations, we can express each of x, y, and z in terms of the three cumulant generating function parameters:

$$z = t, \qquad y = \theta_Q e^{-\mu t}, \qquad x = \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right).$$

Thus, we can solve for $G_\gamma(\theta_\nu, \theta_Q, t)$ via

$$\begin{aligned}G_\gamma(\theta_\nu, \theta_Q, t) &= g\left(\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right), \theta_Q e^{-\mu t}, t\right)\\ &= \nu_0\theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\nu_0\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right) + Q_0\theta_Q e^{-\mu t}\\ &\quad + \left(\theta_\nu - \frac{\theta_Q}{\beta-\alpha}\right)^2\left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{1 - e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}\\ &\quad + \left(\theta_\nu\theta_Q - \frac{\theta_Q^2}{\beta-\alpha}\right)\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha} + \mu\right)(\nu_\infty - \nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{1 - e^{-(2\mu+\beta-\alpha)t}}{2\mu+\beta-\alpha}\\ &\quad + \theta_Q^2\left(\frac{\gamma\alpha\mu(\nu_\infty - \nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty - \nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{1 - e^{-2\mu t}}{2\mu}.\end{aligned}$$

By Lemma 3, we have that $\hat{\mathcal{M}}_0^\infty(\theta_\nu, \theta_Q, t) \le \hat{\mathcal{M}}^\infty(\theta_\nu, \theta_Q, t) \le \hat{\mathcal{M}}_1^\infty(\theta_\nu, \theta_Q, t)$, and since $\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t) = e^{G_\gamma(\theta_\nu, \theta_Q, t)}$, we have completed the proof of the joint moment generating function bounds when $\beta \neq \alpha$. We now apply this to the two marginal generating functions by setting the opposite space parameter to 0. That is, for the intensity we let $\theta_Q = 0$, yielding

$$\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, 0, t) = \exp\left(\nu_0 \theta_\nu e^{-(\mu+\beta-\alpha)t} + \frac{\theta_\nu^2}{2}\left(\gamma\alpha\mu(\nu_\infty - \nu^*) + \alpha^2\nu_\infty\right)\frac{1 - e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}\right),$$

whereas for the queue we take $\theta_\nu = 0$ and obtain

$$\begin{aligned}\hat{\mathcal{M}}_\gamma^\infty(0, \theta_Q, t) = e^{\frac{\nu_0\theta_Q}{\beta-\alpha}\left(e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}\right) + \frac{\theta_Q^2}{(\beta-\alpha)^2}\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2} + \frac{\alpha^2\nu_\infty}{2}\right)\frac{1-e^{-2(\mu+\beta-\alpha)t}}{2(\mu+\beta-\alpha)}} \\ \cdot e^{Q_0\theta_Q e^{-\mu t} - \frac{\theta_Q^2}{\beta-\alpha}\left(\left(\frac{\gamma\alpha\mu}{\beta-\alpha}+\mu\right)(\nu_\infty-\nu^*) + \frac{\alpha\beta\nu_\infty}{\beta-\alpha}\right)\frac{1-e^{-(2\mu+\beta-\alpha)t}}{2\mu+\beta-\alpha}} \\ \cdot e^{\theta_Q^2\left(\frac{\gamma\alpha\mu(\nu_\infty-\nu^*)}{2(\beta-\alpha)^2} + \frac{\mu(\nu_\infty-\nu^*)}{\beta-\alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty\beta^2}{2(\beta-\alpha)^2}\right)\frac{1-e^{-2\mu t}}{2\mu}}.\end{aligned}$$

Now if $\beta = \alpha$, the solution to the characteristic ODE for θ_ν is instead

$$\theta_\nu = xe^{\mu z} - yze^{\mu z},$$

whereas the solutions for θ_Q and t are unchanged: $\theta_Q = ye^{\mu z}$ and $t = z$. This implies that the ODE for g is given by

$$\begin{aligned}\frac{\mathrm{d}g}{\mathrm{d}z}(x, y, z) = &\left(\frac{\gamma\alpha\mu}{2}\left(x^2e^{2\mu z} - 2xyze^{2\mu z} + y^2z^2e^{2\mu z}\right) + \mu\left(xye^{2\mu z} - y^2ze^{2\mu z}\right)\right)(\nu_\infty - \nu^*) \\ &+ \frac{\nu_\infty}{2}\left((2y^2 + \alpha^2x^2 + 2\alpha xy)e^{2\mu z} - 2(\alpha^2xy + \alpha y^2)ze^{2\mu z} + \alpha^2y^2z^2e^{2\mu z}\right),\end{aligned}$$

which yields a solution of

$$\begin{aligned}g(x, y, z) = xv_0 + yQ_0 + &\left(\left(\left(\frac{\gamma\alpha x^2}{2} + xy\right)\frac{e^{2\mu z} - 1}{2} - \left(\gamma\alpha xy + y^2\right)\frac{e^{2\mu z}(2\mu z - 1) + 1}{4\mu}\right.\right. \\ &\left.\left. + \frac{\gamma\alpha y^2}{2}\left(\frac{e^{2\mu z}\left(2\mu z(\mu z - 1) + 1\right) - 1}{4\mu^2}\right)\right)(\nu_\infty - \nu^*) + \frac{\nu_\infty}{2}\left(\left(2y^2 + \alpha^2x^2 + 2\alpha xy\right)\frac{e^{2\mu z} - 1}{2\mu}\right. \\ &\left. - 2(\alpha^2xy + \alpha y^2)\left(\frac{e^{2\mu z}(2\mu z - 1) + 1}{4\mu^2}\right) + \alpha^2y^2\left(\frac{e^{2\mu z}\left(2\mu z(\mu z - 1) + 1\right) - 1}{4\mu^3}\right)\right).\end{aligned}$$

In this case the inverse solutions are

$$z = t, \qquad y = \theta_Q e^{-\mu t}, \qquad x = \theta_\nu e^{-\mu t} + \theta_Q te^{-\mu t},$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

and so $G_\gamma(\theta_\nu, \theta_Q, t)$ is given by

$$
\begin{aligned}
&G_\gamma(\theta_\nu, \theta_Q, t) = g(\theta_\nu e^{-\mu t} + \theta_Q t e^{-\mu t}, \theta_Q e^{-\mu z}, t) \\
&= \nu_0 \theta_\nu e^{-\mu t} + \nu_0 \theta_Q t e^{-\mu t} + Q_0 \theta_Q e^{-\mu t} + \Bigg(\left(\frac{\gamma\alpha(\theta_\nu + \theta_Q t)^2}{2} + \theta_\nu \theta_Q + \theta_Q^2 t \right) \frac{1 - e^{-2\mu t}}{2} \\
&\quad - \left(\gamma\alpha(\theta_\nu \theta_Q + \theta_Q^2 t) + \theta_Q^2 \right) \frac{2\mu t - 1 + e^{-2\mu t}}{4\mu} + \frac{\gamma\alpha\theta_Q^2}{2} \left(\frac{2\mu t(\mu t - 1) + 1 - e^{-2\mu t}}{4\mu^2} \right) \Bigg) (\nu_\infty - \nu^*) \\
&\quad + \frac{\nu_\infty}{2} \Bigg(\left(2\theta_Q^2 + \alpha^2 \theta_\nu^2 + 2\alpha^2 \theta_\nu \theta_Q t + \alpha^2 \theta_Q^2 t^2 + 2\alpha \theta_\nu \theta_\nu + 2\alpha \theta_Q^2 t \right) \frac{1 - e^{-2\mu t}}{2\mu} \\
&\quad - 2 \left(\alpha^2 \theta_\nu \theta_Q + \alpha^2 \theta_Q^2 t + \alpha \theta_Q^2 \right) \left(\frac{2\mu t - 1 + e^{-2\mu t}}{4\mu^2} \right) + \alpha^2 \theta_Q^2 \left(\frac{2\mu t(\mu t - 1) + 1 - e^{-2\mu t}}{4\mu^3} \right) \Bigg).
\end{aligned}
$$

By taking $\hat{\mathcal{M}}_\gamma^\infty(\theta_\nu, \theta_Q, t) = e^{G_\gamma(\theta_\nu, \theta_Q, t)}$, we complete the proof. □

As a consequence of these diffusion approximations, we can give normally distributed approximations for the steady-state distributions of the HESEP intensity and queue length. These are stated below in Corollary 3, again in terms of γ. One can note that the approximate intensity variance in (41) can be used to provide upper and lower bounds on the HESEP variance that may be tighter than the bounds from the ESEP and the Hawkes process in Proposition 14.

Corollary 3. *Let $(\nu_t, Q_{t,\nu})$ be an HESEP with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, decay rate $\beta > 0$, and rate of exponential service $\mu > 0$, with $\mu + \beta > \alpha$. Then the steady-state distributions of the processes ν_t and $Q_{t,\nu}$ are approximated by the random variables $X_\nu(\gamma) \sim N(\nu_\infty, \sigma_\nu^2(\gamma))$ and $X_Q(\gamma) \sim N\big(\frac{\nu_\infty}{\mu}, \sigma_Q^2(\gamma)\big)$, respectively, where*

$$
\sigma_\nu^2(\gamma) = \frac{\gamma\alpha\mu(\nu_\infty - \nu^*) + \alpha^2 \nu_\infty}{2(\mu + \beta - \alpha)}, \tag{41}
$$

and if $\beta \neq \alpha$ then

$$
\begin{aligned}
\sigma_Q^2(\gamma) = &\frac{\gamma\alpha\mu(\nu_\infty - \nu^*) + \alpha^2 \nu_\infty}{2(\beta - \alpha)^2(\mu + \beta - \alpha)} - \frac{(2\gamma\alpha\mu + 2\mu(\beta - \alpha))\,(\nu_\infty - \nu^*) + 2\alpha\beta\nu_\infty}{(\beta - \alpha)^2(2\mu + \beta - \alpha)} \\
&+ \frac{\gamma\alpha\mu(\nu_\infty - \nu^*) + \nu_\infty \beta^2}{2\mu(\beta - \alpha)^2} + \frac{\nu_\infty - \nu^*}{\beta - \alpha} + \frac{\nu_\infty}{2\mu},
\end{aligned} \tag{42}
$$

whereas if $\beta = \alpha$ then

$$
\sigma_Q^2(\gamma) = \left(\frac{1}{2\mu} + \frac{\gamma\alpha}{4\mu^2} \right) (\nu_\infty - \nu^*) + \left(\frac{1}{\mu} + \frac{\alpha}{2\mu^2} + \frac{\alpha^2}{4\mu^3} \right) \nu_\infty, \tag{43}
$$

with $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$ and $\gamma \in [0, 1]$.

In Figures 6 and 7 we plot the simulated steady-state distributions of an HESEP with large baseline intensities, as calculated from 100,000 replications. We then also plot the densities corresponding to the upper and lower approximate diffusion distributions as well as an additional candidate approximation with $\gamma = \frac{\mu}{\mu+\beta}$. We motivate this choice by a ratio of mean

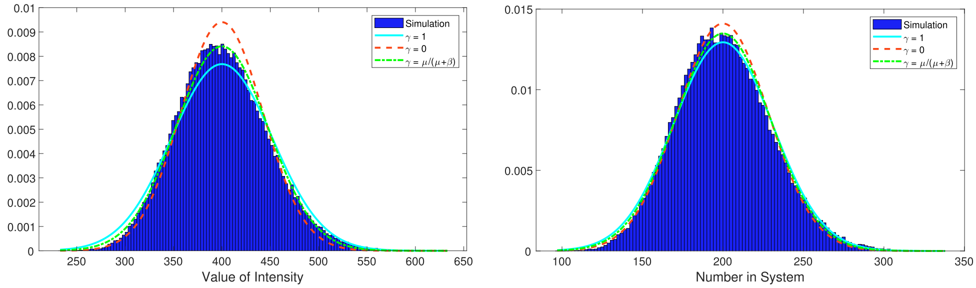


FIGURE 6. Histogram comparing the simulated steady-state HESEP intensity (left) and queue (right) to their diffusion approximations evaluated at multiple values of γ, where $\nu^* = 100$, $\alpha = 3$, $\beta = 2$, and $\mu = 2$.

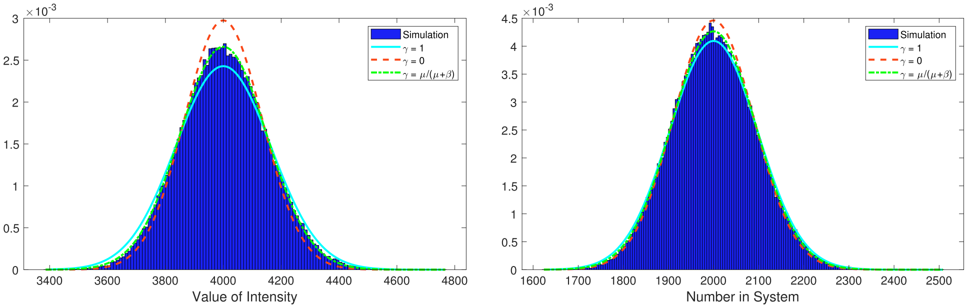


FIGURE 7. Histogram comparing the simulated steady-state HESEP intensity (left) and queue (right) to their diffusion approximations evaluated at multiple values of γ, where $\nu^* = 1,000$, $\alpha = 3$, $\beta = 2$, and $\mu = 2$.

approximations of the terms in (37):

$$\frac{\frac{(\nu_\infty - \nu^*)^2}{\frac{\nu_\infty}{\mu}}}{\alpha(\nu_\infty - \nu^*)} = \frac{\mu(\nu_\infty - \nu^*)}{\alpha\nu_\infty} = \frac{\mu}{\mu + \beta}.$$

In Figure 6 the baseline intensity is equal to 100, whereas in Figure 7 it is 1,000. While there are known limitations of Gaussian approximations for queueing processes, such as those discussed in [48], we see that these approximations appear to be quite close, particularly for the $\nu^* = 1{,}000$ case. The upper and lower bounds predictably over- and under-approximate the tails, while the case of $\gamma = \frac{\mu}{\mu+\beta}$ closely mimics the true distribution.

Appendix C. An ephemeral self-exciting process with finite capacity and blocking

Drawing inspiration from the works that originated queueing theory, we will now consider the change in the ESEP if there is an upper bound on the total number of active exciters. That is, we suppose that there is a finite capacity and no excess buffer beyond them, so that any entities that arrive and find the system full are blocked from entry, thus neither registering an arrival nor causing any excitement. As an employee of the Copenhagen Telephone company, A. K. Erlang developed these pioneering queueing models to determine the probability that a call would be blocked based on the capacity of the telephone network trunk line. Often referred to as the

Erlang-B model, this queueing system remains relevant not just to modern telecommunication systems, but broadly across industries as varied as healthcare operations and transportation. (For English translations of the seminal Erlang papers and a biography of the author, see [14].)

In those original works, Erlang supposed that calls arrive perfectly independently, that they have no influence on or relationship with one another. In the remainder of this subsection we investigate the scenario where the calls instead exhibit self-excitement, which is a potential explanation for the over-dispersion that has been seen in industrial call center data, as detailed in e.g. [38]. Another potential application for this model is a website that may receive viral traffic but is also liable to crash if there are too many simultaneous visitors. Additionally, the finite-capacity model could be used to represent a restaurant that becomes more enticing the more patrons it has in its limited seating area, as we discussed in the introduction.

To begin, we find the steady-state distribution of this process in Proposition 15. Drawing further inspiration from Erlang's work, we will refer to this finite-capacity ESEP model as the *blocking ephemerally self-exciting process* (ESEP-B).

Proposition 15. *Let $\eta_t^{\mathrm{B}} = \eta^* + \alpha Q_t^{\mathrm{B}}$ be an ESEP-B, with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, expiration rate $\beta > \alpha$, and capacity $c \in \mathbb{Z}^+$. That is, if $Q_t^{\mathrm{B}} = c$, any arrivals that occur will be blocked and not recorded. Then the steady-state distribution of the active number in system is given by*

$$\mathbb{P}\left(Q_\infty^{\mathrm{B}} = n\right) = \frac{\mathbb{P}\left(Q_\infty^{\eta} = n\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)} = \frac{\Gamma\left(n + \frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right) n! \left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)} \tag{44}$$

for $0 \le n \le c$ and 0 otherwise, where $\mathbb{P}\left(Q_\infty^{\eta} = n\right)$ is as stated in Theorem 1. Furthermore, the mean and variance of the number in system are given by

$$\mathbb{E}[Q_\infty^B] = \frac{\eta_\infty}{\beta}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right), \tag{45}$$

$$\begin{aligned} Var\left(Q_\infty^B\right) = {} & \frac{\eta_\infty}{\beta}\left(\frac{\eta_\infty}{\beta} + \frac{\alpha}{\beta - \alpha}\right)\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right) - \frac{\eta_\infty^2}{\beta^2}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right)^2 \\ & + \frac{\eta_\infty}{\beta}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right), \end{aligned} \tag{46}$$

where $\eta_\infty = \frac{\beta\eta^*}{\beta-\alpha}$ and

$$I_z(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^z x^{a-1}(1-x)^{b-1}\mathrm{d}x$$

for $z \in [0, 1]$, $a > 0$, and $b > 0$ is the regularized incomplete beta function.

Proof. To prove each of these statements, we first note that for $k \in \mathbb{Z}^+$, $x > 0$, and $p \in (0, 1)$,

$$\sum_{n=0}^{k} \frac{\Gamma\left(n+x\right)}{\Gamma\left(x\right) n!}\left(1-p\right)^x p^n = 1 - I_p\left(k+1, x\right). \tag{47}$$

Hence, we can use (47) to see that

$$\sum_{n=0}^{c} \mathbb{P}\left(Q_\infty^\eta = n\right) = \sum_{n=0}^{c} \frac{\Gamma\left(n+\frac{\eta^*}{\alpha}\right)}{\Gamma\left(\frac{\eta^*}{\alpha}\right) n!} \left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}} \left(\frac{\alpha}{\beta}\right)^n = 1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right).$$

Because the ESEP is a birth–death process, it is reversible. Thus, by truncation we achieve the steady-state distribution; see e.g. Corollary 1.10 in [40]. Then the steady-state mean of the number in system is given by

$$\begin{aligned}
\mathbb{E}[Q_\infty^B] &= \sum_{n=1}^{c} \frac{n\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right) n! \left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)} \\
&= \frac{\frac{\eta^*}{\beta-\alpha}}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)} \sum_{n=1}^{c} \frac{\Gamma\left(n-1+\frac{\eta^*+\alpha}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*+\alpha}{\alpha}}\left(\frac{\alpha}{\beta}\right)^{n-1}}{\Gamma\left(\frac{\eta^*+\alpha}{\alpha}\right)(n-1)!} \\
&= \frac{\eta_\infty}{\beta}\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right),
\end{aligned}$$

where we have again used (47) to simplify the summation. Likewise, the second moment in steady state can be written as

$$\begin{aligned}
\mathbb{E}\left[\left(Q_\infty^B\right)^2\right] &= \sum_{n=1}^{c} \frac{n^2\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right) n! \left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)} \\
&= \sum_{n=2}^{c} \frac{\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right)(n-2)!\left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)} \\
&\quad + \sum_{n=1}^{c} \frac{\Gamma\left(n+\frac{\eta^*}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*}{\alpha}}\left(\frac{\alpha}{\beta}\right)^n}{\Gamma\left(\frac{\eta^*}{\alpha}\right)(n-1)!\left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)} \\
&= \frac{\eta^*(\eta^*+\alpha)}{(\beta-\alpha)^2} \sum_{n=2}^{c} \frac{\Gamma\left(n-2+\frac{\eta^*+2\alpha}{\alpha}\right)\left(\frac{\beta-\alpha}{\beta}\right)^{\frac{\eta^*+2\alpha}{\alpha}}\left(\frac{\alpha}{\beta}\right)^{n-2}}{\Gamma\left(\frac{\eta^*+2\alpha}{\alpha}\right)(n-2)!\left(1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)\right)} + \mathbb{E}[Q_\infty^{\mathrm{B}}] \\
&= \frac{\eta_\infty}{\beta}\left(\frac{\eta_\infty}{\beta} + \frac{\alpha}{\beta-\alpha}\right) \frac{1 - I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)} + \mathbb{E}[Q_\infty^{\mathrm{B}}],
\end{aligned}$$

where once more these sums have been simplified through (47). □

As an illustration of these findings, we now plot both the steady-state distribution and the mean and variance of this blocking system in Figure 8. As can be observed in the figure, the

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

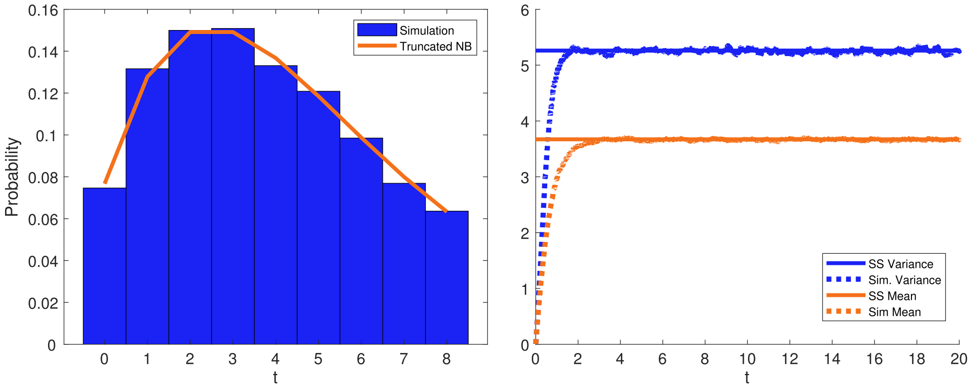


FIGURE 8. Steady-state distribution (left) and mean and variance (right) of the ESEP-B with $\eta^* = 5$, $\alpha = 2$, $\beta = 3$, and $c = 8$, based on 10,000 replications.

system remains over-dispersed even when truncated. We can observe further that this holds in generality as follows. For this, we state two known properties of the regularized incomplete beta function:

$$I_z(a, b) = I_z(a+1, b) + \frac{z^a(1-z)^b}{aB(a, b)}, \qquad I_z(a, b+1) = I_z(a, b) + \frac{z^a(1-z)^b}{bB(a, b)}, \tag{48}$$

where $B(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the beta function. Using these together, we can observe that

$$I_z(a, b) > I_z(a+1, b-1).$$

Thus, we can see that

$$I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right) < I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right) < I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right) < 1,$$

which implies

$$1 > \frac{1 - I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*+\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)} > \frac{1 - I_{\frac{\alpha}{\beta}}\left(c-1, \frac{\eta^*+2\alpha}{\alpha}\right)}{1 - I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}.$$

We now note that the variance is written as the sum of the mean and a positive term and is thus over-dispersed.

We can also note that in the classical Erlang-B model, the famous ‘Poisson arrivals see time averages’ (PASTA) result implies that the steady-state fraction of arrivals that are blocked is equal to the probability that the queue is at capacity in steady-state; see [61]. This is not so for the ESEP-B, as the arrival rate is state-dependent and, more specifically, increases with the queue length. However, in Proposition 16 we find that an equivalent result holds asymptotically as the baseline intensity and the capacity grow large simultaneously. We note that large baseline intensity and capacity are realistic scenarios for many practically relevant applications, including the aforementioned website-crashing scenario.

Proposition 16. *Let $\eta_t^{\mathrm{B}} = \eta^* + \alpha Q_t^{\mathrm{B}}$ be an ESEP-B, with baseline intensity $\eta^* > 0$, intensity jump $\alpha > 0$, exponential service rate $\beta > \alpha$, and capacity $c \in \mathbb{Z}^+$. Then the fraction of arrivals*

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

in steady-state that are blocked, π_{B}, is given by

$$\pi_{\mathrm{B}} = \frac{(\eta^* + \alpha c)\mathbb{P}\left(Q_\infty^\eta = c\right)}{\sum_{k=0}^{c}(\eta^* + \alpha k)\mathbb{P}\left(Q_\infty^\eta = k\right)} = \frac{(\eta^* + \alpha c)\mathbb{P}\left(Q_\infty^{\mathrm{B}} = c\right)}{\eta^* + \alpha\mathbb{E}\left[Q_\infty^{\mathrm{B}}\right]}, \tag{49}$$

where $\mathbb{P}\left(Q_\infty^\eta = k\right)$ is as given in Theorem 1 and $\mathbb{P}\left(Q_\infty^{\mathrm{B}} = c\right)$ and $\mathbb{E}\left[Q_\infty^{\mathrm{B}}\right]$ are as given in Proposition 15. Moreover, if the baseline intensity and the capacity are redefined to be $\eta^* n$ and $c\,n$ for $n \in \mathbb{Z}^+$, then

$$\frac{\pi_{\mathrm{B}}}{\mathbb{P}\left(Q_\infty^{\mathrm{B}} = c\right)} \longrightarrow 1 \tag{50}$$

as $n \to \infty$.

Proof. The expression for the steady-state fraction of arrivals blocked, π_{B}, in (49) follows as a direct consequence from observing that $\eta^* + \alpha k$ is the arrival rate when the queue is in state k. We are thus left to prove (50). By use of (49), we have that the ratio of π_B and $\mathbb{P}\left(Q_\infty^{\mathrm{B}} = c\right)$ is

$$\frac{\pi_{\mathrm{B}}}{\mathbb{P}\left(Q_\infty^{\mathrm{B}} = c\right)} = \frac{\eta^* + \alpha c}{\eta^* + \alpha\mathbb{E}\left[Q_\infty^{\mathrm{B}}\right]} = \frac{\eta^* + \alpha c}{\eta^* + \frac{\alpha\eta^*}{\beta-\alpha}\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(c, \frac{\eta^*}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(c+1, \frac{\eta^*}{\alpha}\right)}\right)},$$

by Proposition 15. Substituting in the scaled forms of the baseline intensity and capacity, $\eta^* n$ and cn, and then dividing the numerator and denominator by cn, this is

$$\frac{\eta^* n + \alpha c n}{\eta^* n + \frac{\alpha\eta^* n}{\beta-\alpha}\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(cn+1, \frac{\eta^* n}{\alpha}\right)}\right)} = \frac{\frac{\eta^*}{c} + \alpha}{\frac{\eta^*}{c} + \frac{\eta^*}{c}\left(\frac{\alpha}{\beta-\alpha}\right)\left(\frac{1-I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha}+1\right)}{1-I_{\frac{\alpha}{\beta}}\left(cn+1, \frac{\eta^* n}{\alpha}\right)}\right)}.$$

From the definition and symmetry of the regularized incomplete beta function, we can note that the ratio of these functions is such that

$$\frac{1 - I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha} + 1\right)}{1 - I_{\frac{\alpha}{\beta}}\left(cn + 1, \frac{\eta^* n}{\alpha}\right)} = \frac{I_{1-\frac{\alpha}{\beta}}\left(\frac{\eta^* n}{\alpha} + 1, cn\right)}{I_{1-\frac{\alpha}{\beta}}\left(\frac{\eta^* n}{\alpha}, cn + 1\right)} = \frac{\alpha c}{\eta^*}\left(\frac{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}}(1-x)^{cn-1}\,\mathrm{d}x}{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}-1}(1-x)^{cn}\,\mathrm{d}x}\right).$$

Now, recognizing an identity for the hypergeometric function ${}_2F_1(a, b; c; z)$, we can re-express this ratio as

$$\begin{aligned}
&\frac{\alpha c}{\eta^*}\left(\frac{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}}(1-x)^{cn-1}\,\mathrm{d}x}{\int_0^{1-\frac{\alpha}{\beta}} x^{\frac{n\eta^*}{\alpha}-1}(1-x)^{cn}\,\mathrm{d}x}\right) \\
&= \frac{\alpha c}{\eta^*}\left(\frac{\frac{1}{\frac{\eta^* n}{\alpha}+1}\left(1-\frac{\alpha}{\beta}\right)^{\frac{\eta^* n}{\alpha}+1}\left(\frac{\alpha}{\beta}\right)^{cn} {}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 2; 1 - \frac{\alpha}{\beta}\right)}{\frac{1}{\frac{\eta^* n}{\alpha}}\left(1-\frac{\alpha}{\beta}\right)^{\frac{\eta^* n}{\alpha}}\left(\frac{\alpha}{\beta}\right)^{cn+1} {}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 1; 1 - \frac{\alpha}{\beta}\right)}\right) \\
&= \frac{\alpha c}{\eta^*}\left(\frac{\eta^* n}{\eta^* n + \alpha}\right)\left(\frac{\beta - \alpha}{\alpha}\right)\frac{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 2; 1 - \frac{\alpha}{\beta}\right)}{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1, 1; cn + 1; 1 - \frac{\alpha}{\beta}\right)}.
\end{aligned}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

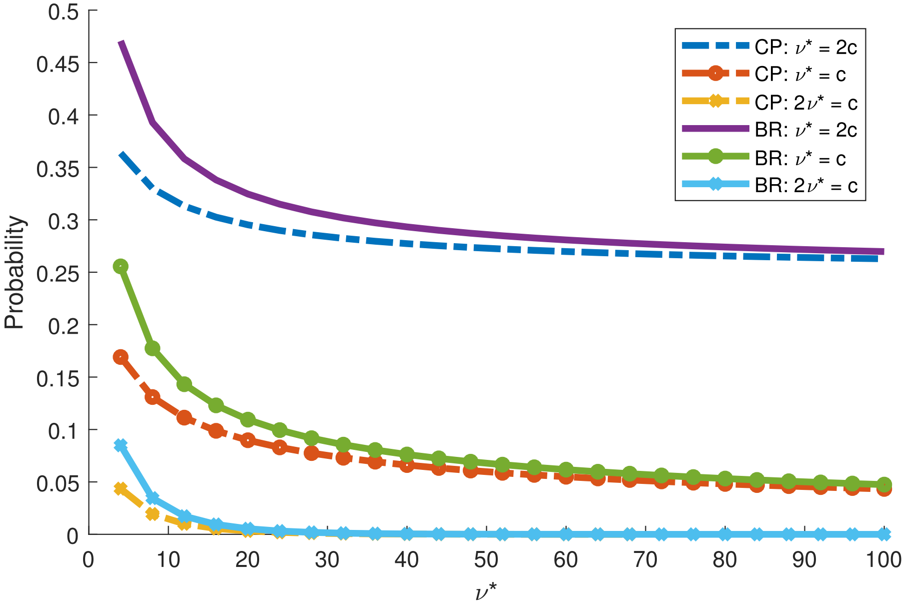


FIGURE 9. Comparison of the ratio of blocked arrivals (BR) and the probability of system being at capacity (CP) when η^* and c are increased simultaneously, for $\alpha = 2$ and $\beta = 3$.

As $n \to \infty$, this yields

$$\frac{\alpha c}{\eta^*}\left(\frac{\eta^* n}{\eta^* n + \alpha}\right)\left(\frac{\beta - \alpha}{\beta}\right)\frac{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1,\, 1; cn + 2; 1 - \frac{\alpha}{\beta}\right)}{{}_2F_1\left(c + \frac{\eta^* n}{\alpha} + 1,\, 1; cn + 1; 1 - \frac{\alpha}{\beta}\right)} \longrightarrow \frac{\alpha c}{\eta^*}\left(\frac{\beta - \alpha}{\alpha}\right),$$

which implies that

$$\frac{\frac{\eta^*}{c} + \alpha}{\frac{\eta^*}{c} + \frac{\eta^*}{c}\left(\frac{\alpha}{\beta - \alpha}\right)\left(\frac{1 - I_{\frac{\alpha}{\beta}}\left(cn, \frac{\eta^* n}{\alpha} + 1\right)}{1 - I_{\frac{\alpha}{\beta}}\left(cn + 1, \frac{\eta^* n}{\alpha}\right)}\right)} \longrightarrow \frac{\frac{\eta^*}{c} + \alpha}{\frac{\eta^*}{c} + \frac{\eta^*}{c}\left(\frac{\alpha}{\beta - \alpha}\right)\frac{\alpha c}{\eta^*}\left(\frac{\beta - \alpha}{\alpha}\right)} = 1.$$

This completes the proof. □

As an example of the convergence stated in Proposition 16, we compare the probability of the system being at capacity and the fraction of blocked arrivals below in Figure 9. In this figure, η^* and c are increased simultaneously according to a fixed ratio. Although at the initial values it is clear that a PASTA-esque result does not hold, as the baseline intensity and capacity both increase one can see that the two curves tend toward one another in each of the different parameter settings.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

Appendix D. Proof of Proposition 3

Proof. Using Proposition 11, we proceed through use of exponential identities for the hyperbolic functions. Specifically, we will make use of the following:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{51}$$

$$\cosh(x) = \frac{e^x + e^{-x}}{2}, \tag{52}$$

and

$$\tanh^{-1}(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right). \tag{53}$$

Using these identities we can further observe that

$$\cosh\left(\tanh^{-1}(x)\right) = \frac{e^{\tanh^{-1}(x)} + e^{-\tanh^{-1}(x)}}{2} = \frac{\left(\frac{1+x}{1-x}\right)^{\frac{1}{2}} + \left(\frac{1-x}{1+x}\right)^{\frac{1}{2}}}{2}.$$

Now, for any time $t \geq 0$ we can note that $N_t = Q_t + D_t$. Thus, we have that

$$\mathbb{E}\left[z^{N_t}\right] = \mathbb{E}\left[z^{Q_t} z^{D_t}\right] = G(z, z, t),$$

where $G(z_1, z_2, t)$ is as given in Proposition 11. Setting $z_1 = z_2 = z$ and $D_0 = N_0 - Q_0$, this is

$$\begin{aligned} G(z, z, t) = &\left(1 - \left(\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z} + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}\right)\right)\right)^2\right)^{\frac{\eta^*}{2\alpha}} \\ &\cdot\left(\frac{\beta+\alpha}{2\alpha} - \frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}\right.\right. \\ &\qquad\left.\left. + \tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}\right)\right)\right)^{Q_0} \\ &\cdot\left(\cosh\left(\tanh^{-1}\left(\frac{2\alpha z - \beta - \alpha}{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}\right)\right)\right)^{\frac{\eta^*}{\alpha}} z^{N_0 - Q_0} e^{\frac{\eta^*(\beta-\alpha)}{2\alpha}t}. \end{aligned} \tag{54}$$

Using the hyperbolic identities and simplifying, this is

$$\begin{aligned} G(z, z, t) = z^{N_0 - Q_0} e^{\frac{\eta^*\eta^*(\beta-\alpha)}{2\alpha}t} &\left(\frac{2e^{\frac{t}{2}\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{\eta^*}{\alpha}} \\ \cdot\left(\frac{\beta+\alpha}{2\alpha} + \frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}{2\alpha}\right.&\left.\left(\frac{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} - \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)^{Q_0}, \end{aligned}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

which is the stated result. However, the simplifications used to reach this form require several steps, which we present individually now. We start with the hyperbolic tangent function that appears on the first and second lines of (54). Using (51) and (53), this is

$$
\begin{aligned}
&-\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)\right)\\
&=-\frac{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}-e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}}{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}+e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}}\\
&=-\frac{e^{\left(t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}-1}{e^{\left(t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}+1}\\
&=\frac{1-\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\\
&=\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}.
\end{aligned}
$$

Thus, the second line of (54) can be simplified to

$$
\begin{aligned}
&\left(\frac{\beta+\alpha}{2\alpha}-\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}\tanh\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\tanh^{-1}\left(\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)\right)\right)^{Q_0}\\
&=\left(\frac{\beta+\alpha}{2\alpha}+\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)^{Q_0}\\
&=\left(\frac{\beta+\alpha}{2\alpha}+\frac{\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}-\frac{\beta+\alpha-2\alpha z}{2\alpha}-\left(\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha}+\frac{\beta+\alpha-2\alpha z}{2\alpha}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}+\left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{Q_0}
\end{aligned}
$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

$$= \left(\frac{\frac{(\beta+\alpha)^2-2\beta\alpha z-2\alpha^2 z}{2\alpha\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} \left(e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-1\right) + \frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} + z - \left(\frac{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}{2\alpha} - z\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right)^{Q_0}$$

$$= \left(\frac{\left(\frac{(\beta-\alpha)z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) \left(e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}-1\right) + z + z e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right)^{Q_0} .$$

Following the same approach, the first line of (54) can be rearranged to

$$\left(1 - \left(\frac{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)} - e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}}{e^{\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)} + e^{-\left(\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}+\frac{1}{2}\log\left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)\right)}} \right)^2 \right)^{\frac{\eta^*}{2\alpha}}$$

$$= \left(1 - \left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} - \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right)^2 \right)^{\frac{\eta^*}{2\alpha}}$$

$$= \left(\frac{4\left(1-\frac{(\beta+\alpha-2\alpha z)^2}{(\beta+\alpha)^2-4\alpha\beta z}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{\left(1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right)^2} \right)^{\frac{\eta^*}{2\alpha}}$$

$$= \left(\frac{\frac{4\alpha\sqrt{z-z^2}}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} e^{\frac{t}{2}\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + \left(1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}\right) e^{t\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} \right)^{\frac{\eta^*}{\alpha}} .$$

Finally, the third line of (54) is simplified through use of (52) and (53). This expression is then given by

$$\left(\frac{\left(\frac{1+\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}} + \left(\frac{1-\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{2\alpha z-\beta-\alpha}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}}}{2} \right)^{\frac{\eta^*}{\alpha}}$$

$$= \left(\left(\frac{\left(\frac{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}} + \left(\frac{1+\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1-\frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}\right)^{\frac{1}{2}}}{2} \right)^2 \right)^{\frac{\eta^*}{2\alpha}}$$

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

$$= \left(\frac{ \frac{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} + 2 + \frac{1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} }{4} \right)^{\frac{\eta^*}{2\alpha}}$$

$$= \left(\frac{ \frac{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} + \frac{1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} + \frac{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} + \frac{1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}}{1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}}} }{4} \right)^{\frac{\eta^*}{2\alpha}}$$

$$= \left(\frac{ 1 - \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} + 1 + \frac{\beta+\alpha-2\alpha z}{\sqrt{(\beta+\alpha)^2-4\alpha\beta z}} }{ 2\left(1 - \frac{(\beta+\alpha-2\alpha z)^2}{(\beta+\alpha)^2-4\alpha\beta z}\right) } \right)^{\frac{\eta^*}{2\alpha}}$$

$$= \left(\frac{\sqrt{(\beta+\alpha)^2 - 4\alpha\beta z}}{2\alpha\sqrt{z - z^2}} \right)^{\frac{\eta^*}{\alpha}} .$$

Together these forms give the stated result. □

Acknowledgements

We are grateful to Robert Hampshire, Emily Fischer, and Sam Gutekunst for helpful discussions. Publicly available drafts of this work used the moniker ‘Queue-Hawkes process’, but we have retitled to avoid confusion and ambiguity. Alas, ‘the process formerly known as the Queue-Hawkes’ is likely not an improvement in this regard.

Funding Information

We acknowledge the generous support of the National Science Foundation (NSF) through Andrew Daw’s Graduate Research Fellowship under grant DGE-1650441, received during his doctoral studies at Cornell University, when this research was initiated.

Competing Interests

There were no competing interests to declare which arose during the preparation or publication process for this article.

References

[1] Ahmed, A. and Xing, E. (2008). Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *Proc. 2008 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Philadelphia, pp. 219–230.

[2] At-Sahalia, Y., Cacho-Diaz, J. and Laeven, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *J. Financial Econom.* **117**, 585–606.

[3] Aldous, D. J. (1985). Exchangeability and related topics. In *École d’ÉtÉ de ProbabilitÉs de Saint-Flour XIII—1983*, Springer, Berlin, Heidelberg, pp. 1–198.

[4] Azizpour, S., Giesecke, K. and Schwenkler, G. (2016). Exploring the sources of default clustering. *J. Financial Econom.* **129**, 154–183.

[5] Bacry, E., Delattre, S., Hoffmann, M. and Muzy, J.-F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stoch. Process. Appl.* **123**, 2475–2499.

[6] Bacry, E., Jaisson, T. and Muzy, J.-F. (2016). Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. *Quant. Finance* **16**, 1179–1201.

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

[8] Ball, F. (1983). The threshold behaviour of epidemic models. *J. Appl. Prob.* **20**, 227–241.

[9] Blackwell, D. (1948). A renewal theorem. *Duke Math. J.* **15**, 145–150.

[10] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–355.

[11] Blanc, P., Donier, J. and Bouchaud, J.-P. (2017). Quadratic Hawkes processes for financial prices. *Quant. Finance* **17**, 171–188.

[12] Blei, D. M. and Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *J. Machine Learning Res.* **12**, 2461–2488.

[13] Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Prob.* **24**, 1563–1588.

[14] Brockmeyer, E., Halstrom, H. and Jensen, A. (1948). *The Life and Works of A. K. Erlang*. Copenhagen Telephone Company, Copenhagen.

[15] Da Fonseca, J. and Zaatour, R. (2014). Hawkes process: fast calibration, application to trade clustering, and diffusive limit. *J. Futures Markets* **34**, 548–579.

[16] Daley, D. J. and Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure*. Springer, New York.

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

[18] Dassios, A. and Zhao, H. (2017). Efficient simulation of clustering jumps with CIR intensity. *Operat. Res.* **65**, 1494–1515.

[19] Davis, M. H. (1984). Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. *J. R. Statist. Soc. B* **46**, 353–388.

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

[21] Daw, A. and Pender, J. (2019). Matrix calculations for moments of Markov processes. Preprint. Available at https://arxiv.org/abs/1909.03320.

[22] Du, N. *et al.* (2015). Dirichlet–Hawkes processes with applications to clustering continuous-time document streams. In KDD '15: *Proc. 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, pp. 219–228.

[23] Eick, S. G., Massey, W. A. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operat. Res.* **41**, 731–742.

[24] Errais, E., Giesecke, K. and Goldberg, L. R. (2010). Affine point processes and portfolio credit risk. *SIAM J. Financial Math.* **1**, 642–665.

[25] Ertekin, Ş., Rudin, C. and McCormick, T. H. (2015). Reactive point processes: a new approach to predicting power failures in underground electrical systems. *Ann. Appl. Statist.* **9**, 122–144.

[26] Farajtabar, M. *et al.* (2017). Fake news mitigation via point process based intervention. In *Proc. 34th International Conference on Machine Learning* (PMLR 70), pp. 1097–1106.

[27] Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*, Vol. **1**, 2nd edn. John Wiley, New York.

[28] Gao, X., Zhou, X. and Zhu, L. (2018). Transform analysis for Hawkes processes with applications in dark pool trading. *Quant. Finance* **18**, 265–282.

[29] 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.

[30] Gao, X. and Zhu, L. (2018). Large deviations and applications for Markovian Hawkes processes with a large initial intensity. *Bernoulli* **24**, 2875–2905.

[31] Guo, X., Ruan, Z. and Zhu, L. (2015). Dynamics of order positions and related queues in a limit order book. Preprint. Available at https://arxiv.org/abs/1505.04810.

[32] Gupta, A., Farajtabar, M., Dilkina, B. and Zha, H. (2018). Discrete interventions in Hawkes processes with applications in invasive species management. In *IJCAI '18: Proc. 27th International Joint Conference on Artificial Intelligence*, AAAI Press, Palo Alto, CA, pp. 3385–3392.

[33] Hale, J. K. and Lunel, S. M. V. (2013). *Introduction to Functional Differential Equations*. Springer, New York.

[34] Halpin, P. F. and De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika* **78**, 793–814.

[35] Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *J. R. Statist. Soc. B* **33**, 438–443.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

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

[37] HAWKES, A. G. AND OAKES, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Prob.* **11**, 493–503.

[38] IBRAHIM, R., YE, H., L'ECUYER, P. AND SHEN, H. (2016). Modeling and forecasting call center arrivals: a literature survey and a case study. *Internat. J. Forecasting* **32**, 865–874.

[39] KARLIS, D. AND XEKALAKI, E. (2005). Mixed Poisson distributions. *Internat. Statist. Rev.* **73**, 35–58.

[40] KELLY, F. P. (2011). *Reversibility and Stochastic Networks*. Cambridge University Press.

[41] KOOPS, D., SAXENA, M., BOXMA, O. AND MANDJES, M. (2018). Infinite-server queues with Hawkes input. *J. Appl. Prob.* **55**, 920–943.

[42] KRUMIN, M., REUTSKY, I. AND SHOHAM, S. (2010). Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Frontiers Comput. Neurosci.* **4**, 147.

[43] LEWIS, E. and MOHLER, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. Preprint. Available at http://paleo.sscnet.ucla.edu/Lewis-Molher-EM_Preprint.pdf.

[44] LI, L. AND ZHA, H. (2018). Energy usage behavior modeling in energy disaggregation via Hawkes processes. *ACM Trans. Intellig. Systems Technol.* 9, 36.

[45] LINDVALL, T. (1977). A probabilistic proof of Blackwell's renewal theorem. *Ann. Prob.* **5**, 482–485.

[46] LUM, K., SWARUP, S., EUBANK, S. AND HAWDON, J. (2014). The contagious nature of imprisonment: an agent-based model to explain racial disparities in incarceration rates. *J. R. Soc. Interface* **11**, 20140409.

[47] MALMGREN, R. D., STOUFFER, D. B., MOTTER, A. E. AND AMARAL, L. A. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proc. Nat. Acad. Sci. USA* 105, 18153–18158.

[48] MASSEY, W. A. AND PENDER, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* **75**, 243–277.

[49] MASUDA, N., TAKAGUCHI, T., SATO, N. AND YANO, K. (2013). Self-exciting point process modeling of conversation event sequences. In *Temporal Networks*, Springer, Berlin, Heidelberg, pp. 245–264.

[50] MEI, H. AND EISNER, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Neural Information Processing Systems, San Diego, CA, pp. 6754–6764.

[51] OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83**, 9–27.

[52] OTTER, R. (1949). The multiplicative process. *Ann. Math. Statist.* **20**, 206–224.

[53] RAMBALDI, M., BACRY, E. AND LILLO, F. (2017). The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quant. Finance* **17**, 999–1020.

[54] RIZOIU, M.-A. *et al.* (2018). SIR–Hawkes: linking epidemic models and Hawkes processes to model diffusions in finite populations. In *WWW '18: Proc. 2018 World Wide Web Conference*, Association for Computing Machinery, New York, pp. 419–428.

[55] RIZOIU, M.-A. *et al.* (2017). Expecting to be hip: Hawkes intensity processes for social media popularity. In *WWW '17: Proc. 26th International Conference on World Wide Web*, Association for Computing Machinery, New York, pp. 735–744.

[56] SINGH, S. AND MYERS, C. R. (2014). Outbreak statistics and scaling laws for externally driven epidemics. *Phys. Rev. E* **89**, 042108.

[57] TRUCCOLO, W. *et al.* (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiology* **93**, 1074–1089.

[58] VAN DER HOFSTAD, R. AND KEANE, M. (2008). An elementary proof of the hitting time theorem. *Amer. Math. Monthly* **115**, 753–756.

[59] VEEN, A. AND SCHOENBERG, F. P. (2008). Estimation of space–time branching process models in seismology using an EM–type algorithm. *J. Amer. Statist. Assoc.* **103**, 614–624.

[60] WILLMOT, G. (1986). Mixed compound Poisson distributions. *ASTIN Bull.* **16**, S59–S79.

[61] WOLFF, R. W. (1982). Poisson arrivals see time averages. *Operat. Res.* **30**, 223–231.

[62] WU, P., RAMBALDI, M., MUZY, J.-F. AND BACRY, E. (2019). *Queue-reactive Hawkes models for the order flow*. Preprint. Available at https://arxiv.org/abs/1901.08938.

[63] XU, H., LUO, D. AND ZHA, H. (2017). Learning Hawkes processes from short doubly-censored event sequences. In *Proc. 34th International Conference on Machine Learning* (PMLR 70), pp. 3831–3840.

[64] XU, H., ZHEN, Y. AND ZHA, H. (2015). Trailer generation via a point process-based visual attractiveness model. In *Proc. Twenty-Fourth International Joint Conference on Artificial Intelligence*, AAAI Press/International Joint Conferences on Artificial Intelligence, Palo Alto, CA, pp. 2198–2204.

[65] YAN, J. *et al.* (2013). Towards effective prioritizing water pipe replacement and rehabilitation. In *Proc. Twenty-Third International Joint Conference on Artificial Intelligence*, AAAI Press/International Joint Conferences on Artificial Intelligence, Menlo Park, CA, pp. 2931–2937.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press

[66] Zhang, X., Blanchet, J., Giesecke, K. and Glynn, P. W. (2015). Affine point processes: approximation and efficient simulation. *Math. Operat. Res.* **40**, 797–819.

[67] Zhang, X.-W., Glynn, P. W., Giesecke, K. and Blanchet, J. (2009). Rare event simulation for a generalized Hawkes process. In *Proc. 2009 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 1291–1298.

[68] Zino, L., Rizzo, A. and Porfiri, M. (2018). Modeling memory effects in activity-driven networks. *SIAM J. Appl. Dynam. Systems* **17**, 2830–2854.

https://doi.org/10.1017/apr.2021.35 Published online by Cambridge University Press