The Queue-Hawkes Process: Ephemeral Self-Excitement

Andrew Daw
School of Operations Research and Information Engineering
Cornell University
257 Rhodes Hall, Ithaca, NY 14853
amd399@cornell.edu

Jamol Pender
School of Operations Research and Information Engineering
Cornell University
228 Rhodes Hall, Ithaca, NY 14853
jjp274@cornell.edu

March 29, 2019

Abstract

Across a wide variety of applications, the self-exciting Hawkes process has been used to model the history of events influencing future occurrences. In this paper, we define a novel generalization of the Hawkes process called the Queue-Hawkes process. This new stochastic process combines the dynamics of a self-exciting process and an infinite server queueing model: arrivals increase the arrival rate, but departures decrease it. By comparison to the Hawkes process, the Queue-Hawkes process is self-excitement on a system rather than on a sequence, making it an ephemerally self-exciting process. Our study of this model includes exploration of the process itself, investigation of relationships between self-exciting processes, and connections to well-known stochastic models such as branching processes, random walks, epidemics, and Bayesian mixture models. Our results for the Queue-Hawkes process include deriving a law of large numbers, fluid limits, and diffusion limit bounds for this new process. Furthermore, we prove a batch scaling construction of general Hawkes processes from a special affine case of the Queue-Hawkes process, which both provides insight into the Hawkes process and motivates the Affine Queue-Hawkes process as an attractive self-exciting process in its own right.

Keywords: Self-exciting processes, Hawkes processes, Queueing theory, Batch scaling.
OR Subjects: Probability: Markov processes, Distributions; Queues: Limit theorems.

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. Hawkes [39] introduced the concept of self-excitation – defining what is now known as the Hawkes process, a model in which “the 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 decays in the interim. In this way, occurrences beget occurrences; hence the term “self-exciting.” Unlike the Poisson process, disjoint increments are not independent in sample paths of 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, such as in Ogata [56], 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 prolific area of work. In these studies, self-excitation is used to capture the often contagious nature of financial activity, see e.g. Errais et al. [28], Bacry et al. [5], Bacry and Muzy [4], Da Fonseca and Zaatour [15], Azizpour et al. [3], Rambaldi et al. [64], Gao et al. [34], Wu et al. [74]. 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 Xu et al. [75], Farajtabar et al. [30], Rizoiu et al. [66, 67]. Notably, this also includes use of Hawkes processes for constructing data-driven methods in the artificial intelligence and machine learning literatures, such as Du et al. [26], Mei and Eisner [54], Xu et al. [76]. Hawkes processes have also recently been used to represent arrivals to service systems in queueing models, e.g. in Gao and Zhu [32, 33], Koops et al. [47], Daw and Pender [21]. 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 [48], environmental management [36], public health [80], energy conservation [50], and industrial preventative maintenance [77].

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, Brémaud and Massoulié [13] introduce a non-linear Hawkes process that adapts the definition of the process intensity to feature a general, non-negative function of the integration over the process history, as opposed to the linear form given originally. Similarly, the quadratic Hawkes process model given by Blanc et al. [10] 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, Dassios and Zhao [17] propose 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 by both the process history and by exogenous shocks. The affine point processes studied in e.g. Errais et al. [28], Zhang et al. [78, 79] 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, imbedding modeling concepts of feedback and dependency into the process intensity. An exact simulation procedure for the Hawkes process with CIR intensity, a generalization of the Hawkes process that is a special case of the affine point process, is shown in Dassios and Zhao [19]. 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, Rizoiu et al. [67] 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, Mei and Eisner [54] introduce 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, Du et al. [26] present 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 Queue-Hawkes process, a generalization of the Hawkes process. Just as many of the preceding works in the generalized Hawkes process literature involve self-exciting dynamics attached to other models, this new process combines self-excitement with infinite server queues. Each event occurrence in the Queue-Hawkes process can be thought of as an entity arriving to a system. Each entity receives service (of potentially random length) and then departs. By comparison to the queueing models driven by Hawkes process arrivals, the excitement generated by an arriving entity in the Queue-Hawkes process is only in effect for the duration of the entity’s time in the system. That is, upon the departure of an entity the intensity of the process will jump downwards. In this way, the self-excitement is ephemeral: it is only in effect as long as the entity is in system. By modeling this phenomenon, the Queue-Hawkes process takes a fundamentally OR-inspired perspective on self-exciting processes, as it captures self-excitement on a system rather than on a sequence.

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”
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 of the spread of the influenza 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 an OR-perspective, the Queue-Hawkes process 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.

Of course, epidemic models need not be exclusively applied to disease spread. These same ideas can be used for information spread and product adoption, such as in the aforementioned Hawkes-infused models in Rizoiu et al. [67] and Zino et al. [80]. 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. We discuss the relevance of the Queue-Hawkes process to epidemics in detail in Subsection 4.3 by relating this model to the Susceptible-Infected-Susceptible (SIS) process through a convergence in distribution. In fact, throughout Section 4 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 modern Bayesian nonparametric models common in artificial intelligence and machine learning models such as the Dirichlet process and the Chinese restaurant process.

In finance, limit order books (LOB’s) are among the many concepts that have been modeled using Hawkes process, such as in Rambaldi et al. [64], Bacry et al. [6]. LOB’s 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 Guo et al. [35], Gao and Zhu [32]. Conceptually, the self-excitement may arise from traders reacting to the activity of other traders, creating runs of transactions. However, the desire to not act on stale information may mean that this excitement only lasts 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 Wu et al. [74], a related work 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 if the pipe is fully, properly repaired it should no longer place strain on its 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 Yan et al. [77], which includes an empirical study. A similar problem for electrical systems is considered in Ertekin et al. [29]. The reactive point process considered in that work is perhaps the model most similar to the Queue-Hawkes process, 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
[29], 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 inter-dependency between these events and to study the resulting ephemerally self-exciting process that arises in these electrical grid failures.

1.2 Organization and Contributions of Paper

Throughout this work we consider three main stochastic models. The first is the Queue-Hawkes process, a novel generalization of the Hawkes process which we define. The two other processes can be seen as special cases of this new process, one in which there is no service and one in which there is no intensity decay. The former results in the original Hawkes process, while the latter is a linear birth-death-immigration process we refer to as the Affine Queue-Hawkes process. Our remaining investigation of these models is organized as follows:

- In Section 2, we introduce the Queue-Hawkes process, bridging a contemporary idea in stochastic models, self-excitement, with one that has a rich history in operations research, queueing theory. Furthermore, we discuss two special cases of this model, the Hawkes and Affine Queue-Hawkes processes, and define the notation used in this paper. Additionally, we relate these three processes to one another through an ordering of their first and second moments.

- In Section 3, we focus on the Affine Queue-Hawkes process. Taking advantage of its analytic tractability, we derive its steady-state distribution, explore the changes to the process under a finite capacity queueing model, and find a matrix calculation for the distribution of the counting process. Finally, we prove that the Affine Queue-Hawkes process converges to the Hawkes process through a batch-scaling, including under general service distributions in the queue and randomly sized batches of arrivals.

- In Section 4, we establish connections from the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes processes to other notable stochastic models. As mentioned previously, this includes classical processes such as branching processes and random walks, as well as modern Bayesian nonparametric models common in artificial intelligence and machine learning contexts such as the Dirichlet process and the Chinese restaurant process. Through these comparisons, we gain added insights for the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes processes.

- In Section 5, we prove limiting results for the Queue-Hawkes process. Specifically, we establish an elementary renewal theorem that results in a law of large numbers for the inter-arrival times. Additionally, we derive a fluid limit of Queue-Hawkes process for a scaling of the baseline intensity, as well as accurate upper and lower approximations for the corresponding diffusion limit.

2 Preliminaries and Models Overview

There are two existing models that are essential to this paper’s ideas: the Hawkes process and the infinite server queue. We now introduce each, beginning with the former. Introduced through the series of papers Hawkes [39, 38], Hawkes and Oakes [40], the Hawkes process is a stochastic intensity point process in which the current rate of arrivals is dependent on the history of arrival process itself. Formally, this is defined as follows: let \((\lambda_t, N^\lambda_t)\) be an intensity and counting
Additional overview of the Hawkes process with the exponential kernel can be found in Section 2 of [21]. 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 [56].

Returning to the latter, let us specify what we refer to as an infinite server queue. In the most general model, often denoted as \( G/G/\infty, \) entities arrive according to some general point process, receive service with random duration drawn from some general distribution, and then depart upon service completion. There are infinitely many servers, so entities do not wait and instead immediately begin receiving service. Throughout this paper, we will use the shorthand “the queue” or “the number in system” when referring to the queue length process, which is the number of entities in the system at the given time. The most elementary infinite server queuing model is the \( M/M/\infty \) queue, in which the first \( M \) denotes that the arrivals occur according to a Poisson process and the second \( M \) means that the service durations are independent, exponentially distributed random variables. In this case, the queue length process is a Markov process, specifically a continuous time Markov chain. In that way, it is useful to think of the \( M \) as standing for either “Markov” or “memoryless.” The memoryless-ness property will be particularly relevant for our investigation. Because the service is memoryless, each entity currently in the system is equally likely to be the next entity to depart, regardless of the order in which they arrived. Furthermore, the overall rate of departures from the system grows linearly
with the number in system. Each of these facts can be seen as consequences of the minimum of \( k \) independent exponential random variables with rate \( \mu \) being exponentially distributed with rate \( k\mu \).

### 2.1 Definition of the Queue-Hawkes Process

Having now reviewed the Hawkes process and the infinite server queue, let us begin to motivate the construction of the Queue-Hawkes process. As discussed in the introduction, we seek an ephemerally self-exciting process in which the arrival of an entity increases the rate of future arrivals, but only while that entity remains in the system. By comparison to the Hawkes-driven queueing systems studied in Gao and Zhu [32, 33], Koops et al. [47], Daw and Pender [21], we want a model in which the arrival intensity responds to both the increases and decreases in the queue length, rather than responding to just the increases in the queue length (or equivalently, the counting process). In this way, we want a process that couples the dynamics of a Hawkes process and an infinite server queue.

Reasoning about this idea, we can form the following desired characteristics for the intensity of our model. Because the self-excitement should be ephemeral, the intensity should jump upwards at arrival epochs and downwards at departures. Furthermore, because the intensity decays between events, the size of the intensity down-jump should decay between events. While this means that the specific size of the down-jump should change with time and with the state of the process, we can specify that when the queue becomes empty, the intensity should return to the baseline. To start simple, we would also like the process to be Markovian. Thus, we can specify further that the size of the down-jump should not depend on the process at any time other than the present.

Using these motivations as a guide, we can now specify the model dynamics in terms of the two event epochs, arrivals and departures, and in terms of the interim. To do so, let us introduce the following notation for the Queue-Hawkes process: let \( \nu_t \) be the arrival intensity, \( Q_{\nu_t} \) be the number in system, and \( N_{\nu_t} \) be the counting process. Borrowing notation from the Markovian Hawkes process, let \( \alpha > 0 \) be the size of the up-jumps of the intensity and let \( \beta \geq 0 \) be the rate of exponential decay in the intensity. Similarly, let \( \nu^* > 0 \) be the baseline intensity. Additionally, let \( \mu \geq 0 \) be the rate of exponential service in the queue. Then, we will define the Queue-Hawkes process behavior as follows:

- At arrival epochs, the intensity jumps by a fixed amount: \( \nu_t \rightarrow \nu_t + \alpha \); the queue increases by one: \( Q_{\nu_t} \rightarrow Q_{\nu_t} + 1 \); the counting process increments by one: \( N_{\nu_t} \rightarrow N_{\nu_t} + 1 \).
- At departure epochs, the intensity decreases by a normalization of the gap between the intensity and its baseline value: \( \nu_t \rightarrow \nu_t - \frac{\nu_t - \nu^*}{Q_{\nu_t}} \); the queue decreases by one: \( Q_{\nu_t} \rightarrow Q_{\nu_t} - 1 \).
- Between events, the intensity decays exponentially towards the baseline intensity: \( \nu_{t+\delta} = \nu^* + (\nu_t - \nu^*)e^{-\beta\delta} \).

Additionally, we can note that the arrival intensity is of course \( \nu_t \), whereas the overall rate of departures is \( \mu Q_{\nu_t} \). With this conceptual reasoning now in hand, we formally define this process below in Definition 2.1.

**Definition 2.1 (The Queue-Hawkes Process).** Let \( t \geq 0 \) and suppose that \( \nu^* > 0 \), \( \alpha > 0 \), \( \beta \geq 0 \), and \( \mu \geq 0 \). Then, define \( \nu_t, N_{\nu_t}, \) and \( Q_{\nu_t} \) such that:

- i) \( N_{\nu_t} \) is an arrival process driven by the intensity \( \nu_t \),
- ii) \( Q_{\nu_t} \) is an infinite server queue driven by the arrivals from \( N_{\nu_t} \) where each entity receives i.i.d. \( \text{Exp}(\mu) \) service,
ii) $\nu_t$ is governed by

$$d\nu_t = \beta(\nu^* - \nu_t)dt + \alpha dN^\nu_t - \frac{\nu_t - \nu^*}{Q^\nu_t}dD^\nu_t$$

where $D^\nu_t = N^\nu_t - Q^\nu_t$.

Then, we say that the intensity-queue-counting process triplet $(\nu_t, Q^\nu_t, N^\nu_t)$ is a **Queue-Hawkes process** with baseline intensity $\nu^*$, intensity jump size $\alpha$, decay rate $\beta$, rate of exponential service $\mu$, and initial values $(\nu_0, Q^\nu_0, N^\nu_0)$.

We can now observe that this meets our desired modeling characteristics. The intensity jumps upwards at arrivals and downwards at departures. This down-jump size decays with the passing of time because $\nu_t$ does while $Q^\nu_t$ does not. Furthermore, if $Q^\nu_t = 1$ this down-jump size is equal to $\nu_t - \nu^*$, meaning that the intensity will return to the baseline whenever the queue returns to empty. While we properly address the Markovian-ness of this process in Proposition 2.1 below, we can quickly note the size of the down-jump does not depend on the process outside of the current time. Moreover, as we have mentioned, each entity currently in system of an infinite server queue is equally likely to be the next to depart because of the memoryless-ness property of the exponential random variable. Because the down-jump size is $\frac{\nu_t - \nu^*}{Q^\nu_t}$, we can note that a similar property holds for the down-jumps of the intensity in Queue-Hawkes process. Each entity currently in the system would have the same effect on the intensity if it is the next to depart, regardless of the order of arrivals. Following this intuition, we now prove that the Queue-Hawkes process satisfies the Markov property.

**Proposition 2.1.** The Queue-Hawkes process $(\nu_t, Q^\nu_t, N^\nu_t)$ is a Markov process.

**Proof.** There are multiple approaches for demonstrating the Markovian nature of this process. Petitioning to the partially deterministic Markov process framework as defined in Davis [20] is perhaps the most direct technique. However, the following arguments are instead rely on a decomposition used in Dassios and Zhao [18] for a procedure for exact simulation of the Hawkes process, as this also motivates an exact simulation procedure for the Queue-Hawkes process [22].

Let $\nu \in [\nu^*, \infty)$ and $Q \in \mathbb{N}$. Let $\mathcal{F}_t$ be the filtration of both the intensity and the queue up to time $t$ and suppose further that $\nu_t = \nu$ and $Q^\nu_t = Q$. Let $S$ represent the time from $t$ until the occurrence of the next event. Then, let $S_{\text{up}}$ and $S_{\text{down}}$ represent the time from $t$ until the next upwards and downwards jumps, respectively. Then, mimicking [18] we decompose $S_{\text{up}}$ into the minimum of two independent random variables $S^{(1)}_{\text{up}}$ and $S^{(2)}_{\text{up}}$ that are also independent of $S_{\text{down}}$, where $P\left(S^{(1)}_{\text{up}} > x \mid \mathcal{F}_t\right) = \exp(-\int_0^x (\nu - \nu^*)^{-\beta}d\xi)$ and $P\left(S^{(2)}_{\text{up}} > x \mid \mathcal{F}_t\right) = e^{-\nu^* x}$. Then, if $\beta > 0$,

$$P\left(S > x \mid \mathcal{F}_t\right) = P\left(S_{\text{up}} > x, S_{\text{down}} > x \mid \mathcal{F}_t\right)$$

$$= e^{-\left(\nu - \nu^*\right)\frac{1 - e^{-\beta x}}{\beta}} \cdot e^{-\nu^* x} \cdot e^{-\mu Q x}$$

$$= P\left(S^{(1)}_{\text{up}} > x \mid \nu_t = \nu, Q^\nu_t = Q\right) \cdot P\left(S^{(2)}_{\text{up}} > x \mid \nu_t = \nu, Q^\nu_t = Q\right)$$

$$\cdot P\left(S_{\text{down}} > x \mid \nu_t = \nu, Q^\nu_t = Q\right)$$

$$= P\left(S > x \mid \nu_t = \nu, Q^\nu_t = Q\right).$$

If $\beta = 0$, we can utilize the affine relationship between the intensity and the queue length,
simplifying this approach to the following:
\[
P(S > x | \mathcal{F}_t) = P(S_{\text{up}} > x, S_{\text{down}} > x | \mathcal{F}_t) \\
= e^{-(\nu^* + \alpha Q)x} \cdot e^{-\mu Qx} \\
= P(S_{\text{up}} > x | \nu_t = \nu, Q_t' = Q) \cdot P(S_{\text{down}} > x | \nu_t = \nu, Q_t'' = Q) \\
= P(S > x | \nu_t = \nu, Q_t' = Q).
\]

We can now conclude that the Queue-Hawkes process is Markovian, as it is constructed through these inter-event times.

For the sake of intuition, we plot a simulated sample path below in Figure 2.1. One can see that at each of the arrival times, denoted by the orange upward triangles, the intensity jumps by the amount \( \alpha = 1 \). At each departure, denoted instead by the grey downward triangles, the intensity jumps down by a normalized amount that varies from epoch to epoch but is always less than \( \alpha \) due to the decay in the intensity. Additionally, when a departure leaves the queue empty, i.e. the number of arrivals and departures becomes equal, the intensity returns to the baseline.

![Figure 2.1: Sample path of the Queue-Hawkes process intensity where \( \nu^* = \alpha = \beta = \mu = 1 \).](image)

### 2.2 Identifying Special Cases of the Queue-Hawkes Process

Let us briefly and loosely review the roles that \( \alpha, \beta, \) and \( \mu \) play in the Queue-Hawkes process. Just like in the Hawkes process, we can observe that in a sense \( \alpha \) is what enacts the self-excitement in the Queue-Hawkes process. The larger the \( \alpha \), the more each arrival will increase the rate of future arrivals. On the other hand, \( \beta \) and \( \mu \) both correspond to ways that the self-excitement can be regulated. As in the Hawkes process, the decay rate \( \beta \) continuously slows the intensity throughout time. By comparison, the service rate \( \mu \) governs the frequency that down-jumps occur, each of which decreases the intensity discontinuously. The larger \( \beta \) is, the less potential each arrival has to excite the process; the larger \( \mu \) is, the less opportunity each arrival has to excite the process.

By isolating either the decay or the service, we can identify two important special cases of the Queue-Hawkes process:

- If there is only decay, i.e. \( \mu = 0 \), then the Queue-Hawkes process is equivalent to a Hawkes process with the same parameters.
• If there is only service, i.e. $\beta = 0$, then at all times the intensity is an affine transformation of the queue length. Because of this, we will refer to this linear birth-death-immigration process as the **Affine Queue-Hawkes** process.

We have reviewed the Hawkes process in the beginning of Section 2 and we will explore the Affine Queue-Hawkes process in detail in Section 3, as we find that it is quite tractable and that it can provide fundamental insights into self-excitement. In fact, throughout the remainder of this paper, we will frequently compare these three processes – the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes processes – and much of the analysis in this work is a result of these comparisons. To distinguish the three cases from one another, we will use the following intensity-based notation shown below in Table 2.1.

| Process                  | Intensity $\nu_t$ | Queue $Q^\nu_t$ | Counting Process $N^\nu_t$ |
|--------------------------|-------------------|-----------------|----------------------------|
| Queue-Hawkes             | $\nu_t$           | $Q^\nu_t$       | $N^\nu_t$                  |
| Hawkes                   | $\lambda_t$       | $Q^\lambda_t$   | $N^\lambda_t$              |
| Affine Queue-Hawkes      | $\eta_t$          | $Q^\eta_t$      | $N^\eta_t$                 |

Table 2.1: Notation for intensity, queue, and counting processes

One can also note that there is also a special case of the Queue-Hawkes process when $\alpha = 0$: the Markovian infinite server queue. If $\alpha = 0$ there is no excitement, and so the arrival process is simply a Poisson process and the queue length process is an $M/M/\infty$. Because this process is absent of self-excitement, this case will not be in the focus of our investigation.

### 2.3 Ordering of the First and Second Moments

To begin forming results for the Queue-Hawkes process and establishing comparisons between it and the Hawkes and Affine Queue-Hawkes processes, we will consider the first and second moments. Because the latter two processes are special cases of the Queue-Hawkes process, we will now derive the transient means for its intensity, queue, and counting processes in Proposition 2.2.

**Proposition 2.2.** Let $(\nu_t, Q^\nu_t, N^\nu_t)$ be a Queue-Hawkes process, then if $\mu + \beta \neq \alpha$ and $\beta \neq \alpha$, the means of the intensity, queue length, and counting process are

\[
E[\nu_t] = \nu_\infty + (\nu_0 - \nu_\infty) e^{-(\mu+\beta-\alpha)t} ,
\]

\[
E[Q^\nu_t] = Q_0 e^{-\mu t} + \frac{\nu_\infty}{\mu} (1 - e^{-\mu t}) - \frac{\nu_0 - \nu_\infty}{\beta - \alpha} \left( e^{-(\mu+\beta-\alpha)t} - e^{-\mu t} \right) ,
\]

\[
E[N^\nu_t] = \nu_\infty t + \frac{\nu_0 - \nu_\infty}{\mu + \beta - \alpha} \left( 1 - e^{-(\mu+\beta-\alpha)t} \right) ,
\]

whereas if $\beta = \alpha$ then

\[
E[Q^\nu_t] = Q_0 e^{-\mu t} + \frac{\nu_\infty}{\mu} (1 - e^{-\mu t}) + (\nu_0 - \nu_\infty) t e^{-\mu t} .
\]

If $\mu + \beta = \alpha$, then the means are given by

\[
E[\nu_t] = \nu_0 + \alpha \nu^* t ,
\]

\[
E[Q^\nu_t] = Q_0 e^{-\mu t} + \frac{\nu_0}{\mu} (1 - e^{-\mu t}) + \alpha \nu^* \left( \frac{t}{\mu} - \frac{1 - e^{-\mu t}}{\mu^2} \right) ,
\]

\[
E[N^\nu_t] = \nu_0 t + \alpha \nu^* \frac{t^2}{2} ,
\]

where $\nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha}$ and $t \geq 0$. 

10
Proof. By use of the infinitesimal generator given in Lemma A.1, we have ordinary differential equations for the means as follows:

$$\frac{d}{dt}E[\nu_t] = (\mu + \beta)\nu^* - (\mu + \beta - \alpha)E[\nu_t], \quad \frac{d}{dt}E[Q_t^\nu] = E[\nu_t] - \mu E[Q_t^\nu], \quad \frac{d}{dt}E[N_t^\nu] = E[\nu_t].$$

Through use of the product rule and the derivatives of exponential functions, we find the stated forms.

As a consequence of Proposition 2.2, we can find the steady-state means of the intensity and the queue by taking the limit as $t \to \infty$. We state this now in Corollary 2.3, which now provides intuition for the term $\nu_\infty$ as the steady-state mean of the intensity.

**Corollary 2.3.** The steady-state means of the intensity and queue length of the Queue-Hawkes process exist if and only if $\mu + \beta > \alpha$. In this case, these are

$$\lim_{t \to \infty} E[\nu_t] = \nu_\infty, \quad \lim_{t \to \infty} E[Q_t^\nu] = \frac{\nu_\infty}{\mu},$$

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

As is known in the literature, the Hawkes process with exponential kernel $g(x) = \alpha e^{-\beta x}$ is stable if and only if $\beta > \alpha$. One can see in Corollary 2.3 that the Queue-Hawkes process has a steady-state mean under a very similar condition, $\mu + \beta > \alpha$. However in this case the ephemerality of the self-excitement creates a weaker requirement with the addition of $\mu$.

To move on to considering higher moments of the process quantities, we can observe through use of the infinitesimal generator given in Lemma A.1 that the time derivative of the second moment of the intensity is given by

$$\frac{d}{dt}E[\nu_t^2] = 2(\mu + \beta)(\nu^*E[\nu_t] - E[\nu_t^2]) + \alpha^2E[\nu_t] + 2\alpha E[\nu_t^2] + \mu E [ (\nu_t - \nu^*)^2].$$

However, this forms a system of differential equations that is not autonomous; we do not have a way to evaluate the lattermost expectation. Thus, we will now develop two pairs of upper and lower bounds on the second moments (or more specifically, the variances) of the Queue-Hawkes process in Propositions 2.4 and 2.5. While the expressions in Proposition 2.5 can be observed to be tighter than those in Proposition 2.4, the latter provides a useful comparison between the three processes studied in this paper: the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes.

**Proposition 2.4.** Let $\alpha > 0$, $\beta > 0$, and $\mu > 0$ be such that $\mu + \beta > \alpha > 0$. Additionally, let $\nu^* > 0$. Let $\nu_t$ be a Queue-Hawkes process with baseline intensity $\nu^*$, intensity jump size $\alpha$, decay rate $\beta$, and service rate $\mu$. Similarly, let $\lambda_t$ be the intensity of a Hawkes process with baseline intensity $\nu^*$, intensity jump $\alpha$, and decay rate $\mu + \beta$. Finally, let $\eta_t$ be a Queue-Hawkes process with baseline intensity $\nu^*$, intensity jump $\alpha$, service rate $\mu + \beta$, and no decay, then the means of these process intensities are all equal:

$$E[\lambda_t] = E[\nu_t] = E[\eta_t].$$

Furthermore, the process variances are ordered such that

$$\text{Var} (\lambda_t) \leq \text{Var} (\nu_t) \leq \text{Var} (\eta_t).$$

Additionally, let $N_t^\nu$, $N_t^\lambda$, and $N_t^\eta$ be the counting processes of the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes processes, respectively. Then, the means of these counting process are equal

$$E \left[ N_t^\lambda \right] = E [N_t^\nu] = E [N_t^\eta].$$
and the variances of these counting processes are again ordered such that

\[ \text{Var} \left( N_t^\lambda \right) \leq \text{Var} \left( N_t^\nu \right) \leq \text{Var} \left( N_t^\eta \right). \]  \hspace{1cm} (2.5)

Finally, the covariances among each intensity and counting process pair are likewise ordered such that

\[ \text{Cov} [\lambda_t, N_t^\lambda] \leq \text{Cov} [\nu_t, N_t^\nu] \leq \text{Cov} [\eta_t, N_t^\eta], \]  \hspace{1cm} (2.6)

where \( t \geq 0 \) and where all intensities have the same initial value.

**Proof.** From Proposition 2.2, we can directly observe that \( E [\nu_t] = E [\eta_t] = E [\lambda_t] \). To show the variance ordering we begin by considering the ODE for the second moment of \( \nu_t \):

\[
\frac{d}{dt} E [\nu_t^2] = 2\beta (\nu^* E [\nu_t] - E [\nu_t^2]) + \alpha^2 E [\nu_t] + 2\alpha E [\nu_t^2] + \mu E \left[ \left( \nu_t - \frac{\nu_t - \nu^*}{Q_t^\nu} \right)^2 - \nu_t^2 \right] Q_t^\nu,
\]

where \( Q_t \) is the queue corresponding to \( \nu_t \). Now, let’s observe that

\[
E \left[ \left( \nu_t - \frac{\nu_t - \nu^*}{Q_t^\nu} \right)^2 - \nu_t^2 \right] Q_t^\nu \geq 2 \left( \nu^* E [\nu_t] - E [\nu_t^2] \right).
\]

This inequality now allows us to directly compare \( \frac{d}{dt} E [\nu_t^2] \) to \( \frac{d}{dt} E [\lambda_t^2] \) and \( \frac{d}{dt} E [\eta_t^2] \). First, we can use Equation 2.7 to see that

\[
\frac{d}{dt} E [\nu_t^2] = 2\beta (\nu^* E [\nu_t] - E [\nu_t^2]) + \alpha^2 E [\nu_t] + 2\alpha E [\nu_t^2] + \mu E \left[ \left( \nu_t - \frac{\nu_t - \nu^*}{Q_t^\nu} \right)^2 - \nu_t^2 \right] Q_t^\nu \\
\geq 2(\mu + \beta) (\nu^* E [\nu_t] - E [\nu_t^2]) + \alpha^2 E [\nu_t] + 2\alpha E [\nu_t^2].
\]

Because we have already shown that \( E [\lambda_t] = E [\nu_t] \), we see that \( \frac{d}{dt} E [\lambda_t^2] \leq \frac{d}{dt} E [\nu_t^2] \) and by Lemma A.2, \( \text{Var} (\lambda_t) \leq \text{Var} (\nu_t) \). By analogous arguments for \( \eta_t \), we achieve the stated result. For the counting process means, we can now observe that all the differential equations are such that

\[
\frac{d}{dt} E [N_t^\lambda] = E [\lambda_t] = \frac{d}{dt} E [N_t^\nu] = E [\nu_t] = \frac{d}{dt} E [N_t^\eta] = E [\eta_t].
\]

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, \( E [\lambda_t] E [N_t^\lambda], E [\nu_t] E [N_t^\nu], \) and \( E [\eta_t] E [N_t^\eta] \), 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

\[
\frac{d}{dt} E [\nu_t N_t^\nu] = -(\mu + \beta - \alpha) E [\nu_t N_t^\nu] + (\mu + \beta) \nu^* E [N_t^\nu] + \alpha E [\nu_t] + E [\nu_t^2],
\]

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, \( E [\nu_t N_t^\nu] \), we can also observe that every function is equivalent across the processes other than the second moments of the intensities. We have
shown that these second moments are in fact ordered and therefore we receive 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

$$\frac{d}{dt} E [(N_t^\nu)^2] = E [\nu_t] + 2E [\nu_t N_t^\nu].$$

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

As a simple consequence of Proposition 2.4, we can note that because the Hawkes process is over-dispersed, i.e. its variance is larger than its mean, so too are the Queue-Hawkes and Affine Queue-Hawkes processes. As discussed, we can also find bounds on the variances of the Queue-Hawkes process quantities without comparison to its special cases but still by use of the comparison lemma, Lemma A.2. The approach in this case follows from recognizing that we can bound the lattermost expression in Equation 2.1 via

$$0 \leq \frac{(\nu_t - \nu^*)^2}{Q_t^\nu} \leq \alpha (\nu_t - \nu^*),$$

as \(0 \leq \nu_t - \nu^* \leq \alpha Q_t^\nu\) by definition. This leads to upper and lower bounds on the transient values of the queue, intensity, and counting process variances we give in Proposition A.4. Because of the length of those expressions, we now simply state the steady-state variance of the intensity in Proposition 2.5.

**Proposition 2.5.** Let \(\nu_t\) be the intensity of a Queue-Hawkes process with baseline intensity \(\nu^* > 0\), intensity jump \(\alpha > 0\), decay rate \(\beta \geq 0\), and service rate \(\mu \geq 0\), where \(\mu + \beta > \alpha\). Then, the variance of the intensity in steady-state is such that

$$\frac{\alpha^2 \nu^*_\infty}{2(\mu + \beta - \alpha)} \leq \lim_{t \to \infty} \text{Var} (\nu_t) \leq \frac{\alpha^2 \nu^*_\infty + \alpha \mu (\nu^*_\infty - \nu^*)}{2(\mu + \beta - \alpha)},$$

where \(\nu^*_\infty = \frac{\mu + \beta - \alpha}{\mu + \beta - \alpha}\) is the steady-state mean.

**Proof.** These bounds follow directly from taking the limit of the expressions in Proposition A.4 as \(t \to \infty\). 

### 3 \(\beta = 0\): The Affine Queue-Hawkes Process

As we have observed in Subsection 2.2, the Queue-Hawkes process has two simpler special cases: the Hawkes process and the Affine Queue-Hawkes process. Since we have reviewed the former in the beginning of Section 2 and since the process has otherwise received a great deal of attention in the literature, we will now focus on the latter. Recall that this special case earns its moniker because the lack of decay means that there is an affine transformation between the intensity and the queue length, i.e.

$$\eta_t = \nu^* + \alpha Q_t^\eta,$$

where \(\nu^* > 0\) is the baseline intensity and \(\alpha > 0\) is the intensity jump size. Here, \(\eta_t\) is the arrival rate to the queue \(Q_t^\eta\) at time \(t \geq 0\), and each entity in the system receives exponentially distributed service at rate \(\mu > 0\). Following the stability conditions we found for the Queue-Hawkes process, we suppose that \(\mu > \alpha\). In this case, \(\alpha\) is both the size of the up-jumps and the down-jumps. This makes it particularly clear that the self-excitement in this process is ephemeral, as entities increase the intensity by \(\alpha\) for only their duration in the system.
We can note that because the value of the intensity is deterministic when given the current queue length (and vice versa) the processes \( \eta_t \) and \( Q_t^\eta \) are each individually Markov processes. For \( N_t^\eta \) as the counting process for the arrival epochs, the pairs \((\eta_t,N_t^\eta)\) and \((Q_t^\eta,N_t^\eta)\) are then Markov processes as well. In Figure 3.1, we show this elegant simplicity of this process through its continuous time Markov chain transition diagram. While linear birth-death-immigration processes have been studied previously, e.g. in Karlin and McGregor [43], Van Doorn [71], it does not appear that they have been considered in the context of self-excitement. In the following subsections we will derive distributional properties for the intensity, queue, and count of the Affine Queue-Hawkes process, consider the change in the process behavior if the queue experiences blocking, and prove that a batch scaling of this process converges to the Hawkes process, including under more general conditions.

3.1 The Affine Queue-Hawkes Intensity and Queue

Because the very definitions of self-exciting processes are concerned with the behavior of their intensities, we will start our investigation of the Affine Queue-Hawkes process by focusing on the intensity and the queue, \( \eta_t \) and \( Q_t^\eta \). We begin by deriving the transient moment generating functions for these two quantities.

**Proposition 3.1.** Let \( \eta_t = \nu^* + \alpha Q_t^\eta \) be the intensity an Affine Queue-Hawkes process with baseline intensity \( \nu^* > 0 \), intensity jump \( \alpha > 0 \), and exponential service rate \( \mu > \alpha \). Then, the moment generating function for the queue length \( Q_t \) is given by

\[
E\left[e^{\theta Q_t^\eta}\right] = \left(\frac{\mu - \alpha e^\theta - \mu(1-e^\theta)e^{-(\mu-\alpha)t}}{\mu - \alpha e^\theta - \alpha(1-e^\theta)e^{-(\mu-\alpha)t}}\right)^Q_0 \left(\frac{\mu}{\mu - \alpha e^\theta} - \frac{\alpha}{\mu - \alpha e^\theta}\left(\frac{\mu - \alpha e^\theta - \mu(1-e^\theta)e^{-(\mu-\alpha)t}}{\mu - \alpha e^\theta - \alpha(1-e^\theta)e^{-(\mu-\alpha)t}}\right)\right)^{\frac{\nu^*}{\alpha}},
\]

for all \( t \geq 0 \) and \( \theta < \log\left(\frac{\mu}{\alpha}\right) \). Then, for \( \eta_t \) the moment generating function is given by

\[
E\left[e^{\theta \eta_t}\right] = \left(\frac{\mu - \alpha e^\theta - \mu(1-e^\theta)e^{-(\mu-\alpha)t}}{\mu - \alpha e^\theta - \alpha(1-e^\theta)e^{-(\mu-\alpha)t}}\right)^{\frac{\nu^*}{\alpha}} \left(\frac{\mu e^{\alpha \theta}}{\mu - \alpha e^{\alpha \theta}} - \frac{\alpha e^{\alpha \theta}}{\mu - \alpha e^{\alpha \theta}}\left(\frac{\mu - \alpha e^{\alpha \theta} - \mu(1-e^{\alpha \theta})e^{-(\mu-\alpha)t}}{\mu - \alpha e^{\alpha \theta} - \alpha(1-e^{\alpha \theta})e^{-(\mu-\alpha)t}}\right)\right)^{\frac{\nu^*}{\alpha}},
\]

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

**Proof.** Using Lemma A.1, we have that the probability generating function for \( Q_t^\eta \), say \( P(z,t) = E\left[z^{Q_t^\eta}\right] \), is given by the solution to the following partial differential equation:

\[
\frac{\partial}{\partial t} E\left[z^{Q_t^\eta}\right] = E\left[(\nu^* + \alpha Q_t^\eta) (z^2 - z) z^{Q_t^\eta-1} + \mu Q_t (1-z) z^{Q_t^\eta-1}\right],
\]
which is equivalently expressed

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

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

$$P(z,t) = \left( \frac{\mu - \alpha z - \mu (1 - z) e^{-(\mu - \alpha)t}}{\mu - \alpha z - \alpha (1 - z) e^{-(\mu - \alpha)t}} \right)^{Q_0} \left( \frac{\mu - \alpha z - \mu (1 - z) e^{-(\mu - \alpha)t}}{\mu - \alpha z - \alpha (1 - z) e^{-(\mu - \alpha)t}} \right)^{\frac{\nu^*}{\alpha}},$$

thus this is the probability generating function for $Q^n_t$. By setting $z = e^\theta$ we receive the moment generating function. Finally, using the affine relationship $\eta_t = \nu^* + \alpha Q^n_t$, we have that

$$E \left[ e^{\theta \eta_t} \right] = E \left[ e^{\theta (\nu^* + \alpha Q^n_t)} \right] = e^{\theta \nu^*} E \left[ e^{\alpha \theta Q^n_t} \right],$$

with $\eta_0 = \nu^* + \alpha Q_0$. \hfill \Box

By taking the limit as time $t \to \infty$, we find the steady-state distributions for each of these processes. Focusing exclusively on the queue length momentarily, we find the moment generating function of the corresponding steady-state distribution in Corollary 3.2.

**Corollary 3.2.** Let $Q^n_t$ be the number in system for an Affine Queue-Hawkes process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and exponential service rate $\mu > \alpha$. The moment generating function for the number in system in steady-state is given by

$$\lim_{t \to \infty} E \left[ e^{\theta Q^n_t} \right] = \left( \frac{\mu - \alpha}{\mu - \alpha e^\theta} \right)^{\frac{\nu^*}{\alpha}},$$

where $\theta < \log \left( \frac{\mu}{\alpha} \right)$.

We can now recognize this form, as it corresponds to a negative binomial distribution. By use of the affine transformation between the queue and the intensity, we now give the precise steady-state distribution for each in Theorem 3.3. Because of the varying definitions of the negative binomial distribution, we state the probability mass function explicitly.

**Theorem 3.3.** Let $\eta_t = \nu^* + \alpha Q^n_t$ be an Affine Queue-Hawkes process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and exponential service rate $\mu > \alpha$. Then, the number in system in steady-state follows a negative binomial distribution with probability of success $\frac{\alpha}{\mu}$ and number of failures $\frac{\nu^*}{\alpha}$, which is to say that the steady-state probability mass function of the queue is

$$P \left( Q^n_\infty = k \right) = \frac{\Gamma \left( k + \frac{\nu^*}{\alpha} \right)}{\Gamma \left( \frac{\nu^*}{\alpha} \right) k!} \left( \frac{\mu - \alpha}{\mu} \right)^{\frac{\nu^*}{\alpha}} \left( \frac{\alpha}{\mu} \right)^k. \quad (3.1)$$

Consequently, the steady-state distribution of the intensity is given by a shifted and scaled negative binomial with probability of success $\frac{\alpha}{\mu}$ and number of failures $\frac{\nu^*}{\alpha}$, shifted by $\nu^*$ and scaled by $\alpha$.

**Proof.** From Corollary 3.2, we can observe that the steady-state moment generating function for the number in queue is equivalent to that of a negative binomial. By the affine transformation $\eta_t = \nu^* + \alpha Q^n_t$, we find the steady-state distribution for the intensity. \hfill \Box
From the known properties of the negative binomial distributions, we note that the steady-state means of the Affine Queue-Hawkes process are

\[ \lim_{t \to \infty} E[\eta_t] = \frac{\mu \nu^*}{\mu - \alpha}, \]

and the steady-state variances are

\[ \lim_{t \to \infty} \text{Var}(\eta_t) = \frac{\alpha^2 \mu \nu^*}{(\mu - \alpha)^2}, \]

\[ \lim_{t \to \infty} \text{Var}(Q_{\eta t}) = \frac{\mu \nu^*}{(\mu - \alpha)^2}. \]

This ease of calculation of the steady-state distribution of a self-exciting process intensity is valuable, as such closed form expressions for the exponential-kernel Hawkes process are not available, despite its popularity in applications. We can also note that the negative binomial distribution is well-studied and has existing inference methods, see for example the discussion of the maximum likelihood estimation in Lloyd-Smith [52], further adding to the practical relevance of the Affine Queue-Hawkes. While maximum likelihood estimation procedures also exist for the Hawkes process, they are prone to challenges such as local maxima and computational difficulty, for reference see Section 1.5 of Rizoiu et al. [65].

### 3.2 Affine Queue-Hawkes Process with Blocking

Drawing inspiration from works that originated queueing theory, we will now consider the change in the Affine Queue-Hawkes process if the queuing system features blocking of arrivals. That is, we suppose that there are finitely many servers and no excess buffer beyond them, so that any entities that arrive and find the queue full are blocked from entry. 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 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 Brockmeyer et al. [14]. In those original works, Erlang supposed that calls arrive perfectly independently, that they have no influence or relationship with one another. In the remainder of this subsection we investigate the scenario where these 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. Ibrahim et al. [42]. 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. To begin, we find the steady-state distribution of this process in Proposition 3.4.

**Proposition 3.4.** Let \( \eta_t^B = \nu^* + \alpha Q_{\eta t}^B \) be an Affine Queue-Hawkes process with blocking, with baseline intensity \( \nu^* > 0 \), intensity jump \( \alpha > 0 \), exponential service rate \( \mu > \alpha \), and capacity \( c \in \mathbb{Z}^+ \). That is, if \( Q_{\eta t}^B = c \) any arrivals that occur will be blocked. Then, the steady-state distribution of the number in system is given by

\[
P(Q_{\infty}^B = n) = \frac{P(Q_{\infty}^B = n)}{1 - I_{\frac{\alpha}{\mu}}(c + 1, \frac{\nu^*}{\alpha})} = \frac{\Gamma \left(n + \frac{\nu^*}{\alpha} \right) \left(\frac{\mu - \alpha}{\mu}\right)^{\frac{\nu^*}{\alpha}} \left(\frac{\alpha}{\mu}\right)^n}{\Gamma \left(\frac{\nu^*}{\alpha}\right) n! \left(1 - I_{\frac{\alpha}{\mu}}(c + 1, \frac{\nu^*}{\alpha})\right)},
\]

for \( 0 \leq n \leq c \) and 0 otherwise, where \( P(Q_{\infty}^B = n) \) is as stated in Theorem 3.3. Furthermore, the mean and variance of the number in system are given by

\[
E(Q_{\infty}^B) = \frac{\eta_{\infty}}{\mu} \left(1 - I_{\frac{\alpha}{\mu}}(c, \frac{\nu^* + \alpha}{\alpha})\right),
\]

\[
\text{Var}(Q_{\infty}^B) = \frac{\mu \nu^*}{(\mu - \alpha)^2}.
\]
\[
\begin{align*}
\text{Var} \left( Q^n_B \right) &= \frac{\eta_\infty}{\mu} \left( \frac{\eta_\infty}{\mu} + \frac{\alpha}{\mu - \alpha} \right) \left( 1 - \frac{I_{\alpha}}{\mu} \left( c - 1, \frac{\nu^* + 2\alpha}{\alpha} \right) \right) - \frac{\eta_\infty^2}{\mu^2} \left( 1 - \frac{I_{\alpha}}{\mu} \left( c, \frac{\nu^* + \alpha}{\alpha} \right) \right)^2 \\
&\quad + \frac{\eta_\infty}{\mu} \left( 1 - I_{\frac{\alpha}{\mu}} \left( c, \frac{\nu^* + \alpha}{\alpha} \right) \right), \\
\end{align*}
\]

where \( \eta_\infty = \frac{\mu^*}{\mu - \alpha} \) and \( I_a(a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \int_0^1 x^{a-1}(1-x)^{b-1} dx \) for \( z \in [0, 1] \), \( a > 0 \) and \( b > 0 \) is the regularized incomplete beta function.

**Proof.** To show each of these, we first note that for \( k \in \mathbb{Z}^+ \), \( x > 0 \), and \( p \in (0, 1) \),
\[
\sum_{n=0}^k \frac{\Gamma(n+x)}{\Gamma(x) n!} (1-p)^n p^n = 1 - I_p(k+1, x) .
\]
Hence, we can use Equation 3.5 to see that
\[
\sum_{n=0}^c P(\eta^n_B = n) = \sum_{n=0}^c \left( \frac{\Gamma(n+\nu^*)}{\Gamma(\nu^*) n!} \left( \frac{\mu - \alpha}{\mu} \right)^{\nu^*} \left( \frac{\alpha}{\mu} \right)^n \right) = 1 - I_p \left( c + 1, \frac{\nu^*}{\alpha} \right).
\]
Because the Affine Queue-Hawkes process is a birth-death process it is reversible. Thus, by truncation we achieve the steady-state distribution, see e.g. Corollary 1.10 in Kelly [45]. Then, the steady-state mean of the number in system is given by
\[
E \left[ Q^n_B \right] = \sum_{n=1}^c \frac{n \Gamma(n+\nu^*)}{\Gamma(\nu^*) n!} \left( \frac{\mu - \alpha}{\mu} \right)^{\nu^*} \left( \frac{\alpha}{\mu} \right)^n = \frac{\nu^*}{1 - I_p \left( c + 1, \frac{\nu^*}{\alpha} \right)} - \frac{\Gamma(n+\nu^*)}{\Gamma(\nu^*) n!} \left( \frac{\mu - \alpha}{\mu} \right)^{\nu^*} \left( \frac{\alpha}{\mu} \right)^{n-1}
\]
where we have again used Equation 3.5 to simplify the summation. Likewise, the second moment in steady-state can be written
\[
E \left[ (Q^n_B)^2 \right] = \sum_{n=1}^c \frac{n^2 \Gamma(n+\nu^*)}{\Gamma(\nu^*) n!} \left( \frac{\mu - \alpha}{\mu} \right)^{\nu^*} \left( \frac{\alpha}{\mu} \right)^n = \frac{\nu^*/(\mu - \alpha)}{(\mu - \alpha)^2} \sum_{n=2}^c \frac{\Gamma(n+\nu^*)}{\Gamma(\nu^*) (n-2)!} \left( \frac{\mu - \alpha}{\mu} \right)^{\nu^*} \left( \frac{\alpha}{\mu} \right)^n + \sum_{n=1}^c \frac{\Gamma(n+\nu^*)}{\Gamma(\nu^*) (n-1)!} \left( \frac{\mu - \alpha}{\mu} \right)^{\nu^*} \left( \frac{\alpha}{\mu} \right)^{n-1} + E \left[ Q^B_\infty \right]
\]
where once more these sums have been simplified through Equation 3.5. \( \square \)
As a demonstration of these findings, we now plot both the steady-state distribution and the mean and variance of this blocking system in Figure 3.2. As can be observed in the figure, this system remains over-dispersed even when truncated. We can observe further that this holds in generality as follows. To observe this, we state two known properties of the regularized incomplete beta function:

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

(3.6)

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 $I_{\mu}(c + 1, \frac{\nu^*}{\alpha}) < I_{\mu}(c, \frac{\nu^* + \alpha}{\alpha}) < I_{\mu}(c - 1, \frac{\nu^* + 2\alpha}{\alpha}) < 1$, and this implies

$$1 > \frac{1 - I_{\mu}(c, \frac{\nu^* + \alpha}{\alpha})}{1 - I_{\mu}(c + 1, \frac{\nu^*}{\alpha})} > \frac{1 - I_{\mu}(c - 1, \frac{\nu^* + 2\alpha}{\alpha})}{1 - I_{\mu}(c + 1, \frac{\nu^*}{\alpha})}.$$

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

![Figure 3.2: Steady-state distribution (left) and mean and variance (right) of the blocking Affine Queue-Hawkes process with $\nu^* = 5$, $\alpha = 2$, $\mu = 3$, and $c = 8$ (Right), based on 10,000 replications.](image)

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 Wolff [73]. This is not so for Affine Queue-Hawkes process with blocking, as the arrival rate is state-dependent and, more specifically, increases with the queue length. However, in Proposition 3.5 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 3.5. Let $\eta^B = \nu^* + \alpha Q^B$ be an Affine Queue-Hawkes process with blocking, with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, exponential service rate $\mu > \alpha$, and capacity $c \in \mathbb{Z}^+$. Then, the fraction of arrivals in steady-state that are blocked $\pi_B$ is given by

$$
\pi_B = \frac{(\nu^* + \alpha c)P(Q^B_\infty = c)}{\sum_{k=0}^{\infty}(\nu^* + \alpha k)P(Q^B_\infty = k)} = \frac{(\nu^* + \alpha c)P(Q^B_\infty = c)}{\nu^* + \alpha c \mathbb{E}[Q^B_\infty]},
$$

(3.7)

where $P(Q^B_\infty = k)$ is as given in Theorem 3.3 and $P(Q^B_\infty = c)$ and $\mathbb{E}[Q^B_\infty]$ are as given in Proposition 3.4. Moreover, if the baseline intensity and the capacity are redefined to be $\nu^* n$ and $cn$ for $n \in \mathbb{Z}^+$, then

$$
\frac{\pi_B}{P(Q^B_\infty = c)} \to 1,
$$

(3.8)

as $n \to \infty$.

Proof. The expression for steady-state fraction of arrivals blocked $\pi_B$ in Equation 3.7 follows as a direct consequence from observing that the $\nu^* + \alpha k$ is the arrival rate when the queue is in state $k$. We are thus left to show Equation 3.8. By use of Equation 3.7, we have that the ratio of $\pi_B$ and $P(Q^B_\infty = c)$ is

$$
\frac{\pi_B}{P(Q^B_\infty = c)} = \frac{\nu^* + \alpha c}{\nu^* n + \alpha \nu^* n/\mu - \alpha} \left(1 - \frac{1}{\mu} \left(c, \frac{\nu^* n}{\alpha} + 1\right)\right) = \frac{\nu^* + \alpha c}{\nu^* + \alpha \mathbb{E}[Q^B_\infty]}.
$$

By use of Proposition 3.4. Substituting in the scaled forms of the baseline intensity and capacity $\nu^* n$ and $cn$ and then dividing the numerator and denominator by $cn$, this is

$$
\frac{\nu^* n + \alpha cn}{\nu^* n + \alpha \nu^* n/\mu - \alpha} = \frac{\nu^* + \alpha c}{\nu^* + \alpha (1 - x)^{cn-1} \mathrm{d}x}.
$$

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

$$
\frac{1 - I_{\frac{\nu^* n}{\mu}} (cn, \frac{\nu^* n}{\alpha} + 1)}{1 - I_{\frac{\nu^* n}{\mu}} (cn + 1, \frac{\nu^* n}{\alpha} + 1)} = \frac{\nu^* + \alpha c}{\nu^* + \alpha (1 - x)^{cn-1} \mathrm{d}x}.
$$

We can now recognize an identity for the hypergeometric function $2F_1(a, b; c; z)$, and thus re-express this ratio as

$$
\frac{\alpha c}{\nu^*} \left(\nu^* n + \alpha\right) \left(\frac{\nu^* n}{\nu^* n + \alpha}\right) \left(\frac{\mu - \alpha}{\mu}\right) \frac{2F_1\left(c + \frac{\nu^* n}{\alpha} + 1, 1; cn + 2; 1 - \frac{\alpha}{\mu}\right)}{2F_1\left(c + \frac{\nu^* n}{\alpha} + 1, 1; cn + 1; 1 - \frac{\alpha}{\mu}\right)} \to \frac{\alpha c}{\nu^*} \left(\frac{\mu - \alpha}{\alpha}\right),
$$

as $n \to \infty$. 

19
which thus implies that
\[
\frac{\nu^* + \alpha}{\nu^* + \frac{\alpha}{\mu - \alpha} \left( \frac{1 - I_{\mu}^k (cn, \frac{\nu^*}{\mu} + 1)}{1 - I_{\mu}^k (cn + 1, \frac{\nu^*}{\mu})} \right)} \rightarrow \frac{\nu^* + \alpha}{\nu^* + \frac{\alpha}{\mu - \alpha} \left( \frac{\mu - \alpha}{\mu - \alpha} \right)} = 1,
\]
and this completes the proof. □

As an example of the convergence stated in Proposition 3.5, we compare the probability of the queue being at capacity and the fraction of blocked arrivals below in Figure 3.3. In this figure, \( \nu^* \) 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.

![Figure 3.3: Comparison of the ratio of blocked arrivals (BR) and the probability of queue being at capacity (CP) when increasing \( \nu^* \) and \( c \) simultaneously, where \( \alpha = 2 \) and \( \mu = 3 \).](image)

3.3 The Affine Queue-Hawkes Counting Process

With having studied the queue length and the intensity for the Affine Queue-Hawkes process we now turn our attention to the counting process. To begin, we find the transient mean and variance of this process in Proposition 3.6, and we note that by doing so we are explicitly stating the upper bound of the Queue-Hawkes counting process variance in Proposition 2.4.

**Proposition 3.6.** Let \( N_t^\eta \) be the number of arrivals by time \( t \) in an Affine Queue-Hawkes process with baseline intensity \( \nu^* > 0 \), intensity jump \( \alpha > 0 \), and exponential service rate \( \mu > \alpha \). Then, the mean and variance of \( N_t \) are given by
\[
E[N_t^\eta] = \eta_\infty t + \frac{\nu_0 - \eta_\infty}{\mu - \alpha} \left( 1 - e^{-(\mu - \alpha)t} \right),
\]
(3.9)
and

$$\text{Var} \left( N_t^\eta \right) = \frac{(\mu^2 + \alpha^2) \eta_\infty}{(\mu - \alpha)^2} t - \frac{2\alpha \mu (\eta_0 - \eta_\infty)}{(\mu - \alpha)^3} \left( e^{-(\mu - \alpha)t} + (\mu - \alpha)t e^{-(\mu - \alpha)t} \right)$$

$$+ \left( \frac{\nu_0 - \eta_\infty}{\mu - \alpha} - \frac{\alpha \nu_0}{(\mu - \alpha)^3} - \frac{\alpha^2 + \alpha \mu}{(\mu - \alpha)^3} \right) (1 - e^{-(\mu - \alpha)t})$$

$$+ \left( \frac{2(\alpha^2 + \alpha \mu) \nu_0}{2(\mu - \alpha)^3} - \frac{\alpha \nu_0}{(\mu - \alpha)^3} \right) (1 - e^{-2(\mu - \alpha)t}) \right), \quad (3.10)$$

for all $t \geq 0$ and with $\eta_\infty = \frac{\mu \nu_0}{\mu - \alpha}$.

**Proof.** Using Lemma A.1, we find these quantities through the solutions to the arising system of differential equations.

Now, we can also move beyond the first and second moment to give the probability generating function of the counting process in closed form below in Proposition 3.7. One can note that by comparison, the generating functions of the Hawkes process are instead only expressible as functions of ordinary differential equations with no known closed form solutions, see for example Subsection 3.5 of [21].

**Proposition 3.7.** Let $N_t^\eta$ be the number of arrivals by time $t \geq 0$ in a Queue-Hawkes process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and exponential service rate $\mu > \alpha$. Then, the probability generating function of $N_t^\eta$ is given by

$$E \left[ z^{N_t^\eta} \right] = e^{\nu^*(\mu - \alpha)t} \left( \frac{2e^{t\sqrt{\mu^2 - 4\alpha \mu z}}}{(\mu + \alpha - 2\alpha z) \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} + (1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{\mu^2 - 4\alpha \mu z}} e^{t\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}})} \right)$$

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

(3.11)

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

**Proof.** See Appendix A.2.

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 3.8.

**Proposition 3.8.** Let $N_t^\eta$ be the number of arrivals by time $t$ in a Queue-Hawkes process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and exponential service rate $\mu > \alpha$. Further, let $Q_0 = k$ be the initial number in system. Then for $i \in \mathbb{N}$, define the matrices $D_i \in \mathbb{R}^{k+i+1 \times k+i+1}$ and $S_i \in \mathbb{R}^{k+i+1 \times k+i+2}$ as

$$D_i = \begin{bmatrix} -((\nu^* + (k+i)(\alpha + \mu)) & (k+i)\mu \\ -((\nu^* + (k+i-1)(\alpha + \mu)) & -((\nu^* + \alpha + \mu) & \mu \\ & & -\nu^* \end{bmatrix},$$

$$S_i = \begin{bmatrix} ((\nu^* + (k+i)(\alpha + \mu)) & (k+i)\mu \\ -((\nu^* + (k+i-1)(\alpha + \mu)) & -((\nu^* + \alpha + \mu) & \mu \\ \cdots & \cdots & -\nu^* \end{bmatrix},$$

$$Q_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}.$$
and

\[
S_k = \begin{bmatrix}
\nu^* + \alpha (k + 1) \\
\vdots \\
\nu^* + \alpha (k + i - 1) \\
\nu^* + \alpha \\
\nu^*
\end{bmatrix}
\]

Further, let \(Z_n \in \mathbb{R}^{d_n \times d_n}\) for \(d_n = \frac{n(n+1)}{2} + (n+1)(k+1)\) be a matrix such that

\[
Z_n = \begin{bmatrix}
D_0 & S_0 & D_1 & S_1 \\
& D_2 & S_2 & \ddots \\
& & \ddots & \ddots \\
& & & D_{n-2} & S_{n-2} \\
& & & & D_{n-1} & S_{n-1} \\
& & & & & D_{n}
\end{bmatrix}
\]

Then, the probability that \(N_t = n\) is given by

\[
P(N_t = n) = v_1^T e^{Z_n t} v_j,
\]

where \(v_j \in \mathbb{R}^{d_n}\) is the unit column vector for the \(j^{th}\) coordinate and \(v_j = \sum_{j=0}^{k+n} v_{d_n-j}\).

Proof. This follows directly from viewing \(Z_n\) as a sub-matrix of the generator matrix of a CTMC, much like one can do to calculate probabilities of phase-type distributions. Specifically, the sub-generator matrix is defined on the state space \(S = \bigcup_{i=0}^{n} \{(0, i), (1, i), \ldots, (k + i - 1, i), (k + i, i)\}\). 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, \(D_j\) 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, \(S_i\) is for transitions from states in \(\{(k + i, i), (k + i - 1, i), \ldots, (1, i), (0, i)\}\) to states in \(\{(k + i + 1, i + 1), (k + i, i + 1), \ldots, (1, i + 1), (0, i + 1)\}\). Then, one can consider this from an absorbing CTMC 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 \(Z_n\) to consider up to \(n\) arrivals. Then, \(e^{Z_n t}\) is the sub-matrix for probabilities of transitions among states in \(S\), 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 states 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 \(\{(n + n, n), (n + n - 1, n), \ldots, (1, n), (0, n)\}\). 

So far in our exploration of the Affine Queue-Hawkes process, we have found that this special case of the Queue-Hawkes process shows analytic tractability exceeding that of the Hawkes process, itself another special case of the Queue-Hawkes process. In the following subsection, we find that in fact the Affine Queue-Hawkes process can be viewed as a pre-limit object that converges to the Hawkes process, unifying these two special cases and providing fundamental insight into the original self-exciting process.

### 3.4 Batch Scaling of the Affine Queue-Hawkes

To motivate the analysis in this subsection, consider the following. Suppose that \(\eta_t(n)\) is the intensity for a variant of the Affine Queue-Hawkes process in which arrivals occur in batches of
size \( n \in \mathbb{Z}^+ \). Specifically, if \( Q_t^n(n) \) is the queue length corresponding to this process, then \( \eta_t(n) \) is given by

\[
\eta_t(n) = \nu^* + \frac{\alpha}{n} Q_t^n(n).
\]

Note that when a batch arrives the group of entities collectively increases the intensity by \( \alpha \). Each entity is then served by one of infinitely many servers. When an entity completes service and departs, the intensity decreases by \( \frac{\alpha}{n} \). Assuming independent, exponentially distributed service at rate \( \mu \), we can write the time derivative of the moment generating function for the intensity and the counting process \( N_t^n(n) \)

\[
\frac{d}{dt} E \left[ e^{\theta_1 \eta_t(n) + \theta_2 N_t^n(n)} \right] = E \left[ \eta_t(n) \left( e^{\alpha \theta_1 + \theta_2} - 1 \right) e^{\theta_1 \eta_t(n) + \theta_2 N_t^n(n)} + n \mu \frac{\eta_t(n) - \nu^*}{\alpha} \left( e^{-\alpha \theta_1 \frac{n}{n}} - 1 \right) \right] e^{\theta_1 \eta_t(n) + \theta_2 N_t^n(n)}.
\]

If we define \( M_n(\theta_1, \theta_2, t) = E \left[ e^{\theta_1 \eta_t(n) + \theta_2 N_t^n(n)} \right] \), we can view this as a partial differential equation:

\[
\frac{\partial}{\partial t} M_n(\theta_1, \theta_2, t) = \left( e^{\alpha \theta_1 + \theta_2} - 1 + \frac{n \mu}{\alpha} \left( e^{-\alpha \theta_1 \frac{n}{n}} - 1 \right) \right) \frac{\partial}{\partial \theta_1} M_n(\theta_1, \theta_2, t) - \frac{n \mu \nu^*}{\alpha} \left( e^{-\alpha \theta_1 \frac{n}{n}} - 1 \right) M_n(\theta_1, \theta_2, t).
\]

If the batch size \( n \) grows infinitely large, one can observe that this PDE converges to

\[
\frac{\partial}{\partial t} M_\infty(\theta_1, \theta_2, t) = \left( e^{\alpha \theta_1 + \theta_2} - 1 - \mu \theta_1 \right) \frac{\partial}{\partial \theta_1} M_\infty(\theta_1, \theta_2, t) + \mu \nu^* \theta_1 M_\infty(\theta_1, \theta_2, t),
\]

and this is the partial derivative for the intensity and count of a Hawkes process, say \((\lambda_t, N_t^\lambda)\) with baseline intensity \( \nu^* \), intensity jump \( \alpha \), and decay rate \( \mu \); this can also be derived by an infinitesimal generator approach, see for example Theorem 4 of [21]. By Theorem 7.16 of Rudin [68], we know that because these differential equations converge, the corresponding moment generating functions also converge. Using Mukherjea et al. [55], this implies that the cumulative distribution functions also converge to those of the Hawkes intensity and counting process, demonstrating that this batch scaling of the Affine Queue-Hawkes process yields the Hawkes process, i.e.

\[
(\eta_t(n), N_t^n(n)) \xrightarrow{D} (\lambda_t, N_t^\lambda).
\]

We summarize this result below in Theorem 3.9.

**Theorem 3.9.** Let the \( n \)th batch-scaled Affine Queue-Hawkes process \((\eta_t(n), N_t^n)\) be defined such that \( N_t^n \) is the counting process for arrival epochs occurring according to the intensity \( \eta_t(n) = \nu^* + \frac{\alpha}{n} Q_t^n(n) \) where \( Q_t^n(n) \) is such that arrivals occur in batches of size \( n \) at the \( N_t^n(n) \) epochs and then individually depart after i.i.d. exponential service with rate \( \mu > 0 \). Then, for all \( t \geq 0 \),

\[
\eta_t(n) \xrightarrow{D} \lambda_t \text{ and } N_t^n(n) \xrightarrow{D} N_t^\lambda
\]

as \( n \to \infty \) with \((\lambda_t, N_t^\lambda)\) as the intensity and counting process of a Hawkes process with baseline intensity \( \nu^* \), intensity jump size \( \alpha \), and decay rate \( \mu \).

Through this batch scaling we now have a pre-limit object that motivates the Hawkes process from a simple Markov chain form. This show that the Hawkes process can be constructed and understood from stochastic processes that are more common or elementary. Additionally,
this limit suggests the Affine Queue-Hawkes (or scalings of it) as potential approximations or substitutions for the Hawkes itself. As we discussed in the introduction to this section, we believe that the Affine Queue-Hawkes process holds great promise for wide use as a self-exciting process.

For additional intuition for this batch scaling, we look to the law of large numbers. That is, note that we can express the intensity as

\[ \eta_t(n) = \nu^* + \frac{\alpha}{n} Q_t(n) = \nu^* + \frac{\alpha}{n} \sum_{i=1}^{N_t} \sum_{j=1}^{n} 1 \{ t < A_i + S_{i,j} \} , \]

where \( A_i \) is the arrival epoch for the \( i \)th batch, \( S_{i,j} \) is the service duration of the \( j \)th entity within that batch, and \( N^\eta_t \) is the number of batches to have arrived by time \( t \). Supposing that a batch arrived at time 0, the remaining excitement from these entities as time \( t \) is

\[ \frac{\alpha}{n} \sum_{j=1}^{n} 1 \{ t < S_j \} \rightarrow \alpha P (S_1 > t) = \alpha e^{-\mu t} , \]

almost surely as \( n \rightarrow \infty \). We can now observe that this law of large numbers scaling has no reliance upon the exponential distribution. Thus, if we utilize other service duration distributions in the Affine Queue-Hawkes process we can yield different decay kernels for the Hawkes process in the batch scaling limit. Using this intuition as a guide, we prove a generalized batch scaling in Theorem 3.10 that incorporates random batch distributions and general service to construct marked, general decay Hawkes processes.

**Theorem 3.10.** For \( t \geq 0 \) and \( n \in \mathbb{Z}^+ \), define \( \eta_t(n) \) and \( Q^\eta_t(n) \) such that \( Q^\eta_t(n) \) is a queue length process with batch arrivals occurring at rate

\[ \eta_t(n) = \nu^* + \frac{\alpha}{n} Q^\eta_t(n) , \quad (3.14) \]

for \( \nu^* > 0 \) and \( \alpha > 0 \). Let the batch sizes be drawn from an i.i.d. sequence of non-negative, discrete random variables \( \{ B_i \mid i \in \mathbb{Z}^+ \} \) and let \( Q_0(n) = 0 \). Furthermore, suppose that there are infinitely many servers, and that the service durations are i.i.d. with cumulative distribution function \( G(x) \). Let \( N^\eta_t(n) \) be the counting process of the resulting arrival epochs. Then, for all \( t \geq 0 \),

\[ \eta_t(n) \overset{D}{\Rightarrow} \lambda_t \text{ and } N^\eta_t(n) \overset{D}{\Rightarrow} N^\lambda_t \quad (3.15) \]

as \( n \rightarrow \infty \), where \( (\lambda_t, N^\lambda_t) \) is the general Hawkes process intensity and counting process pair such that

\[ \lambda_t = \nu^* + \sum_{i=1}^{N^\lambda_t} M_i \tilde{G}(t - A_i) , \quad (3.16) \]

where \( \{ A_i \mid i \in \mathbb{Z}^+ \} \) are the Hawkes process arrival epochs, \( \tilde{G}(x) = 1 - G(x) \), and \( \{ M_i \mid i \in \mathbb{Z}^+ \} \) is an i.i.d. sequence of random variables such that \( \frac{\alpha B_i}{n} \overset{D}{\Rightarrow} M_1 \) and \( \frac{B_i}{n^2} \overset{p}{\rightarrow} 0 \).

**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 \( A^\eta_i \) for \( i \in \mathbb{Z}^+ \) be the time of the \( i \)th arrival in the batch scaled Affine Queue-Hawkes process (where \( n \) is given in context) and similarly let \( A^\lambda_i \) be the \( i \)th 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 \( n \)-batch-scaled Affine Queue-Hawkes processes,

\[
P(A^0_1 > x) = e^{-\nu^* x},
\]
as \( Q_0(n) = 0 \) and thus the first arrival is driven by the external baseline rate. Likewise for the Hawkes process, since Equation 3.16 implies that \( \lambda_t = \nu^* \) for \( 0 \leq t < A^\lambda_1 \), we can see that

\[
P(A^\lambda_1 > x) = e^{-\nu^* x},
\]
and thus \( P(A^\lambda_1 > x) = P(A^0_1 > x) \). As an inductive hypothesis, we now assume that \( \{A^\lambda_1, \ldots, A^\lambda_N\} \) converge in joint and marginal distributions to \( \{A^\lambda_1, \ldots, A^\lambda_k\} \) where \( k \in \mathbb{Z}^+ \). Now, for the Hawkes process we can observe that

\[
P_k \left( A^\lambda_{k+1} - A^\lambda_k > x \right) := P \left( A^\lambda_{k+1} - A^\lambda_k > x \mid \{A^\lambda_1, \ldots, A^\lambda_k\} \right) = E_k \left[ e^{-\int_0^x \lambda \lambda_{\lambda+k+t}^\lambda \lambda dr} \right],
\]
because when conditioned on the arrival history, the Hawkes process behaves like an imhomogeneous Poisson process until the next arrival occurs. Using Equation 3.16, we can express this as

\[
E_k \left[ e^{-\int_0^x \lambda \lambda_{\lambda+k+t}^\lambda \lambda dr} \right] = e^{-\nu^* x} E_k \left[ e^{-\int_0^x M_i G(A_k - A_i + t) dt} \right] = e^{-\nu^* x} \prod_{i=1}^k E_k \left[ e^{-M_i \int_0^x G(A_k - A_i + t) dt} \right].
\]

Turning to the Affine Queue-Hawkes process, we define \( N^\eta_{i,j} \left( (t, t + s) \right) \) as the number of arrivals on the time interval \( (t, t + s) \) that are generated by the excitement caused by the \( j \)th entity within the \( i \)th batch. Furthermore, let \( N^\eta_i \left( (t, t + s) \right) \) be the number of arrivals on \( (t, t + s) \) that are generated by the external, baseline rate \( \nu^* \). Then, using this notation we have that

\[
P_k \left( A^\eta_{k+1} - A^\eta_k > x \mid \{A^\eta_1, \ldots, A^\eta_k\} \right)
= P_k \left( \cap_{i=1}^k \prod_{j=1}^{B_i} \left\{ N^\eta_{i,j} \left( (A^\eta_k, A^\eta_k + x) \right) = 0 \right\} \cap \left\{ N^\eta_i \left( (A^\eta_k, A^\eta_\lambda + x) = 0 \right) \right\} \right)
= E_k \left( \prod_{i=1}^k \prod_{j=1}^{B_i} 1 \left\{ N^\eta_{i,j} \left( (A^\eta_k, A^\eta_k + x) = 0 \right) \right\} \right).
\]

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 receive

\[
E_k \left( \prod_{i=1}^k \prod_{j=1}^{B_i} 1 \left\{ N^\eta_{i,j} \left( (A^\eta_k, A^\eta_k + x) = 0 \right) \right\} \right)
= e^{-\nu^* x} \prod_{i=1}^k E_k \left( \prod_{j=1}^{B_i} 1 \left\{ N^\eta_{i,j} \left( (A^\eta_k, A^\eta_k + x) = 0 \right) \right\} \right).
\]

Consider an arbitrary entity, say the \( j \)th entity in the \( i \)th batch. Let \( S_{i,j} \) be its service duration. If this entity has departed from the queue before \( A^\eta_k \), then it cannot generate further arrivals and thus

\[
P_k \left( N^\eta_{i,j} \left( (A^\eta_k, A^\eta_k + x) = 0 \mid S_{i,j} \leq A^\eta_k - A^\eta_k \right) = 1.
\]
Likewise, if it does not depart until after $A_k^n + x$, then the probability that it generates an arrival on $(A_k^n, A_k^n + x]$ is

$$P_k \left( N_{i,j}^n \left( (A_k^n, A_k^n + x) \right) = 0 \mid S_{i,j} \geq A_k^n - A_i^n + x \right) = e^{-\frac{\alpha}{\pi} x}.$$

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

$$P_k \left( N_{i,j}^n \left( (A_k^n, A_k^n + x) \right) = 0 \mid S_{i,j} = A_k^n - A_i^n + z \right) = e^{-\frac{\alpha}{\pi} z},$$

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

$$e^{-\nu x} \prod_{i=1}^{k} E_k \left[ \prod_{j=1}^{B_i} 1 \{ N_{i,j}^n \left( (A_k^n, A_k^n + x) \right) = 0 \} \right]$$

$$= e^{-\nu x} \prod_{i=1}^{k} E_k \left[ \prod_{j=1}^{B_i} \left( G(A_k^n - A_i^n) + e^{-\frac{\alpha}{\pi} x} G(A_k^n - A_i^n) + \int_0^x e^{-\frac{\alpha}{\pi} z} g(A_k^n - A_i^n + z) dz \right) \right]$$

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 rather just the batch itself, we can evaluate this inside the expectation as

$$e^{-\nu x} \prod_{i=1}^{k} E_k \left[ \left( G(A_k^n - A_i^n) + e^{-\frac{\alpha}{\pi} x} G(A_k^n - A_i^n) + \int_0^x e^{-\frac{\alpha}{\pi} z} g(A_k^n - A_i^n + z) dz \right) \right]$$

$$= e^{-\nu x} \prod_{i=1}^{k} E_k \left[ \left( G(A_k^n - A_i^n) + e^{-\frac{\alpha}{\pi} x} G(A_k^n - A_i^n) + \int_0^x e^{-\frac{\alpha}{\pi} z} g(A_k^n - A_i^n + z) dz \right) \right].$$

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

$$G(A_k^n - A_i^n) + e^{-\frac{\alpha}{\pi} x} G(A_k^n - A_i^n) + \int_0^x e^{-\frac{\alpha}{\pi} z} g(A_k^n - A_i^n + z) dz = e^{-\frac{\alpha}{\pi} x} + \frac{\alpha}{n} \int_0^x e^{-\frac{\alpha}{\pi} z} G(A_k^n - A_i^n + z) dz.$$

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

$$e^{-\frac{\alpha}{\pi} x} + \frac{\alpha}{n} \int_0^x e^{-\frac{\alpha}{\pi} z} G(A_k^n - A_i^n + z) dz = 1 - \frac{\alpha}{n} \int_0^x e^{-\frac{\alpha}{\pi} z} G(A_k^n - A_i^n + z) dz.$$

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
\[
e^{-\nu x} \prod_{i=1}^{k} E_k \left[ \left( G(A^\eta_k - A^\eta_i) + e^{-\frac{\alpha}{n} y} \tilde{G}(A^\eta_k - A^\eta_i) + \int_{0}^{x} e^{-\frac{\alpha}{n} z} g(A^\eta_k - A^\eta_i + z) dz \right) ^{B_i} \right]
\]
\[
= e^{-\nu x} \prod_{i=1}^{k} E_k \left[ e^{-\frac{\alpha}{n} B_i \int_{0}^{x} e^{-\frac{\alpha}{n} y} \tilde{G}(A^\eta_k - A^\eta_i + z) dz + O\left( \frac{B_i}{n^2} \right) } \right] .
\]

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
\[
e^{-\nu x} \prod_{i=1}^{k} E_k \left[ e^{-\frac{\alpha}{n} B_i \int_{0}^{x} e^{-\frac{\alpha}{n} y} \tilde{G}(A^\eta_k - A^\eta_i + z) dz + O\left( \frac{B_i}{n^2} \right) } \right] = e^{-\nu x} \prod_{i=1}^{k} E_k \left[ e^{-\frac{\alpha}{n} M_i \int_{0}^{x} e^{-\frac{\alpha}{n} y} \tilde{G}(A^\eta_k - A^\eta_i + z) dz + O\left( \frac{B_i}{n^2} \right) } \right] ,
\]

We can now take the limit as \( n \to \infty \) and observe that
\[
e^{-\nu x} \prod_{i=1}^{k} E_k \left[ e^{-\frac{\alpha}{n} M_i \int_{0}^{x} e^{-\frac{\alpha}{n} y} \tilde{G}(A^\eta_k - A^\eta_i + z) dz + O\left( \frac{B_i}{n^2} \right) } \right] \to e^{-\nu x} \prod_{i=1}^{k} E_k \left[ e^{-M_i \int_{0}^{x} e^{-\frac{\alpha}{n} y} \tilde{G}(A^\eta_k - A^\eta_i + z) dz } \right] ,
\]
as we have that \( \frac{\alpha}{n} B_1 \overset{D}{\to} M_1 \) and \( \frac{B_i}{n^2} \overset{D}{\to} 0 \). This is now equal to the Hawkes process inter-arrival probability \( P_k(A_{k+1}^\lambda - A_k^\lambda > x) \). 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 now show that the dynamics of the processes converge when we condition on having the same fixed arrival times, which we now denote \( \{ A_i | i \in \mathbb{Z}^+ \} \) for both processes. Since \( N^\eta_i(n) \) is defined as the counting process of arrival epochs rather than total number of arrivals, \( N^\eta_i(n) = N^\Lambda_i \) 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 \( \lambda_{A^-} = \inf_{A_{k-1} \leq t < A_k} \lambda_t \) and \( \eta_{A^-}(n) = \inf_{A_{k-1} \leq t < A_k} \eta_t(n) \) for all \( n \), where \( A_0 = 0 \). Then, the jump in the \( n \)th Affine Queue-Hawkes intensity at the \( k \)th jump is such that
\[
\eta_{A_k}(n) - \eta_{A^-}(n) = \frac{\alpha}{n} B_k \overset{D}{\to} M_k = \lambda_{A_k} - \lambda_{A^-} ,
\]
as \( n \to \infty \). For the behavior between arrival times we first note that for \( S_j \) independent and distributed with CDF \( G(\cdot) \) for all \( j \in \mathbb{Z}^+ \), the probability generating function of \( \frac{1}{n} \sum_{j=1}^{B_1} 1\{ y < S_j \} \) is
\[
E\left[ z^{\frac{1}{n} \sum_{j=1}^{B_1} 1\{ y < S_j \}} \right] = E\left[ \left( G(y) + \tilde{G}(y)z^{\frac{1}{n}} \right)^{B_1} \right] = E\left[ \left( 1 - \tilde{G}(y) \left( 1 - e^{\frac{1}{n} \log z} \right) \right)^{B_1} \right] ,
\]
and by a Taylor expansion approach similar to what we used in the proof’s first part, we can see that
\[
E\left[ \left( 1 - \tilde{G}(y) \left( 1 - e^{\frac{1}{n} \log z} \right) \right)^{B_1} \right] = E\left[ e^{-B_1 \tilde{G}(y) \left( 1 - e^{\frac{1}{n} \log z} \right) + O\left( \frac{B_1}{n^2} \right) } \right] = E\left[ e^{\frac{1}{n} B_1 \tilde{G}(y) \log(z) + O\left( \frac{B_1}{n^2} \right) } \right] .
\]
Taking the limit as \( n \to \infty \), this yields \( E\left[ z^{\frac{1}{n} \sum_{j=1}^{B_1} 1\{ y < S_j \}} \right] \to E\left[ z^{\tilde{G}(y)M_1} \right] \), which is to say that
\[
\frac{1}{n} \sum_{j=1}^{B_1} 1\{ y < S_j \} \overset{D}{\to} \tilde{G}(y)M_1 .
\]
Using this, we can now see that for \( k \in \mathbb{Z}^+ \) and \( 0 \leq x < A_{k+1} - A_k \), the intensity of the \( n \)th batch scaled Affine Queue-Hawkes satisfies

\[
\eta_{A_k + x}(n) = \nu^* + \frac{\alpha}{\eta} \sum_{i=1}^{k} \sum_{j=1}^{B_i} \mathbb{1}_{\{A_k + x < A_i + S_{i,j}\}} = \nu^* + \sum_{i=1}^{k} M_i \tilde{G}(A_k - A_i + x) = \lambda_{A_k + x},
\]

as \( n \to \infty \). Thus, both the jump sizes of \( \eta_t(n) \) and the behavior of \( \eta_t(n) \) between jumps converge to that of \( \lambda_t \), completing the proof.

We can note that the formal proof of Theorem 3.9 now follows as a direct consequence of Theorem 3.10. For an empirical demonstrate of this convergence, in Figure 3.4 we plot cumulative distribution functions for the intensity of the Markovian Affine Queue-Hawkes process across multiple batch sizes and compare them to the empirical distribution of the Markovian Hawkes process. As one can see, in each of the two parameter settings with \( n = 8 \), the distribution of the batch scaled Affine Queue-Hawkes intensity is quite close to that of the Hawkes intensity. Loosely speaking, one can also note that the convergence appears to be faster in the case displayed on the right hand side, which has larger parameter values. In future work, it will be of interest to consider the rate of convergence for this batch-scaling and how those depend on the process parameters.

![Figure 3.4: Empirical steady-state CDF of the scaled Affine Queue-Hawkes process intensity where \( \nu^* = \alpha = 1 \) and \( \mu = 2 \) (left); and where \( \nu^* = 5, \alpha = 2 \) and \( \mu = 3 \) (right), based on 10,000 replications.](image)

As a reference, we list the components of the general Affine Queue-Hawkes process and their corresponding limiting quantities in the general Hawkes process below in Table 3.1. We can note that because the limiting excitation kernel given in Theorem 3.10 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 Section 2. However, it does not include kernels that have a “hump remote from the origin” that Hawkes mentions briefly in the original paper [39]. If desired, this can be remedied through extension to multi-phase service in the general Affine Queue-Hawkes process, with the intensity defined as an affine relationship with one of the later phases. Furthermore, while the generalized Affine Queue-Hawkes process is certainly of interest, we will direct the scope of the remainder to this paper.
to the simple Markovian case for the sake of initial exploration. Nevertheless, dedicated study of the general case is an intriguing course of future work that we intend to pursue.

\[ n \rightarrow \infty \]

| Batch | Service | Affine QH |
|-------|---------|-----------|
| Mark  | Decay   | Hawkes    |

Table 3.1: Overview of convergence details in the batch-scaling of the Affine Queue-Hawkes process

Before concluding this section, let us remark that in addition to providing conceptual understanding into the Hawkes process itself, this batch scaling is also of practical relevance in explaining the use of the Hawkes process in many application settings. For example, in biological applications such as the environmental management problem considered in [36], one of the invasive species studied may produce multiple offspring simultaneously but only for the duration of its life cycle. Similarly, when a user shares a post on a social media platform, it is immediately and simultaneously dispatched to the real-time feeds of many other users, prompting further posts in response. Furthermore, posts are only available for a finite amount of time in an increasing number of popular social media apps. Responses can thus only happen in this specified time window, creating an ephemeral effect. Building on these ideas, we will establish additional connections to applications through a series of comparisons to other notable stochastic processes in Section 4.

4 Insights from 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 Hawkes and Oakes [40]. By viewing a portion of arrivals as immigrants – externally driven and stemming from a homogenous 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 three self-exciting processes that are the subject of this paper, the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes processes.

4.1 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 processes 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 ephemerally self-exciting processes, the Queue-Hawkes and Affine Queue-Hawkes processes, this equates to 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 4.1 we calculate these distributions by way of inhomogeneous Poisson processes, yielding a Poisson mixture form for each.

**Proposition 4.1.** Let $X_\nu$ be the number of new arrivals generated by the excitement caused throughout the duration of an arbitrary initial arrival in the system in a Queue-Hawkes process. Let $\alpha > 0$ be the jump size, $\beta > 0$ be the decay rate, and $\mu > 0$ be the service rate. Then, the probability mass function of this offspring distribution is given by

$$P(X_\nu = k) = \frac{\alpha^k \mu \Gamma \left( \frac{k}{\beta} \right)}{\beta^{k+1} \Gamma \left( k + 1 + \frac{k}{\beta} \right)} _1F_1 \left( k + 1; k + 1 + \frac{\mu}{\beta}; -\frac{\alpha}{\beta} \right).$$ (4.1)

Similarly, let $X_\lambda$ be the number of new arrivals generated by the excitement caused by an arbitrary initial arrival in a Hawkes process with jump size $\alpha$ and decay rate $\beta$. This offspring distribution is then Poisson distributed with probability mass function

$$P(X_\lambda = k) = e^{-\frac{\alpha}{\beta}} \left( \frac{\alpha}{\beta} \right)^k.$$ (4.2)

Finally, let $X_\eta$ be the number of new arrivals generated by the excitement caused by an arbitrary initial arrival throughout its duration in the system in an Affine Queue-Hawkes process with jump size $\alpha$ and service rate $\mu$. Then, this offspring distribution is geometrically distributed with probability mass function

$$P(X_\eta = k) = \left( \frac{\mu}{\alpha + \mu} \right) \left( \frac{\alpha}{\alpha + \mu} \right)^k, \quad \text{where all } k \in \mathbb{N}.$$ (4.3)

**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 $\alpha e^{-\beta t} 1\{t < S_1\}$, $\alpha e^{-\beta t}$, and $\alpha 1\{t < S_2\}$ for the Queue-Hawkes, Hawkes, and Affine Queue-Hawkes processes, respectively, where $S_1, S_2 \sim \text{Exp}(\mu)$ are independent. Using Daley and Vere-Jones [16], one can note that the offspring distributions across all time can then be expressed as

$$X_\nu \sim \text{Pois} \left( \alpha \int_0^\infty e^{-\beta t} 1\{t < S_1\} dt \right), \quad X_\lambda \sim \text{Pois} \left( \alpha \int_0^\infty e^{-\beta t} dt \right), \quad X_\eta \sim \text{Pois} \left( \alpha \int_0^\infty 1\{t < S_2\} dt \right),$$

which are equivalently stated $X_\nu \sim \text{Pois} \left( \frac{\alpha}{\beta} (1 - e^{-\beta S_1}) \right)$, $X_\lambda \sim \text{Pois} \left( \frac{\alpha}{\beta} \right)$, and $X_\eta \sim \text{Pois} (\alpha S_2)$. This now immediately yields the stated distributions for $X_\lambda$ and $X_\eta$, as the Poisson-Exponential mixture is known to yield a geometric distribution, see for example the overview of Poisson mixtures in Karlis and Xekalaki [44]. The probability mass function for $X_\nu$ is then found through conditioning on $S_1$.

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 by
comparison 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 4.1, the Hawkes descendant process is a Poisson branching process and, similarly, the Affine Queue-Hawkes descendant process 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 4.2.

**Lemma 4.2 (Hitting Time Theorem).** The total progeny $Z$ of a branching process with descendant distribution equivalent to $X_1$ is

$$P(Z = k) = \frac{1}{k} P(X_1 + X_2 + \ldots X_k = k - 1),$$

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

**Proof.** See Otter [58] for the original statement and proof in terms of random walks; a review and elementary proof are given in the brief note Van der Hofstad and Keane [70].

We now use the hitting time theorem to give the total descendants distributions for the Hawkes and Affine Queue-Hawkes processes in Proposition 4.3. This is a standard technique for branching processes, and it now yields valuable insight into these two self-exciting models. One can note it is difficult to compute convolutions of the offspring distribution for the general Queue-Hawkes process given in Proposition 4.1. Furthermore, because the down-jump mechanics in the general Queue-Hawkes process depend on the full state of the process at that time, there could be multiple ways of defining the total number of descendants of an arrival and thus progeny is not well-defined. Because of this, we instead focus on the two special cases of it in the following analysis.

**Proposition 4.3.** Let $Z_\lambda$ be a random variable for the total progeny of an arbitrary arrival in a Hawkes process with intensity jump $\alpha > 0$ and decay rate $\beta > \alpha$. Likewise, let $Z_\eta$ be a random variable for the total progeny of an arbitrary arrival in an Affine Queue-Hawkes process with intensity jump $\alpha$ and service rate $\mu > \alpha$. Then, the probability mass functions for $Z_\lambda$ and $Z_\eta$ are given by

$$P(Z_\lambda = k) = e^{-\frac{1}{\beta} \alpha k} \frac{k!}{\beta^k} \left( \frac{\alpha}{\beta} \right)^{k-1}$$

and

$$P(Z_\eta = k) = \frac{1}{k} \left( \frac{2k - 2}{k - 1} \right) \left( \frac{\mu}{\mu + \alpha} \right)^{k-1} \left( \frac{\alpha}{\mu + \alpha} \right)^{1},$$

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

**Proof.** This follows by applying Lemma 4.2 to Proposition 4.1. Because the sum of independent Poisson random variables is Poisson distributed with the sum of the rates, we have that

$$\frac{1}{k} P(X_{\lambda,1} + X_{\lambda,2} + \ldots X_{\lambda,k} = k - 1) = \frac{1}{k} P(K_1 = k - 1),$$

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

$$\frac{1}{k} P(X_{\eta,1} + X_{\eta,2} + \ldots X_{\eta,k} = k - 1) = \frac{1}{k} P(K_2 = k - 1),$$

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

31
For a visual comparison of the descendants in the Affine Queue-Hawkes and Hawkes processes, we plot these two progeny distributions for equivalent parameters in Figure 4.1. As suggested by the variance ordering in Proposition 2.4, the tail of the Affine Queue-Hawkes process progeny distribution is heavier than that of the Hawkes process.

![Figure 4.1: Progeny distributions for Affine Queue-Hawkes and Hawkes processes with \( \alpha = \beta = \frac{1}{2} \).](image)

We can note that while one can calculate the mean of each progeny via the probability mass functions in Proposition 4.3, they can also easily be found using Wald’s identity. We know from Proposition 3.6 that the expected number of arrivals (including by immigration) in the Affine Queue-Hawkes process is

\[
\mathbb{E} \left[ N^\eta_t \right] = \frac{\mu^* t}{\mu - \alpha} + \frac{\nu_0 - \nu_\infty}{\mu - \alpha} \left( 1 - e^{- (\mu - \alpha) t} \right).
\]

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

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

where \( M_t \) is a Poisson process with rate \( \nu^* \) and \( Z_i(t) \) are the total progeny up to time \( t \geq 0 \) that descend from the \( i \)-th immigrant arrival. Now, by applying Wald’s identity to the limit of \( \frac{1}{t} \mathbb{E} \left[ N^\eta_t \right] \) as \( t \to \infty \), we see that

\[
\frac{\mu^*}{\mu - \alpha} = \lim_{t \to \infty} \frac{\mathbb{E} \left[ N^\eta_t \right]}{t} = \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \left[ \sum_{i=1}^{M_t} Z_i(t) \right] = \nu^* \mathbb{E} \left[ Z_0 \right],
\]

and so \( \mathbb{E} \left[ Z_\lambda \right] = \frac{\nu^*}{\mu - \alpha} \). By analogous arguments for the Hawkes process, we see that \( \mathbb{E} \left[ Z_\lambda \right] = \frac{\beta}{\beta - \alpha} \).
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 Affine Queue-Hawkes and Hawkes processes. That is, let the first entity be the first generation, its offspring be the second generation, their offspring the third, and so on. In Proposition 4.4 we find the probability mass function for the Affine Queue-Hawkes process in closed form and a recurrence relation for the cumulative distribution function for the Hawkes process.

**Proposition 4.4.** Let \( G_\lambda \) 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 \( \beta > \alpha \). Then, \( G_\lambda \) has cumulative distribution function \( F_{G_\lambda}(k) = P(G_\lambda \leq k) \) satisfying the recursion
\[
F_{G_\lambda}(k) = e^{-\frac{\alpha}{\beta}(1-F_{G_\lambda}(k-1))},
\]
where \( F_{G_\lambda}(0) = 0 \). Likewise, let \( G_\eta \) be the number of distinct arrival generations across in full the progeny of an initial arrival in an Affine Queue-Hawkes process with intensity jump \( \alpha \) and service rate \( \mu > \alpha \). Then, the probability mass function for \( G_\eta \) is given by
\[
P(G_\eta = k) = \frac{\alpha^{k-1}(\mu - \alpha)}{\mu^k - \alpha^k} - \frac{\alpha^k(\mu - \alpha)}{\mu^{k+1} - \alpha^{k+1}},
\]
where all \( k \in \mathbb{Z}^+ \).

**Proof.** Let \( Y^\lambda_k \) and \( Y^\eta_k \) be Galton-Watson branching processes defined as
\[
Y^\lambda_k = \sum_{i=1}^{Y^\lambda_{k-1}} X^{(k)}_{\lambda,i}, \quad Y^\eta_k = \sum_{i=1}^{Y^\eta_{k-1}} X^{(k)}_{\eta,i},
\]
with \( X^{(k)}_{\lambda,i} \sim \text{Pois} \left( \frac{\alpha}{\beta} \right) \), \( X^{(k)}_{\eta,i} \sim \text{Geo} \left( \frac{\alpha}{\alpha + \mu} \right) \), and \( Y^\lambda_0 = Y^\eta_0 = 1 \). These processes then have probability generating functions
\[
P^\lambda_k(z) = \sum_{j=0}^{\infty} z^j P(Y^\lambda_k = j) \quad \text{and} \quad P^\eta_k(z) = \sum_{j=0}^{\infty} z^j P(Y^\eta_k = j),
\]
that are given by the recursions \( P^\lambda_{k+1}(z) = P_{X\lambda}(P^\lambda_k(z)) \) and \( P^\eta_{k+1}(z) = P_{X\eta}(P^\eta_k(z)) \) with \( P^\lambda_1(z) = P_{X\lambda}(z) \) and \( P^\eta_1(z) = P_{X\eta}(z) \), where \( P_{X\lambda}(z) \) and \( P_{X\eta}(z) \) are the probability generating functions of \( X^{(1)}_{\lambda,i} \) and \( X^{(1)}_{\eta,i} \), respectively; see e.g. Section XII.5 of Feller [31]. One can then use induction to observe that
\[
P^\eta_k(z) = 1 - \frac{\alpha^k(1-z)}{\beta^k + \sum_{j=1}^{k} \alpha^j \beta^{k-j}(1-z)},
\]
whereas \( P^\lambda_k(z) = e^{-\frac{\alpha}{\beta}(1-P^\lambda_{k-1}(z))} \), with \( P^\lambda_1(z) = e^{-\frac{\alpha}{\beta}(1-z)} \). Because of their shared offspring distribution constructions, the number of the progeny in the \( k \)th arrival generations of the Hawkes and Affine Queue-Hawkes processes are equivalent in distribution to \( Y^\lambda_k \) and \( Y^\eta_k \), respectively. In this way, we can express \( G_\lambda \) and \( G_\eta \) as
\[
G_\lambda = \inf\{ k \in \mathbb{Z}^+ \mid Y^\lambda_k = 0 \} \quad \text{and} \quad G_\eta = \inf\{ k \in \mathbb{Z}^+ \mid Y^\eta_k = 0 \}.
\]
This leads us to observe that the events \( \{ G_\lambda = j \} \) and \( \{ Y^\lambda_j = 0, Y^\lambda_{j-1} > 0 \} \) are equivalent, as are \( \{ G_\eta = j \} \) and \( \{ Y^\eta_j = 0, Y^\eta_{j-1} > 0 \} \). Focusing for now on \( G_\lambda \), we have that
\[
P \left( Y^\lambda_j = 0, Y^\lambda_{j-1} > 0 \right) = \sum_{i=1}^{\infty} P \left( X^{(1)}_{\lambda,i} = 0 \right)^i P \left( Y^\lambda_{j-1} = i \right) = \mathcal{P}^\lambda_{j-1} \left( P \left( X^{(1)}_{\lambda,1} = 0 \right) \right) - P \left( Y^\lambda_{j-1} = 0 \right),
\]
and since \( P(K = 0) = P(0) \) for any non-negative discrete random variable \( K \) with probability generating function \( P(z) \), this yields

\[
P(G_\lambda = j) = P_j^\lambda(0) - P_{j-1}^\lambda(0).
\]

Using \( P_0(0) = 0 \), this telescoping sum now produces the stated form of the cumulative distribution function for \( G_\lambda \). By analogous arguments for \( G_\eta \), we complete the proof.

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

### 4.2 Similarities with Bayesian Statistics and Machine Learning Models

In the branching process perspective of the Affine Queue-Hawkes process, 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 progeny 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 11.19 in Aldous [2]. The CRP models the successive arrival of customers to the restaurant that has 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 \( n \)th 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 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 Affine Queue-Hawkes process, a new arrival at time \( t \geq 0 \) was generated as part of active excitement family \( i \) with probability \( \frac{\alpha Q_{t,i}}{\alpha Q_{t} + \nu} \) and otherwise was an externally generated arrival with probability \( \frac{\nu}{\alpha Q_{t} + \nu} \), where \( Q_{t,i} \) is the number of entities in the system at time \( t \) in the \( i \)th excitement family with \( Q_t = \sum_i Q_{t,i} \). By normalizing the numerator and denominator of these probabilities by \( \frac{1}{Q_t} \), we see that these dynamics match the CRP almost exactly. The difference is hardly a novel idea for restaurants – in the Affine Queue-Hawkes process entities 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. Blackwell et al. [9]. The CRP is also established enough to have its own generalizations, such as the distance dependent CRP in Blei and Frazier [11], in which the probability a customer joins a table is dependent on a distance metric, and the recurrent CRP in Ahmed and Xing [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 Affine Queue-Hawkes process, we investigate the distribution of the number of active families in the Affine Queue-Hawkes. Equivalently stated, this is 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 4.5. Suppose that an Affine Queue-Hawkes process 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 $\tau$ be a stopping time such that $\tau = \inf\{t \geq 0 \mid X_t = 0\}$. Then, the expected value of $\tau$ is

$$E[\tau] = \frac{1}{\alpha} \log \left( \frac{\mu}{\mu - \alpha} \right),$$

where $\alpha > 0$ is the intensity jump size and $\mu > \alpha$ is the service 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 non-negative integers. State 0 is the single absorbing state and in any other state $j$ the two possible transitions are to $j+1$ at rate $\alpha j$ and to $j-1$ at rate $\mu j$, as visualized below.

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

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

and this simplifies to the stated result.

Proposition 4.5 gives the expectation of the total time of an excitement family is active in the system. Using this, in Proposition 4.6 we now employ a classical queueing theory result to find the exact distribution of the number of active families simultaneously in the system in steady-state.

Proposition 4.6. Let $B$ be the number of distinct excitement families that have progeny active in the system in steady-state of an Affine Queue-Hawkes process with baseline intensity $\nu^* > 0$, intensity jump $\alpha > 0$, and service rate $\mu > \alpha$. Then, $B \sim \text{Pois} \left( \frac{\nu^*}{\alpha} \log \left( \frac{\mu}{\mu - \alpha} \right) \right)$.

Proof. We first note that new excitement families are started when a baseline-generated arrival occurs, which follows a Poisson process with rate $\nu^*$. The duration excitement family’s time in system then has mean given by Proposition 4.5. 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 $M/G/\infty$ queue in Kendall notation. This process is known to have Poisson distributed steady-state distribution, see e.g. Eick et al. [27], 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 \cite{72}. For

\[ B \sim \text{Pois} \left( \frac{\nu^*}{\alpha} \log \left( \frac{\mu}{\mu - \alpha} \right) \right), \quad Q \sim \text{NegBin} \left( \frac{\alpha}{\mu}, \frac{\nu^*}{\alpha} \right), \quad \text{and} \quad L_i \sim \text{Log} \left( \frac{\alpha}{\mu} \right), \]

then one can observe that

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

where \( P (L_1 = k) = \left( \frac{\alpha}{\mu} \right)^k \left( k \log \left( \frac{\mu}{\mu - \alpha} \right) \right)^{-1} \) 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.

### 4.3 Connections to Epidemic Models

As a final observation regarding the Affine Queue-Hawkes 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 infection through her contact with others. Furthermore when a person recovers from a disease such as the flu, she is no longer contagious and thus she no longer contributes to the rate of disease spread. While we have discussed that this scenario has the hallmarks of self-excitement, 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 \( S_t \) 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 infection 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 CTMC 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, and Figure 4.2 shows 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 \( \nu^* \), the epidemic infection rate as \( \alpha \), and the recovery rate as \( \mu \).

![Figure 4.2: Stochastic SIS model with exogenous infections](image)

One can note that there are immediate similarities between this process and the Affine Queue-Hawkes process. 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 Affine Queue-Hawkes process does not. In Proposition 4.7 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 Affine Queue-Hawkes process.

**Proposition 4.7.** 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 \( \nu^* > 0 \), epidemic infection rate \( \alpha > 0 \), and recovery rate \( \mu > 0 \). Then, as \( N \to \infty \)

\[
I_t \xrightarrow{D} Q^0_t,
\]

where \( Q^0_t \) is the number in system at time \( t \) for an Affine Queue-Hawkes process with baseline intensity \( \nu^* \), intensity jump \( \alpha \), and service rate \( \mu > 0 \).

**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

\[
\frac{d}{dt} E[e^{\theta I_t}] = E\left[ \frac{\alpha I_t S_t}{N} \left( e^\theta - 1 \right) e^{\theta I_t} + \mu I_t \left( e^{-\theta} - 1 \right) e^{\theta I_t} + \frac{\nu^* S_t}{N} \left( e^\theta - 1 \right) e^{\theta I_t} \right]
\]

\[
= E\left[ \frac{\alpha I_t (N - I_t)}{N} \left( e^\theta - 1 \right) e^{\theta I_t} \right] + E \left[ \mu I_t \left( e^{-\theta} - 1 \right) e^{\theta I_t} \right] + E \left[ \frac{\nu^* (N - I_t)}{N} \left( e^\theta - 1 \right) e^{\theta I_t} \right],
\]

which we can re-express in partial differential equation form as

\[
\frac{\partial E[e^{\theta I_t}]}{\partial t} = \left( \alpha \left( e^\theta - 1 \right) + \mu \left( e^{-\theta} - 1 \right) - \frac{\nu^*}{N} \left( e^\theta - 1 \right) \right) \frac{\partial E[e^{\theta I_t}]}{\partial \theta} - \frac{\alpha}{N} \left( e^\theta - 1 \right) \frac{\partial^2 E[e^{\theta I_t}]}{\partial \theta^2}
\]

\[
+ \nu^* \left( e^\theta - 1 \right) E[e^{\theta I_t}].
\]

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

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

which we can recognize as the partial differential equation for the moment generating function of the Affine Queue-Hawkes process based on our analyses in Section 3. \( \square \)

As a demonstration of this convergence, we plot the empirical steady-state distribution of the SIS process for increasing population size below in Figure 4.3. Note that in this example the distributions appear quite close for populations of size 1,000 or larger. On the scale of the populations of cities (or even some larger high schools), this is quite small.

One can note that Proposition 4.7 can also serve as motivation for use of the general Queue-Hawkes process in modeling of infectious diseases, as the decay can represent a person’s decreasing contagiousness as they recover. We would be remiss if we did not note that connections from epidemic models to birth-death processes are not new. For example, Ball [7] demonstrated that epidemic models converge to birth-death processes, and Singh and Myers [69] even noted that the exogenously driven Susceptible-Infected-Recovered (SIR) model – that is, people cannot become re-infected – converges to a linear birth-death-immigration process; however, these works did not outright form connections to self-exciting processes. In Rizoiu et al. [67], 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, the topics considered in
these prior works serve to expand the practical relevance of the Affine Queue-Hawkes process, as they note that these epidemic models are also of use outside of public health. For example, the contagious nature of these models has also been used to study topics like product adoption, idea spread, and social influence. These all also naturally relate to the concept of self-excitement, and in Proposition 4.7 we observe that this connection can be formalized.

5 Limits of the Queue-Hawkes Process

In this final section of analysis in this paper, we obtain limiting results for the general Queue-Hawkes process. We begin with an elementary renewal theorem result that we then use to find a strong law of large numbers for the inter-arrival times. By extension to the special cases, this applies to both the Hawkes process and Affine Queue-Hawkes process. In the final two results we derive approximations for the distributions of the Queue-Hawkes process. By comparison to the special case Affine Queue-Hawkes process, neither the general Queue-Hawkes process nor the Hawkes process special case offer the tractability that, for example, allowed us to calculate the steady-state distribution for the affine case. Hence, we turn to approximating the distribution. Drawing inspiration from the queueing theory literature, we find fluid and diffusion approximations of the Queue-Hawkes process when scaling the baseline intensity.
5.1 Strong Convergence of the Queue-Hawkes Counting Process

We begin with the almost sure convergence of the ratio of the Queue-Hawkes counting process and time, which is an elementary renewal result in the style of Blackwell [8] or Lindvall [51], for example. However, by comparison to the context of such works, we know the mean and variance of the process via Proposition 3.6 and we are instead solely interested in establishing the convergence, as we will obtain additional results by consequence. Using these expressions for the first two moments, the proof of Theorem 5.1 follows standard approaches using the Borel-Cantelli lemma. In Corollary 5.2 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 Queue-Hawkes process by way of the continuous mapping theorem, which is another standard technique.

**Theorem 5.1.** Let \((\nu_t, Q^\nu_t, N^\nu_t)\) be a Queue-Hawkes counting process with baseline intensity \(\nu^*\), intensity jump \(\alpha > 0\), intensity decay rate \(\beta \geq 0\), and rate of exponentially distributed service \(\mu \geq 0\), where \(\mu + \beta > \alpha\). Then,

\[
\frac{N^\nu_t}{t} \xrightarrow{a.s.} \nu_\infty
\]

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

*Proof. See Appendix A.3.*

**Corollary 5.2.** Let \((\nu_t, Q^\nu_t, N^\nu_t)\) be an Queue-Hawkes 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 \(S^\nu_k\) denote the \(k\)th inter-arrival time for \(k \in \mathbb{Z}^+\). Then,

\[
\frac{1}{n} \sum_{k=1}^n S^\nu_k \xrightarrow{a.s.} \frac{1}{\nu_\infty}
\]

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

*Proof. Let \(A^\nu_n\) denote the time of the \(n\)th arrival for each \(n \in \mathbb{Z}^+\), which is to say that \(A^\nu_n = \sum_{k=1}^n S^\nu_k\). Now, observe that the time of the most recent arrival up to time \(t\), \(A^\nu_{N^\nu_t}\), can be bounded as

\[
t - S^\nu_{N^\nu_t + 1} \leq A^\nu_{N^\nu_t} \leq t,
\]

since if \(t - S^\nu_{N^\nu_t + 1} > A^\nu_{N^\nu_t}\), then arrival \(N^\nu_t + 1\) would have occurred before time \(t\). Now, we also note that because \(\nu^* > 0\) then \(N^\nu_t \to \infty\) as \(t \to \infty\) and this implies that

\[
\frac{S^\nu_{N^\nu_t + 1}}{N^\nu_t} \xrightarrow{a.s.} 0
\]

as \(t \to \infty\). From Proposition 5.1 and the continuous mapping theorem, we know that \(\frac{t}{N^\nu_t} \to \frac{1}{\nu_\infty}\) and \(\frac{t - S^\nu_{N^\nu_t + 1}}{N^\nu_t} \to \frac{1}{\nu_\infty}\) almost surely. By the sandwiching \(A^\nu_{N^\nu_t}\), this yields the stated result.*
Corollary 5.3. Let \((\lambda_t, N^\lambda_t)\) be the intensity and count of a Hawkes process with baseline intensity \(\lambda^* > 0\), intensity jump rate \(\beta > \alpha\). Similarly, let \((\eta_t, N^\eta_t)\) be the intensity and counting process pair for an Affine Queue-Hawkes process with baseline intensity \(\nu^* > 0\), intensity jump rate \(\alpha > 0\), and rate of exponentially distributed service \(\mu > \alpha\). Then, for \(S^\lambda_k\) and \(S^\eta_k\) as the \(k\)th inter-arrival times for the Hawkes and Affine Queue-Hawkes processes, respectively, we have that

\[
\frac{N^\lambda_t}{t} \overset{a.s.}{\to} \lambda_\infty, \quad \frac{N^\eta_t}{t} \overset{a.s.}{\to} \eta_\infty, \tag{5.3}
\]

and

\[
\frac{1}{n} \sum_{k=1}^n S^\lambda_k \overset{a.s.}{\to} \frac{1}{\lambda_\infty}, \quad \frac{1}{n} \sum_{k=1}^n S^\eta_k \overset{a.s.}{\to} \frac{1}{\eta_\infty}, \tag{5.4}
\]

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

5.2 Baseline Fluid Limit of the Queue-Hawkes

In this subsection and in the sequel, we consider a baseline scaling of the Queue-Hawkes process. 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 5.4. 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 \(n\)

\[
z_n(t) + \frac{C_1}{n} \leq n \cdot \left( e^{\frac{z_n(t)}{n}} - 1 \right) \leq z_n(t) + \frac{C_2}{n}. \tag{5.5}
\]

Proof. The proof follows by 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 5.5. 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 to see, we find that the fluid limit converges to the means of the intensity and queue as given in Proposition 2.2.

Theorem 5.5. For \(n \in \mathbb{Z}\), let the \(n\)th fluid-scaled Queue-Hawkes process \((u_n(t), Q^\nu_n(t))\) be defined such that the baseline intensity is \(nu^*\), 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 \((\frac{u^\nu_n(t)}{n}, \frac{Q^\nu_n(t)}{n})\), the limit of the moment generating function

\[
\tilde{M}^\infty(t, \theta, \theta_Q) \equiv \lim_{n \to \infty} \mathbb{E}\left[ e^{\theta \frac{u^\nu_n(t)}{n} + \theta \frac{Q^\nu_n(t)}{n}} \right], \tag{5.6}
\]

is given by

\[
\tilde{M}^\infty(t, \theta, \theta_Q) = e^{\theta u^\nu E[u^\nu] + \theta Q^\nu E[Q^\nu]}, \tag{5.7}
\]

for all \(t \geq 0\).

Proof. See Appendix A.4.
5.3 Baseline Diffusion Limit of the Queue-Hawkes

To now consider a diffusion limit we will still scale the baseline intensity by $n$, but we now instead scale the process intensity and the queue length by $\frac{1}{\sqrt{n}}$. More specifically, we scale the centered version 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 5.5, the diffusion scaling also involves second order terms. As in the context of the variance bounds we discussed in Subsection 2.3, it is challenging to calculate such quantities for the Queue-Hawkes process. Thus, we will use the same idea from the variance bounds in Proposition 2.5 and bound the quantities above and below via

$$0 \leq \frac{(\nu_t - \nu^*)^2}{Q_t^\nu} \leq \alpha(\nu_t - \nu^*).$$

(5.8)

By doing so, we create upper and lower bounds for the true diffusion limit of the Queue-Hawkes process. 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 $\gamma = 1$ as the upper.

**Theorem 5.6.** For $n \in \mathbb{Z}$, let the $n^{th}$ diffusion-scaled Queue-Hawkes process $(\nu_t(n), Q_t^\nu(n))$ be defined such that the baseline intensity is $n\nu^*$, the intensity jump size is $\alpha > 0$, the intensity decay rate is $\beta > 0$, and the rate of exponentially distributed service is $\mu > 0$, where $\mu + \beta > \alpha$. For the scaled quantities $(\frac{\nu_t(n)}{\sqrt{n}}, \frac{Q_t^\nu(n)}{\sqrt{n}})$, let $\tilde{\mathcal{M}}^\infty(t, \theta^\nu, \theta_Q)$ be defined:

$$\tilde{\mathcal{M}}^\infty(t, \theta^\nu, \theta_Q) \equiv \lim_{n \to \infty} E\left[ \frac{d\nu_t}{\sqrt{n}}(\nu_t(n) - n
\nu^*) + \frac{\theta_Q}{\sqrt{n}}(Q_t^\nu(n) - \frac{\nu}{\nu}) \right].$$

(5.9)

Then for $\beta \neq \alpha$, this is bounded above and below by $B_0 \leq \tilde{\mathcal{M}}^\infty(t, \theta^\nu, \theta_Q) \leq B_1$, where $B_\gamma$ is given by

$$B_\gamma = e^{\nu_0\theta_t e^{-(\mu + \beta - \alpha)t}} + \frac{\nu_0\theta_t}{\beta - \alpha}(e^{-\mu t} - e^{-(\mu + \beta - \alpha)t}) + Q_0\theta_t e^{-\mu t} + \left(\theta^Q - \frac{\theta^Q}{\beta - \alpha}\right)^2 \frac{2(\gamma e^{\nu_\infty - \nu^*})}{2(\beta - \alpha)^2} + \frac{\theta^Q_\infty}{\beta - \alpha} \left(\frac{\gamma e^{\nu_\infty - \nu^*}}{2(\beta - \alpha)^2} \right) + \frac{\theta^Q_\infty}{\beta - \alpha} \left(\frac{\gamma e^{\nu_\infty - \nu^*}}{2(\beta - \alpha)^2} \right),$$

(5.10)

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

$$B_\gamma = e^{\nu_0\theta_t e^{-\mu t} + \nu_0\theta_t e^{-\mu t} + Q_0\theta_t e^{-\mu t} + \left(\frac{\gamma e^{\nu_\infty - \nu^*}}{e^{\mu t} + \theta^Q_\infty} \right) + \frac{\theta^Q_\infty}{\beta - \alpha}} \left(\frac{\gamma e^{\nu_\infty - \nu^*}}{\beta - \alpha} \right) + \frac{\theta^Q_\infty}{\beta - \alpha} \left(\frac{\gamma e^{\nu_\infty - \nu^*}}{\beta - \alpha} \right) + \frac{\theta^Q_\infty}{\beta - \alpha} \left(\frac{\gamma e^{\nu_\infty - \nu^*}}{\beta - \alpha} \right),$$

(5.11)

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

**Proof.** See Appendix A.5.

As a consequence of these diffusion approximations, we can give normally distributed approximations for the steady-state distributions of the Queue-Hawkes intensity and queue length. These are stated below in Corollary 5.7 again in terms of $\gamma$. One can note that the approximate intensity variance in Equation 5.12 matches the upper and lower bounds given in Proposition 2.5 for $\gamma = 1$ and $\gamma = 0$, respectively.
Corollary 5.7. Let \((\nu_t, Q^\nu_t)\) be a Queue-Hawkes process 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 processes \(\nu_t\) and \(Q^\nu_t\) are approximated by the random variables \(X_\nu(\gamma) \sim N(\nu_\infty, \sigma^2_\nu(\gamma))\) and \(X_Q(\gamma) \sim N\left(\frac{\mu}{\mu + \beta}, \sigma^2_Q(\gamma)\right)\), respectively, where

\[
\sigma^2_\nu(\gamma) = \frac{\gamma \alpha \mu (\nu_\infty - \nu^*) + \alpha^2 \nu_\infty}{2(\mu + \beta - \alpha)},
\]

and if \(\beta \neq \alpha\) then

\[
\sigma^2_Q(\gamma) = \frac{\gamma \alpha \mu (\nu_\infty - \nu^*) + \alpha^2 \nu_\infty}{2(\beta - \alpha)(\mu + \beta - \alpha)} - \frac{2\gamma \alpha \mu + 2\mu (\beta - \alpha)}{(\beta - \alpha)^2(2\mu + \beta - \alpha)} (\nu_\infty - \nu^*) + 2\alpha \beta \nu_\infty
\]

\[
+ \frac{\gamma \alpha \mu (\nu_\infty - \nu^*) + \nu_\infty \beta^2}{2\mu(\beta - \alpha)^2} + \frac{\nu_\infty - \nu^*}{\beta - \alpha} + \frac{\nu_\infty}{2\mu},
\]

whereas if \(\beta = \alpha\) then

\[
\sigma^2_Q(\gamma) = \left(1 + \frac{\gamma \alpha}{4\mu^2}\right) (\nu_\infty - \nu^*) + \left(1 + \frac{\alpha}{2\mu^2} + \frac{\alpha^2}{4\mu^3}\right) \nu_\infty,
\]

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

In Figures 5.1 and 5.2 we plot the simulated steady-state distributions of a Queue-Hawkes process 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 approximations of the terms in Equation 5.8:

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

In Figure 5.1 the baseline intensity is equal to 100, whereas in Figure 5.2 it is 1,000. While there are known limitations of Gaussian approximations for queueing processes such as is discussed in Massey and Pender [53], we see that these approximations appear to be quite close, particularly so 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.

Figure 5.1: Histogram comparing the simulated steady-state intensity (left) and queue (right) to their diffusion approximations evaluated at multiple values of \(\gamma\), where \(\nu^* = 100\), \(\alpha = 3\), \(\beta = 2\), and \(\mu = 2\).
6 Conclusion and Final Remarks

In this paper we have considered ephemerally self-exciting processes. By uniting the dynamics of a Hawkes process and an infinite server queue, we defined the Queue-Hawkes process, a new generalization of the Hawkes process in which arriving entities increase the rate of future arrivals only as long as they are in the system. We identified two notable special cases of the Queue-Hawkes process: the Hawkes process, which corresponds to having no service in the Queue-Hawkes process, and the Affine-Queue Hawkes process, which is the Queue-Hawkes process with no decay. These three processes constitute the core of what we have studied in this paper, and the relationships between the models are summarized below in Figure 6.1.

\[
\text{Queue-Hawkes} \quad \beta = 0 \quad \mu = 0 \\
\text{Affine QH} \quad \text{Batch scaling} \quad \text{Hawkes}
\]

Figure 6.1: Relating the Queue-Hawkes, Affine Queue-Hawkes, and Hawkes processes

Our analysis for these three models includes both exploration of the models individually, such as the limiting results for the Queue-Hawkes process in Section 5, and comparisons between the processes, such as the construction of the Hawkes process formed from the batch scaling of the Affine Queue-Hawkes process shown in Section 3. We have also made several connections from these self-exciting processes to other well known stochastic models including branching processes, random walks, epidemics, and Bayesian mixture models, as was discussed in Section 4.

For future work, modern uses of stochastic models prioritize the investigation of multidimensional, non-Markovian, and marked versions of these processes, similar to what we considered in Subsection 3.4. We are quite interested in pursuing these generalizations, but because the present subject has already led to a lengthy analysis we instead will study them in subsequent research. Additionally, while we have studied a finite server variant of the Affine Queue-Hawkes in the blocking model in Subsection 3.2, we are also interested in studying other types of finite server scenarios in the future. One approach for doing so could be to have a rate of abandonment for the entities that are awaiting service, in which we could employ techniques from
Similarly, one can consider how this process changes when dealing with systems with finite populations, such as in the transitory queue models given in Honnappa et al. [41] or the SIR-Hawkes process in Rizoiu et al. [67]. Additionally, one could pursue queueing models driven by arrivals from a Queue-Hawkes process; in this case one could potentially use the heavy traffic limits in Pang and Whitt [59, 60] to aid analysis. One could also consider a new generalization of self-excitement based on delayed information, such as in the queueing works Pender et al. [61, 63, 62]. For future lines of theoretical work, we are quite interested in further exploring the batch scaling in Theorem 3.10. We can note that there are similar results connecting batch arrival, Poisson driven, infinite server queues to Poisson shot-noise processes in the works de Graaf et al. [25], Daw and Pender [24]. We are interested in adapting and extending these ideas to other models in self-excitement and forming comparisons between self-exciting models and externally excited models like the shot-noise process, which has recently been used in queueing and service system contexts in e.g. Oreshkin et al. [57], L’Ecuyer et al. [49], Boxma et al. [12], Koops et al. [46]. Furthermore, the problem of distinguishing self-excited data from externally excited data is both an open and intriguing one, and so in future work we will explore what aid the batch scaling constructions of these processes can provide in resolving this question.

Acknowledgements

We acknowledge the generous support of the National Science Foundation (NSF) for Andrew Daw’s Graduate Research Fellowship under grant DGE-1650441. Additionally, we are grateful for helpful discussions with Robert Hampshire at the University of Michigan and Emily Fischer at Cornell University, particularly in regards to the contents of Section 4.

References

[1] Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In Proceedings of the 2008 SIAM International Conference on Data Mining, pages 219–230. SIAM, 2008.

[2] David J Aldous. Exchangeability and related topics. In École d’Été de Probabilités de Saint-Flour XIII1983, pages 1–198. Springer, 1985.

[3] Shahriar Azizpour, Kay Giesecke, and Gustavo Schwenkler. Exploring the sources of default clustering. Journal of Financial Economics, 2016.

[4] Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. Quantitative Finance, 14(7):1147–1166, 2014.

[5] Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-François Muzy. Some limit theorems for Hawkes processes and application to financial statistics. Stochastic Processes and their Applications, 123(7):2475–2499, 2013.

[6] Emmanuel Bacry, Thibault Jaisson, and Jean-François Muzy. Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. Quantitative Finance, 16(8):1179–1201, 2016.

[7] Frank Ball. The threshold behaviour of epidemic models. Journal of Applied Probability, 20 (2):227–241, 1983.
[8] David Blackwell. A renewal theorem. *Duke Mathematical Journal*, 15(1):145–150, 1948.

[9] David Blackwell, James B MacQueen, et al. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[10] Pierre Blanc, Jonathan Donier, and J-P Bouchaud. Quadratic Hawkes processes for financial prices. *Quantitative Finance*, 17(2):171–188, 2017.

[11] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(Aug):2461–2488, 2011.

[12] Onno Boxma, Offer Kella, and Michel Mandjes. Infinite-server systems with Coxian arrivals. Working paper. URL https://scholars.huji.ac.il/sites/default/files/offerkella/files/oom1905181.pdf.

[13] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

[14] E Brockmeyer, HL Halstrom, and Arne Jensen. *The life and works of AK Erlang*. Copenhagen: Copenhagen Telephone Co, 1948.

[15] José Da Fonseca and Riadh Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

[16] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[17] Angelos Dassios and Hongbiao Zhao. A dynamic contagion process. *Advances in applied probability*, 43(3):814–846, 2011.

[18] Angelos Dassios and Hongbiao Zhao. Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18, 2013.

[19] Angelos Dassios and Hongbiao Zhao. Efficient simulation of clustering jumps with CIR intensity. *Operations Research*, 65(6):1494–1515, 2017.

[20] Mark HA Davis. Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 353–388, 1984.

[21] Andrew Daw and Jamol Pender. Queues driven by Hawkes processes. *Stochastic Systems*, 8(3):192–229, 2018.

[22] Andrew Daw and Jamol Pender. Poster: Exact simulation of the Queue-Hawkes process. In *2018 Winter Simulation Conference (WSC)*, pages 4234–4235. IEEE, 2018.

[23] Andrew Daw and Jamol Pender. New perspectives on the Erlang-A queue. *Advances in Applied Probability (to appear)*, 2019.

[24] Andrew Daw and Jamol Pender. On the distributions of infinite server queues with batch arrivals. *Queueing Systems (to appear)*, 2019.

[25] WF de Graaf, Werner RW Scheinhardt, and Richard J Boucherie. Shot-noise fluid queues and infinite-server systems with batch arrivals. *Performance evaluation*, 116:143–155, 2017.
[26] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song. Dirichlet-Hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 219–228. ACM, 2015.

[27] Stephen G Eick, William A Massey, and Ward Whitt. The physics of the \( M_t/G/\infty \) queue. *Operations Research*, 41(4):731–742, 1993.

[28] Eymen Errais, Kay Giesecke, and Lisa R Goldberg. Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1(1):642–665, 2010.

[29] Şeyda Ertekin, Cynthia Rudin, Tyler H McCormick, et al. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9(1):122–144, 2015.

[30] Mehrdad Farajtabar, Jiachen Yang, Xiaoqing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1097–1106. JMLR. org, 2017.

[31] William Feller. *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, 2 edition, 1957.

[32] Xuefeng Gao and Lingjiong Zhu. Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems*, pages 1–46.

[33] Xuefeng Gao and Lingjiong Zhu. Large deviations and applications for Markovian Hawkes processes with a large initial intensity. *Bernoulli*, 24(4A):2875–2905, 2018.

[34] Xuefeng Gao, Xiang Zhou, and Lingjiong Zhu. Transform analysis for Hawkes processes with applications in dark pool trading. *Quantitative Finance*, 18(2):265–282, 2018.

[35] Xin Guo, Zhao Ruan, and Lingjiong Zhu. Dynamics of order positions and related queues in a limit order book. *arXiv preprint arXiv:1505.04810*, 2015.

[36] Amrita Gupta, Mehrdad Farajtabar, Bistra Dilkina, and Hongyuan Zha. Discrete interventions in Hawkes processes with applications in invasive species management. In *IJCAI*, pages 3385–3392, 2018.

[37] Jack K Hale and Sjoerd M Verduyn Lunel. *Introduction to functional differential equations*, volume 99. Springer Science & Business Media, 2013.

[38] Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 438–443, 1971.

[39] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[40] Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974. doi:10.2307/3212693.

[41] Harsha Honnappa, Rahul Jain, and Amy R Ward. On transitory queueing. *arXiv preprint arXiv:1412.2321*, 2014.
[42] Rouba Ibrahim, Han Ye, Pierre L'Ecuyer, and Haipeng Shen. Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3):865–874, 2016.

[43] Samuel Karlin and James McGregor. The classification of birth and death processes. *Transactions of the American Mathematical Society*, 86(2):366–400, 1957.

[44] Dimitris Karlis and Evdokia Xekalaki. Mixed poisson distributions. *International Statistical Review*, 73(1):35–58, 2005.

[45] Frank P. Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[46] David T Koops, Onno J Boxma, and MRH Mandjes. Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems*, 86(3-4):301–325, 2017.

[47] DT Koops, Mayank Saxena, OJ Boxma, and Michel Mandjes. Infinite-server queues with Hawkes input. *Journal of Applied Probability*, 55(3):920–943, 2018.

[48] Michael Krumin, Inna Reutsky, and Shy Shoham. Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Frontiers in computational neuroscience*, 4:147, 2010.

[49] Pierre L’Ecuyer, Klas Gustavsson, and Leif Olsson. Modeling bursts in the arrival process to an emergency call center. In *2018 Winter Simulation Conference (WSC)*, pages 525–536. IEEE, 2018.

[50] Liangda Li and Hongyuan Zha. Energy usage behavior modeling in energy disaggregation via Hawkes processes. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(3):36, 2018.

[51] Torgny Lindvall. A probabilistic proof of Blackwell’s renewal theorem. *The Annals of Probability*, 5(3):482–485, 1977.

[52] James O Lloyd-Smith. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PloS one*, 2(2):e180, 2007.

[53] William A Massey and Jamol Pender. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75(2-4):243–277, 2013.

[54] Hongyuan Mei and Jason M Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.

[55] A Mukherjea, M Rao, and S Suen. A note on moment generating functions. *Statistics & probability letters*, 76(11):1185–1189, 2006.

[56] Yoshihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[57] Boris N Oreshkin, Nazim Réegnard, and Pierre L’Ecuyer. Rate-based daily arrival process models with application to call centers. *Operations Research*, 64(2):510–527, 2016.
[58] Richard Otter. The multiplicative process. *The Annals of Mathematical Statistics*, pages 206–224, 1949.

[59] Guodong Pang and Ward Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65(4):325–364, 2010.

[60] Guodong Pang and Ward Whitt. Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*, 73(2):119–146, 2013.

[61] Jamol Pender, Richard H Rand, and Elizabeth Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 27(04):1730016, 2017.

[62] Jamol Pender, Richard H Rand, and Elizabeth Wesson. Strong approximations for queues with customer choice and constant delays. 2017.

[63] Jamol Pender, Richard H Rand, and Elizabeth Wesson. An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynamics*, 91(4):2411–2427, 2018.

[64] Marcello Rambaldi, Emmanuel Bacry, and Fabrizio Lillo. The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance*, 17(7):999–1020, 2017.

[65] Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. Hawkes processes for events in social media. In *Frontiers of Multimedia Research*, pages 191–218. Association for Computing Machinery and Morgan & Claypool, 2017.

[66] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be HIP: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pages 735–744. International World Wide Web Conferences Steering Committee, 2017.

[67] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 419–428. International World Wide Web Conferences Steering Committee, 2018.

[68] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[69] Sarabjeet Singh and Christopher R Myers. Outbreak statistics and scaling laws for externally driven epidemics. *Physical Review E*, 89(4):042108, 2014.

[70] Remco Van der Hofstad and Michael Keane. An elementary proof of the hitting time theorem. *The American Mathematical Monthly*, 115(8):753–756, 2008.

[71] Erik A Van Doorn. On the $\alpha$-classification of birth-death and quasi-birth-death processes. *Stochastic models*, 22(3):411–421, 2006.

[72] Gord Willmot. Mixed compound poisson distributions. *ASTIN Bulletin: The Journal of the IAA*, 16(S1):S59–S79, 1986.

[73] Ronald W Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.
[74] Peng Wu, Marcello Rambaldi, Jean-François Muzy, and Emmanuel Bacry. Queue-reactive Hawkes models for the order flow. arXiv preprint arXiv:1901.08938, 2019.

[75] Hongteng Xu, Yi Zhen, and Hongyuan Zha. Trailer generation via a point process-based visual attractiveness model. In Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[76] Hongteng Xu, Dixin Luo, and Hongyuan Zha. Learning Hawkes processes from short doubly-censored event sequences. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3831–3840. JMLR. org, 2017.

[77] Junchi Yan, Yu Wang, Ke Zhou, Jin Huang, Chunhua Tian, Hongyuan Zha, and Weishan Dong. Towards effective prioritizing water pipe replacement and rehabilitation. In Twenty-Third International Joint Conference on Artificial Intelligence, 2013.

[78] Xiao-Wei Zhang, Peter W Glynn, Kay Giesecke, and Jose Blanchet. Rare event simulation for a generalized Hawkes process. In Proceedings of the 2009 Winter Simulation Conference (WSC), pages 1291–1298. IEEE, 2009.

[79] Xiaowei Zhang, Jose Blanchet, Kay Giesecke, and Peter W Glynn. Affine point processes: approximation and efficient simulation. Mathematics of Operations Research, 40(4):797–819, 2015.

[80] Lorenzo Zino, Alessandro Rizzo, and Maurizio Porfiri. Modeling memory effects in activity-driven networks. SIAM Journal on Applied Dynamical Systems, 17(4):2830–2854, 2018.

A Appendices

A.1 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 Queue-Hawkes process is Markov, and it supports much of our analysis throughout this work.

Lemma A.1. For a sufficiently regular function $f : (\mathbb{R}^+ \times \mathbb{N} \times \mathbb{N}) \rightarrow \mathbb{R}$, the generator of the Queue-Hawkes process is given by

$$
\mathcal{L}f(\nu_t, Q_t, N_t) = \beta(\nu^* - \nu_t) \frac{\partial f(\nu_t, Q_t, N_t)}{\partial \nu_t} + \sum_{i=1}^{n} \nu_t (f(\nu_t + \alpha, Q_t + 1, N_t + 1) - f(\nu_t, Q_t, N_t))
$$

$$
+ \mu Q_t \left( f \left( \frac{\nu_t - \nu^*}{Q_t}, Q_t - 1, N_t \right) - f(\nu_t, Q_t, N_t) \right) \tag{A.1}
$$

Excitation Decay

Arrivals

Departures

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

$$
\frac{d}{dt} \mathbb{E} [f(\nu_t, Q_t, N_t)] = \mathbb{E} [\mathcal{L}f(\nu_t, Q_t, N_t)] \tag{A.2}
$$

for all $t \geq 0$. 

49
This is a direct result of the Queue-Hawkes process belonging to the family of piecewise deterministic Markov processes, as defined in Davis [20]. Moreover, the specific regularity conditions are given in Theorem 5.5 of that work.

Throughout this work we make comparisons between different processes, in particular between the Queue-Hawkes process, the Hawkes process, and the Affine Queue-Hawkes process. One way that we do this is to investigate the differential equations found with use of Lemma A.1. In Lemma A.2 we provide the method by which we make such comparisons.

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

\[
\frac{dx(t)}{dt} = f(t, x(t)), \quad x(0) = x_0
\]

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

\[
\frac{dy(t)}{dt} \leq f(t, y(t)) \quad \text{for } t \in [0, T] \text{ and } y(0) \leq x_0
\]

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

**Proof.** The the proof of this result is given in Hale and Lunel [37].

We now give a result in Proposition A.3 regarding the size of the down-jumps. This observation is similar to the discussions of Section 2 in that it is a direct observation of the Queue-Hawkes process. However, because it is not of use in any of our later analysis we provide it here.

**Proposition A.3.** Let \( \phi_t = \frac{\mu - \nu^*}{Q_t} \) be the size of a down-jump occurring at time \( t \geq 0 \). Suppose that \( b \geq a \geq 0 \) is such that \( Q_t \) is positive for all \( t \in [a, b] \). Then, the \( \phi_t \) has no downward jumps on \([a, b] \).

**Proof.** Suppose that \([a, b] \) is such as interval, and then for \( t \in [a, b] \) we note that

\[
\frac{\nu_t - \frac{\mu - \nu^*}{Q_t} - \nu^*}{Q_t - 1} = \frac{Q_t \nu_t - \nu_t + \nu^* - \frac{Q_t \nu^*}{Q_t - 1}}{Q_t} = \frac{\nu_t - \nu^*}{Q_t},
\]

and this is equal to \( \phi_t \).

**Proposition A.4.** Let \((\nu_t, Q_t^\nu, N_t^\nu)\) be a Queue-Hawkes process with baseline intensity \( \nu^* > 0 \), intensity jump \( \alpha > 0 \), decay rate \( \beta \geq 0 \), and service rate \( \mu \geq 0 \), where \( \mu + \beta > \alpha \). Then for all \( t \geq 0 \), the variance of the intensity satisfies

\[
\overline{\psi}_1(t) \leq \text{Var}(\nu_t) \leq \underline{\psi}_1(t),
\]

where \( \overline{\psi}_1(\cdot) \) is defined

\[
\overline{\psi}_1(t) = \frac{\alpha^2 \nu_\infty + \alpha \mu (\nu_\infty - \nu^*)}{2(\mu + \beta - \alpha)} (1 - e^{-2(\mu + \beta - \alpha)t}) + \frac{\alpha^2 + \alpha \mu}{\mu + \beta - \alpha} (\nu_0 - \nu_\infty) (e^{-(\mu + \beta - \alpha)t} - e^{-2(\mu + \beta - \alpha)t}) + 2(\nu_0 - \nu_\infty) \nu_\infty e^{-2(\mu + \beta - \alpha)t},
\]

and \( \underline{\psi}_1(\cdot) \) is such that

\[
\underline{\psi}_1(t) = \frac{\alpha^2 \nu_\infty}{2(\mu + \beta - \alpha)} (1 - e^{-2(\mu + \beta - \alpha)t}) + \frac{\alpha^2 (\nu_0 - \nu_\infty)}{\mu + \beta - \alpha} (e^{-(\mu + \beta - \alpha)t} - e^{-2(\mu + \beta - \alpha)t}) + 2(\nu_0 - \nu_\infty) \nu_\infty e^{-2(\mu + \beta - \alpha)t},
\]
with \( \nu_\infty = \frac{(\mu+\beta)\nu^*}{\mu+\beta-\alpha} \). Then, for the functions \( \psi_2(t,f(\cdot)) \) and \( \psi_3(t,f(\cdot)) \) defined as

\[
\psi_2(t,f(\cdot)) = (\mu + \beta)\nu^* e^{-(2\mu+\beta-\alpha)t} \int_0^t e^{(2\mu+\beta-\alpha)s} E[Q^\nu_t] ds - \frac{\mu\nu^*}{2\mu+\beta-\alpha} \left(1 - e^{-(2\mu+\beta-\alpha)t}\right) - E[\nu_t]E[Q^\nu_t] + (\mu + \alpha) e^{-(2\mu+\beta-\alpha)t} \int_0^t e^{(2\mu+\beta-\alpha)s} E[\nu_s] ds + e^{-(2\mu+\beta-\alpha)t} \int_0^t e^{(2\mu+\beta-\alpha)s} \left(f(s) + E[\nu_s]^2\right) ds + \nu_0 Q_0 e^{-(2\mu+\beta-\alpha)t},
\]

\[
\psi_3(t,f(\cdot)) = (\mu + \beta)\nu^* e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} E[N^\nu_s] ds + \alpha e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} E[N^\nu_s] ds - E[\nu_t]E[N^\nu_t] + e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} \left(f(s) + E[\nu_s]^2\right) ds,
\]

the covariance of the intensity and the queue and of the intensity and the counting process are bounded above and below by

\[
\psi_2(t,\psi_1(\cdot)) \leq \text{Cov}[\nu_t, Q^\nu_t] \leq \psi_2(t,\psi_1(\cdot)) \quad \text{and} \quad \psi_3(t,\psi_1(\cdot)) \leq \text{Cov}[\nu_t, N^\nu_t] \leq \psi_3(t,\psi_1(\cdot)),
\]

for all \( t \geq 0 \). Finally, the variances of the queue and the counting process are such that

\[
\psi_4(t,\psi_1(\cdot)) \leq \text{Var}(Q^\nu_t) \leq \psi_4(t,\psi_1(\cdot)) \quad \text{and} \quad \psi_5(t,\psi_1(\cdot)) \leq \text{Var}(N^\nu_t) \leq \psi_5(t,\psi_1(\cdot)),
\]

where \( \psi_4(t,f(\cdot)) \) and \( \psi_5(t,f(\cdot)) \) are defined as

\[
\psi_4(t,f(\cdot)) = \nu_0 Q_0 e^{-2\mu t} + e^{-2\mu t} \int_0^t e^{2\mu s} E[\nu_s] ds + \mu e^{-2\mu t} \int_0^t e^{2\mu s} E[Q^\nu_s] ds - E[Q^\nu_t] + 2\mu e^{-2\mu t} \int_0^t e^{2\mu s} (\psi_2(s,f(\cdot)) + E[\nu_s]E[Q^\nu_s]) ds,
\]

\[
\psi_5(t,f(\cdot)) = \nu_\infty t + \frac{\nu_0 - \nu_\infty}{\mu + \beta - \alpha} \left(1 - e^{-(\mu+\beta-\alpha)t}\right) - \frac{\nu_0 - \nu_\infty}{\mu + \beta - \alpha} \left(1 - 2e^{-(\mu+\beta-\alpha)t} + e^{-2(\mu+\beta-\alpha)t}\right) + 2 \int_0^t (\psi_3(s,f(\cdot)) + E[\nu_s]E[N^\nu_s]) ds,
\]

for all \( t \geq 0 \).

**Proof.** We know from Equation 2.1 that the time derivative of the second moment of the Queue-Hawkes intensity is

\[
\frac{d}{dt} E[\nu^2_t] = (\alpha^2 + 2(\mu + \beta)\nu^*) E[\nu_t] - 2(\mu + \beta - \alpha) E[\nu^2_t] + \mu E \left[ Q^\nu_t \left( \frac{\nu_t - \nu^*}{Q^\nu_t} \right)^2 \right].
\]

By recalling that the down-jump size \( \frac{\nu_0 - \nu_\infty}{Q^\nu_t} \) is no greater than the up-jump size \( \alpha \), we can observe that

\[
Q^\nu_t \left( \frac{\nu_t - \nu^*}{Q^\nu_t} \right)^2 = (\nu_t - \nu^*) \left( \frac{\nu_t - \nu^*}{Q^\nu_t} \right) \leq \alpha (\nu_t - \nu^*),
\]

and so we can bound the time derivative of \( E[\nu^2_t] \) by

\[
\frac{d}{dt} E[\nu^2_t] \leq (\alpha^2 + \alpha \mu + 2(\mu + \beta)\nu^*) E[\nu_t] - 2(\mu + \beta - \alpha) E[\nu^2_t] - \alpha \mu \nu^* = (\alpha^2 + \alpha \mu + 2(\mu + \beta)\nu^*) \left( \nu_\infty + (\nu_0 - \nu_\infty) e^{-(\mu+\beta-\alpha)t} \right) - 2(\mu + \beta - \alpha) E[\nu^2_t] - \alpha \mu \nu^*.
\]
By solving the ordinary differential equation given by this upper bound and initial value $\nu_0^2$, we can use Lemma A.2 to see that

$$E [\nu_t^2] \leq \nu_0^2 e^{-2(\mu+\beta-\alpha)t} + \left( \frac{\alpha^2 + \alpha \mu}{\mu + \beta - \alpha} + 2\nu_\infty \right) (\nu_0 - \nu_\infty) \left( e^{-(\mu+\beta-\alpha)t} - e^{-2(\mu+\beta-\alpha)t} \right)$$

$$+ \left( \frac{(\alpha^2 + \alpha \mu)\nu_\infty - \alpha \mu \nu^*}{2(\mu + \beta - \alpha)} + \nu_\infty \right) \left( 1 - e^{-2(\mu+\beta-\alpha)t} \right),$$

for all $t \geq 0$. By subtracting the square of the mean from this expression, we achieve the stated result, as the left-hand side of this inequality is equal to $\psi_1(t) + E [\nu_t^2]$. We can now again use Lemma A.1 to note that

$$\frac{d}{dt} E \left[ (N_t^\nu)^2 \right] = E [\nu_t] + 2E [\nu_t N_t^\nu],$$

$$\frac{d}{dt} E [\nu_t N_t^\nu] = (\mu + \beta)\nu^* E [N_t^\nu] - (\mu + \beta - \alpha)E [\nu_t N_t^\nu] + \alpha E [\nu_t] + E [\nu_t^2],$$

$$\frac{d}{dt} E \left[ (Q_t^\nu)^2 \right] = E [\nu_t] + \mu E [Q_t^\nu] + 2E [\nu_t Q_t^\nu] - 2\mu E \left[ (Q_t^\nu)^2 \right],$$

$$\frac{d}{dt} E [\nu_t Q_t^\nu] = (\mu + \beta)\nu^* E [Q_t^\nu] - \mu \nu^* + (\mu + \alpha)E [\nu_t] + E [\nu_t^2] - (2\mu + \beta - \alpha)E [\nu_t Q_t^\nu],$$

and we can bound the equations in this system through the upper bound we have found for $E [\nu_t^2]$. Using this new system and the initial values implied by assuming the counting process starts at 0 and the queue starts at $Q_0$, we apply Lemma A.2, which yields

$$E [\nu_t N_t^\nu] \leq (\mu + \beta)\nu^* e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} E [N_s] ds + \alpha e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} E [\nu_s] ds$$

$$+ e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} \left( \psi_3(s) + E [\nu_s]^2 \right) ds,$$

$$E [\nu_t Q_t^\nu] \leq (\mu + \beta)\nu^* e^{-(\mu+\beta-\alpha)t} \int_0^t e^{(\mu+\beta-\alpha)s} E [Q_s] ds - \frac{\mu \nu^*}{2(\mu + \beta - \alpha)} \left( 1 - e^{-(2\mu + \beta - \alpha)t} \right)$$

$$+ (\mu + \alpha) e^{-(2\mu + \beta - \alpha)t} \int_0^t e^{(2\mu + \beta - \alpha)s} E [\nu_s] ds + e^{-(2\mu + \beta - \alpha)t} \int_0^t e^{(2\mu + \beta - \alpha)s} \left( \psi_3(s) + E [\nu_s]^2 \right) ds$$

$$+ \nu_0 Q_0 e^{-(2\mu + \beta - \alpha)t},$$

where the left-hand side in the latter inequality is $\psi_2(t, \overline{\psi}_1(\cdot)) + E [\nu_t] E [Q_t^\nu]$ while that of the former is $\psi_3(t, \overline{\psi}_3(\cdot)) + E [\nu_t] E [N_t^\nu]$. Thus, we then bound the second moments by

$$E \left[ (N_t^\nu)^2 \right] \leq \int_0^t E [\nu_s] ds + 2 \int_0^t \left( \psi_3(s, \overline{\psi}_3(\cdot)) + E [\nu_s] E [N_s] \right) ds,$$

$$E \left[ (Q_t^\nu)^2 \right] \leq \nu_0 Q_0 e^{-2\mu t} + e^{-2\mu t} \int_0^t e^{2\mu s} E [\nu_s] ds + \mu e^{-2\mu t} \int_0^t e^{2\mu s} E [Q_s] ds$$

$$+ 2\mu e^{-2\mu t} \int_0^t e^{2\mu s} \left( \psi_2(s, \overline{\psi}_2(\cdot)) + E [\nu_s] E [Q_s] \right) ds,$$

and this completes the proof of the upper bounds. The lower bounds follow by similar arguments in which $E [\nu_t^2]$ is instead bounded below by

$$\frac{d}{dt} E [\nu_t^2] \geq (\alpha^2 + 2(\mu + \beta)\nu^*) E [\nu_t] - 2(\mu + \beta - \alpha)E [\nu_t^2]$$

$$= (\alpha^2 + 2(\mu + \beta)\nu^*) \left( \nu_\infty + (\nu_0 - \nu_\infty) e^{-(\mu+\beta-\alpha)t} \right) - 2(\mu + \beta - \alpha)E [\nu_t^2],$$

52
where in this case we have instead bounded the quantity $Q_t^\nu\left(\frac{\nu - \nu^*}{Q_t^\nu}\right)^2$ below by 0. Using Lemma A.2, we have

$$
E[\nu_t^2] \geq \frac{\alpha^2(\nu_0 - \nu_\infty)}{\mu + \beta - \alpha} \left(e^{-(\mu + \beta - \alpha)t} - e^{-2(\mu + \beta - \alpha)t}\right) + \left(\frac{\alpha^2\nu_\infty}{2(\mu + \beta - \alpha)} + \nu_\infty^2\right) \left(1 - e^{-2(\mu + \beta - \alpha)t}\right) + \nu_\infty^2 e^{-2(\mu + \beta - \alpha)t}.
$$

This now yields $\psi_t(t) + E[\nu_t]^2$, and thus the remainder of the proof follows analogously.

As another auxiliary result, in Proposition A.5 we give the probability generating function for the number in system and the number of departures in the Affine Queue-Hawkes process. The departure process is largely outside of the scope of this work, but this result is instrumental in the proof of the probability generating function for the counting process in Proposition 3.7, which is given in Section A.2.

**Proposition A.5.** Let $Q_t^\nu$ be the queue length of an Affine Queue-Hawkes process with baseline intensity $\nu^* > 0$, intensity jump size $\alpha > 0$, and exponential service rate $\mu > \alpha$. Then, let $D_t^\nu$ be the number of entities that have departed from the queue by time $t$. Then, the joint probability generating function of $Q_t^\nu$ and $D_t^\nu$ $G(z_1, z_2, t) \equiv E\left[\frac{Q_t^\nu}{z_1} \frac{D_t^\nu}{z_2}\right]$ is given by

$$
G(z_1, z_2, t) = \frac{D_0}{z^2} e^{\frac{\nu^*(\mu - \alpha)}{2\alpha} t} \left(1 - \left(\frac{\mu + \alpha}{2} - 4\alpha\mu z_2 + \tanh^{-1}\left(\frac{\sqrt{\mu + \alpha} - 2\alpha z_1}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z_2}\right)\right)^2\right) \left(\frac{\mu + \alpha - 2\alpha z_1}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z_2}\right) \left(\frac{\mu + \alpha - 2\alpha z_1}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z_2}\right) \left(\frac{2\alpha z_1 - \mu - \alpha}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z_2}\right) \left(\frac{2\alpha z_1 - \mu - \alpha}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z_2}\right) \left(\frac{\nu^*}{E}\right)^{Q_0}.
$$

where $Q_0$ and $D_0$ are the 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 A.1 that

$$
\frac{d}{dt} E\left[\frac{Q_t^\nu}{z_1} \frac{D_t^\nu}{z_2}\right] = E\left[\nu^* + \alpha Q_t^\nu(z_1 - 1)z_1^2 + \mu Q_t^\nu(z_1 - 1)^2 + \mu Q_t^\nu(z_1 - z_2)\right],
$$

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

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

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

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

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

\[
\begin{align*}
\frac{dz_1}{ds}(r, s) &= \alpha(z_1 - z_1^2) + \mu(z_1 - z_2), & z_1(r, 0) = r \\
\frac{dt}{ds}(r, s) &= 1, & t(r, 0) = 0 \\
\frac{dQ_0}{ds}(r, s) &= \nu^*(z_1 - 1), & Q_0(r, 0) = \log(r^Q_{Q_0} z_2^D_{D_0}).
\end{align*}
\]
Solving the first two equations we see that
\[
z_1(r, s) = \frac{\mu + \alpha}{2\alpha} + \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2} \tanh \left( \frac{s}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2} - \tanh^{-1} \left( \frac{\mu + \alpha - 2\alpha r}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2}} \right) \right)
\]
\[
t(r, s) = s,
\]
which allows us to now solve for \( g(r, s) \). Using the solution to \( z_1(r, s) \), the ordinary differential equation for \( g(r, s) \) is given by
\[
\frac{dg}{ds}(r, s) = \frac{\nu^* \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2}}{2\alpha} \tanh \left( \frac{s}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2} - \tanh^{-1} \left( \frac{\mu + \alpha - 2\alpha r}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2}} \right) \right) + \frac{\nu^*(\mu - \alpha)}{2\alpha},
\]
which yields a solution of
\[
g(r, s) = \log(nQ_0 + D_0) + \frac{\nu^*(\mu - \alpha)}{2\alpha} s + \frac{\nu^*}{2\alpha} \log \left( 1 - \frac{(\mu + \alpha - 2\alpha r)^2}{(\mu + \alpha)^2 - 4\alpha \mu z_2} \right) + \frac{\nu^*}{\alpha} \log \left( \cosh \left( \frac{s}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2} - \tanh^{-1} \left( \frac{\mu + \alpha - 2\alpha r}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2}} \right) \right) \right).
\]
Now, from these equations we can express the characteristics variables in terms of the original arguments as \( s = t \) and
\[
r = \frac{\mu + \alpha}{2\alpha} - \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2} \tanh \left( \frac{t}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z_2} - \tanh^{-1} \left( \frac{2\alpha z_1 - \mu - \alpha}{2\alpha - 2\alpha \mu} \right) \right).
\]
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 A.6 we give the steady-state moment generating function for the batch scaled Affine Queue-Hawkes process with batch size \( n = 2 \)

**Proposition A.6.** Let the 2nd batch-scaled Affine Queue-Hawkes process be defined as follows: \( \nu_1(2) = \nu^* + \alpha Q^0_1(2) \) where \( Q^0_1(2) \) is such that arrivals occur in batches of size \( n \) and each depart after i.i.d. exponential service with rate \( \mu > \alpha > 0 \). Then, the steady-state moment generating function of \( Q^0_1(2) \) is given by
\[
E \left[ e^{\theta Q^0_1(2)} \right] = \lim_{t \to \infty} E \left[ e^{\theta Q^1_1(t)} \right] = \exp \left( \frac{2\nu^*}{\sqrt{\alpha(\alpha + 8\mu)}} \tanh^{-1} \left( \frac{2e^\theta + 1}{\alpha} \sqrt{\frac{\alpha}{\alpha + 8\mu}} \right) + \tanh^{-1} \left( \frac{3\sqrt{\frac{\alpha}{\alpha + 8\mu}}}{\alpha + 8\mu} \right) \right) \left( \frac{2\theta - 2\alpha}{2\mu - 2\alpha (e^{\theta e^{2\theta}})} \right)^{\frac{\alpha + 8\mu}{\alpha}} \tag{A.6}
\]

**Proof.** Using Lemma A.1, we see that the moment generating function will be given by the solution to
\[
\frac{d}{dt} E \left[ e^{\theta Q^1_1(t)} \right] = E \left[ \left( \nu^* + \frac{\alpha Q^1_1(t)}{2} \right) \left( e^{\theta (Q^1_1(t) + 2)} - e^{\theta Q^1_1(t)} \right) \right] + \mu Q^0_1(t) \left( e^{\theta (Q^1_1(t) - 1)} - e^{\theta Q^1_1(t)} \right).\]
which can be equivalently expressed in PDE form as
\[
\frac{\partial}{\partial t} M_2(\theta, t) = \nu^* \left( e^{2\theta} - 1 \right) M_2(\theta, t) + \left( \frac{\alpha}{2} \left( e^{2\theta} - 1 \right) + \mu \left( e^{-\theta} - 1 \right) \right) \frac{\partial}{\partial \theta} M_2(\theta, t),
\]
where \( M_2(\theta, t) = E \left[ e^{\theta Q_t(2)} \right] \). To solve for the steady-state moment generating function we consider the ODE given by
\[
\frac{d}{d\theta} M_2(\theta, \infty) = \frac{\nu^* \left( 1 - e^{2\theta} \right) M_2(\theta, \infty)}{\frac{\theta}{2} \left( e^{2\theta} - 1 \right) + \mu \left( e^{-\theta} - 1 \right)},
\]
with the initial condition that \( M_2(0, \infty) = 1 \). Through taking the derivative of the expression in Equation A.6, we verify the result. \( \square \)

### A.2 Proof of Proposition 3.7

**Proof.** Using Proposition A.5, we proceed through use of exponential identities for the hyperbolic functions. Specifically, we will make use of the following:

\[
\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad (A.7)
\]
\[
\cosh(x) = \frac{e^x + e^{-x}}{2}, \quad (A.8)
\]

and
\[
\tanh^{-1}(x) = \frac{1}{2} \log \left( \frac{1 + x}{1 - x} \right). \quad (A.9)
\]

Using these identities we can further observe that
\[
\cosh \left( \tanh^{-1}(x) \right) = \frac{e^{\tanh^{-1}(x)} + e^{-\tanh^{-1}(x)}}{2} = \frac{(1+x)}{2} + \frac{(1-x)}{2} \cdot \frac{1}{2}.
\]

Now, for any time \( t \geq 0 \) we can note that \( N_t = Q_t + D_t \). Thus, we have that
\[
E \left[ z^{N_t} \right] = E \left[ z^{Q_t} z^{D_t} \right] = G(z, z, t),
\]
where \( G(z_1, z_2, t) \) is as given in Proposition A.5. Setting \( z_1 = z_2 = z \) and \( D_0 = N_0 - Q_0 \), this is
\[
G(z, z, t) = z^{N_0 - Q_0} e^{\frac{\nu^* (z - \alpha)}{2 \alpha} t} \left( 1 - \left( \tanh \left( \frac{t}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} + \tanh^{-1} \left( \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \right) \right) \right) \right)^{2 \frac{\nu^*}{\alpha}} \cdot \left( \frac{\mu + \alpha}{2\alpha} - \frac{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}{2\alpha} \tanh \left( \frac{t}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} + \tanh^{-1} \left( \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \right) \right) \right) \cdot \left( \cosh \left( \tanh^{-1} \left( \frac{2\alpha z - \mu - \alpha}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \right) \right) \right)^{\frac{\nu^*}{\alpha}}. \quad (A.10)
\]
Using the hyperbolic identities and simplifying, this is

\[ G(z, z, t) = z^{N_0}Q_0e^{\frac{\alpha^2}{2\alpha}} \left( \frac{2e^{\frac{t}{2}}\sqrt{\mu + \alpha}^2 - 4\alpha \mu z}{1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z} + \left( 1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z} \right) e^{\frac{t}{2}}\sqrt{\mu + \alpha}^2 - 4\alpha \mu z} \right) \]

\begin{align*}
\frac{\mu + \alpha}{2\alpha} + \sqrt{\frac{(\mu + \alpha)^2}{2} - 4\alpha \mu z} & \left( \frac{1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z}}{1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{\mu + \alpha}^2 - 4\alpha \mu z}} \right) e^{\frac{t}{2}}\sqrt{\mu + \alpha}^2 - 4\alpha \mu z \\
\end{align*}

which is the stated result. However, the simplifications used to reach this form require multiple parts and several steps and so we can these individually now. We start with the hyperbolic tangent function that appears on the first and second lines of Equation A.10. Using Equations A.7 and A.9, this is

\[- \tanh \left( \frac{t}{2} \sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} \right) + \tanh^{-1} \left( \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \right) \]

\[- e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} + \frac{1}{2} \log \left( \frac{1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}}{1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}} \right) \]

\[= - e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} + \frac{1}{2} \log \left( \frac{1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}}{1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}} \right) \]

\[= - \frac{t\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} + \log \left( \frac{1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}}{1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}} \right)}{e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} - 1 \]

\[1 - \frac{1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}}{e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} = 1 + \frac{1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}}}{e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \]

\[1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} - \left( 1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \right) e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z} \]

\[1 - \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} + \left( 1 + \frac{\mu + \alpha - 2\alpha z}{\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}} \right) e^{\frac{t}{2}}\sqrt{(\mu + \alpha)^2 - 4\alpha \mu z}.
\]
Thus, the second line of Equation A.10 simplifies as

\[
\left(\mu + \alpha - \frac{\sqrt{\mu + \alpha^2 - 4\alpha \mu z}}{2\alpha}\right) \left(1 - \frac{\mu + \alpha - 2az}{\sqrt{\mu + \alpha^2 - 4\alpha \mu z}}\right) e^{t\sqrt{\mu + \alpha^2 - 4\alpha \mu z}} \left(\frac{z}{2\alpha} \sqrt{\mu + \alpha^2 - 4\alpha \mu z} + \text{tan}^{-1}\left(\frac{\mu + \alpha - 2az}{\sqrt{\mu + \alpha^2 - 4\alpha \mu z}}\right)\right)
\]

Following the same approach, the first line of Equation A.10 rearranges to

\[
\left(\frac{\mu + \alpha - 2az}{\sqrt{\mu + \alpha^2 - 4\alpha \mu z}}\right) e^{t\sqrt{\mu + \alpha^2 - 4\alpha \mu z}}\left(\frac{z}{2\alpha} \sqrt{\mu + \alpha^2 - 4\alpha \mu z} + \text{tan}^{-1}\left(\frac{\mu + \alpha - 2az}{\sqrt{\mu + \alpha^2 - 4\alpha \mu z}}\right)\right)
\]
Finally, the third line of Equation A.10 is simplified through use of Equations A.8 and A.9. This expression is then given by

\[
\left(\frac{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}\right)^{\frac{\nu^*}{\alpha}} \left(\frac{1-\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}{1-\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}\right)^{\frac{\nu^*}{\alpha}} \left(\frac{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}\right) = \left(\frac{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}\right)^{\frac{\nu^*}{2\alpha}} \left(\frac{1-\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}{1-\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}\right)^{\frac{\nu^*}{2\alpha}} \left(\frac{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}{1+\frac{\mu+\alpha-2az}{\sqrt{\mu+\alpha}^2-4a\mu z}}\right)
\]

Together these forms give the stated result. \(\square\)

### A.3 Proof of Theorem 5.1

**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

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

We now note that \(N_t^\nu - E[N_t^\nu]\) is a martingale by definition, and so \(|N_t^\nu - E[N_t^\nu]|\) is a sub-martingale. Additionally, we can observe that

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

because \(s^2 \leq t\) for any \(t\). By Doob’s martingale inequality, we have

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

From Proposition 2.4, we note that the variance of a Queue-Hawkes counting process with baseline intensity \(\nu^*\), intensity jump size \(\alpha\), decay rate \(\beta\), and service rate \(\mu\) is upper-bounded
by the variance of an Affine Queue-Hawkes counting process with baseline \( \nu^* \), jump size \( \alpha \), and service rate \( \mu + \beta \). Using the explicit form of the Affine Queue-Hawkes counting process variance from Proposition 3.6, we have the bound

\[
\text{Var} \left( \frac{N^{\nu}_{(s+1)^2}}{N^{\nu}_{(s+1)^2}} \right) \leq \text{Var} \left( \frac{N^{\eta}_{(s+1)^2}}{N^{\eta}_{(s+1)^2}} \right) = \frac{(\mu + \beta)^2 + \alpha^2}{(\mu + \beta - \alpha)^2} \cdot \left( s + 1 \right)^2 - \frac{2\alpha\mu(\nu_0 - \nu_\infty)}{(\mu + \beta - \alpha)^3} \cdot e^{-(\mu + \beta - \alpha)(s+1)^2} \\
\quad + \frac{2(\alpha^2 + \alpha(\mu + \beta))\nu_0}{(\mu + \beta - \alpha)^3} \cdot \left( 1 - e^{-(\mu + \beta - \alpha)(s+1)^2} \right).
\]

Together, this implies that \( \text{P} \left( E_s \right) \in O \left( \frac{1}{s^3} \right) \). Therefore \( \sum_{s=0}^\infty \text{P} \left( E_s \right) < \infty \), and so by the Borel-cantelli lemma, \( \frac{\sum_{s=0}^\infty E \left[ N^{\nu}_s \right]}{t} \xrightarrow{a.s.} 0 \). Since \( \lim_{t \to \infty} \frac{E \left[ N^{\nu}_s \right]}{t} = \nu_\infty \), we complete the proof.

\[\text{A.4 Proof of Theorem 5.5}\]

\(\text{Proof.}\) The proof will follow in two steps. The first step is to show that the limiting moment generating function converges to a PDE given by \( \mathcal{M}_\infty \) using properties of the exponential function and Lemma 5.4. 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 fluid limit. We begin with the infinitesimal generator form which simplifies through the linearity of expectation as

\[
\frac{\partial}{\partial t} \tilde{\mathcal{M}}^n(t, \theta, \theta Q) = \frac{\partial}{\partial t} \mathbb{E} \left[ \frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n} \right] \\
= \mathbb{E} \left[ \beta (\nu^* n - \nu_t(n)) \frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n} \right] + \mathbb{E} \left[ \nu_t(n) \left( e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} - 1 \right) e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right] \\
+ \mathbb{E} \left[ \frac{\mu Q_t(n)}{n} \left( e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} - 1 \right) e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right] \\
+ \mathbb{E} \left[ \frac{\nu_t(n)}{n} \left( e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} - 1 \right) e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right] \\
= \beta \nu^* \theta_\nu \mathbb{E} \left[ e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right] - \beta \theta_\nu \mathbb{E} \left[ \frac{\nu_t(n)}{n} e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right] \\
+ n \left( e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} - 1 \right) \mathbb{E} \left[ \frac{\nu_t(n)}{n} \frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n} \right] \\
+ \frac{\mu}{n} \mathbb{E} \left[ Q_t(n) \left( e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} - 1 \right) e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right] \\
= \beta \nu^* \theta_\nu \tilde{\mathcal{M}}(t, \theta, \theta Q) + \left( n \left( e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} - 1 \right) - \beta \theta_\nu \right) \frac{\partial}{\partial \theta_\nu} \tilde{\mathcal{M}}(t, \theta, \theta Q) \\
+ \frac{\mu}{n} \mathbb{E} \left[ Q_t(n) \left( \frac{-\theta_\nu (\nu_t(n) - \nu^* n)}{Q_t(n)} - \theta_\theta \right) e^{\frac{\theta_\nu n \nu_t(n) + \theta Q Q_t(n)}{n}} \right],
\]
where the last equality holds for sufficiently large \( n \), where \( \epsilon_n \) is in some bounded interval as according to Lemma 5.4. Then, by rearranging further we can see that in limit this becomes

\[
\frac{\partial}{\partial t} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) = \beta \nu^* \theta \nu M^n(t, \theta, \theta_Q) + \left( n \left( \frac{\alpha_n}{n} + \frac{\alpha_Q}{n} - 1 \right) - \beta \theta \nu \right) \frac{\partial}{\partial \theta} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) - \mu \theta \nu \theta \nu E \left[ \frac{n}{n} e^{\frac{\mu}{n} \nu (n) + \frac{\alpha_Q}{n} Q^+ (n)} \right] \\
+ \mu \theta \nu \nu E \left[ \frac{\partial}{\partial \theta} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) \right] - \mu \theta \nu E \left[ \frac{\partial}{\partial \nu} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) \right] \\
= (\mu + \beta) \nu^* \theta \nu \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) + \left( n \left( \frac{\alpha_n}{n} + \frac{\alpha_Q}{n} - 1 \right) - (\mu + \beta) \theta \nu \right) \frac{\partial}{\partial \theta} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) \\
- \mu \theta Q \frac{\partial}{\partial \theta \nu} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) + \mu \nu Q \frac{\partial}{\partial \theta \nu} \tilde{\mathcal{M}}^n(t, \theta, \theta_Q) \\
\rightarrow \infty \ (\mu + \beta) \nu^* \theta \nu \tilde{\mathcal{M}}^\infty(t, \theta, \theta_Q) + \left( \theta Q - (\mu + \beta - \alpha) \theta \nu \right) \frac{\partial}{\partial \theta} \tilde{\mathcal{M}}^\infty(t, \theta, \theta_Q) - \mu \theta Q \frac{\partial}{\partial \theta \nu} \tilde{\mathcal{M}}^\infty(t, \theta, \theta_Q).
\]

We now solve this partial differential equation for \( \tilde{\mathcal{M}}^\infty(t, \theta, \theta_Q) \) through the method of characteristics. For simplicity’s sake we will instead use this procedure to solve for \( G(t, \theta, \theta_Q) = \log \left( \frac{\tilde{\mathcal{M}}^\infty(t, \theta, \theta_Q)}{\tilde{\mathcal{M}}^\infty(0, \theta, \theta_Q)} \right) \). This PDE is given by

\[
(\mu + \beta) \nu^* \theta \nu = \frac{\partial}{\partial t} G(t, \theta, \theta_Q) + \mu \theta Q \frac{\partial}{\partial \theta Q} G(t, \theta, \theta_Q) + ((\mu + \beta - \alpha) \theta \nu - \theta Q) \frac{\partial}{\partial \theta \nu} G(t, \theta, \theta_Q),
\]

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

\[
\frac{d\theta Q}{dz}(x, y, z) = \mu \theta Q, \quad \theta Q(x, y, 0) = x \\
\frac{d\theta}{dz}(x, y, z) = (\mu + \beta - \alpha) \theta \nu - \theta Q, \quad \theta \nu(x, y, 0) = y \\
\frac{d}{dz}(x, y, z) = 1, \quad t(x, y, 0) = 0 \\
\frac{dg}{dz}(x, y, z) = (\mu + \beta) \nu^* \theta \nu = (\mu + \beta - \alpha) \nu \nu_\infty \theta \nu, \quad g(x, y, 0) = x Q_0 + y \nu_0.
\]

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

\[
\theta Q(x, y, z) = xe^{\mu z}, \\
\theta \nu(x, y, z) = ye^{(\mu+\beta-\alpha)z} + \frac{x}{\beta - \alpha} \left( e^{\mu z} - e^{(\mu+\beta-\alpha)z} \right), \\
t(x, y, z) = z, \\
g(x, y, z) = x Q_0 + y \nu_0 + \nu \nu_\infty \left( y - \frac{x}{\beta - \alpha} \right) \left( e^{(\mu+\beta-\alpha)z} - 1 \right) + \frac{x \nu \nu (\mu + \beta - \alpha) (e^{\mu z} - 1)}{\mu (\beta - \alpha)}.
\]

Now, we can solve for the characteristic variables in terms of the original variables and find \( x = \theta_Q e^{-\mu t} \), \( y = \theta \nu e^{-((\mu+\beta-\alpha)t)} + \frac{\theta Q}{\beta - \alpha} (e^{-\mu t} - e^{-(\mu+\beta-\alpha)t}) \), and \( z = t \), so this gives a PDE
solution of
\[ G(t, \theta_Q, \theta_{\nu}) = g \left( \theta_Q e^{-\mu t}, \theta_{\nu} e^{-(\mu + \beta - \alpha)t} + \frac{\theta_Q}{\beta - \alpha} \left( e^{-\mu t} - e^{-(\mu + \beta - \alpha)t} \right), t \right) \]
\[ = Q_0 \theta_Q e^{-\mu t} + \nu_0 \left( \theta_{\nu} e^{-(\mu + \beta - \alpha)t} + \frac{\theta_Q}{\beta - \alpha} \left( e^{-\mu t} - e^{-(\mu + \beta - \alpha)t} \right) \right) \]
\[ + \nu_{\infty} \left( \theta_{\nu} - \frac{\theta_Q}{\beta - \alpha} \right) \left( 1 - e^{-(\mu + \beta - \alpha)t} \right) + \frac{\theta_Q \nu_{\infty} (\mu + \beta - \alpha)(1 - e^{-\mu t})}{\mu (\beta - \alpha)}. \]

If instead \( \beta = \alpha \), the solutions to the characteristic ODE’s are as follows:
\[ \theta_Q(x, y, z) = xe^{\mu z}, \]
\[ \theta_{\nu}(x, y, z) = e^{\mu z} (y - xz), \]
\[ t(x, y, z) = z, \]
\[ g(x, y, z) = xQ_0 + y\nu_0 + \nu_{\infty} y (e^{\mu z} - 1) - \frac{\nu_{\infty} \theta_Q}{\mu} (e^{\mu z} (\mu z - 1) + 1). \]

This makes our expressions for the characteristic variables \( x = \theta_Q e^{-\mu t}, y = \theta_{\nu} e^{-\mu t} + \theta_Q e^{-\mu t}, \) and \( z = t \). This now makes the PDE solution
\[ G(t, \theta_Q, \theta_{\nu}) = g \left( \theta_Q e^{-\mu t}, \theta_{\nu} e^{-\mu t} + \theta_Q e^{-\mu t}, t \right) \]
\[ = Q_0 \theta_Q e^{-\mu t} + \nu_0 \theta_{\nu} e^{-\mu t} + \nu_0 \theta_Q e^{-\mu t} + \nu_{\infty} (\theta_{\nu} + \theta_Q t) (1 - e^{-\mu t}) - \frac{\nu_{\infty} \theta_Q}{\mu} (\mu t - 1 + e^{-\mu t}), \]
and we can observe that each of these cases simplify to the corresponding means of the queue and the intensity, which yields the stated result. \( \square \)

### A.5 Proof of Theorem 5.6

**Proof.** The upper and lower bounds of this approximation follow in a similar manner to Propositions 2.5 and A.4: bounding the quantity \( Q_t^\nu(n) \left( \frac{\nu(n) - \nu^*}{Q_t^\nu(n)} \right)^2 \) above and below by observing
\[ 0 \leq Q_t^\nu(n) \left( \frac{\nu(n) - \nu^*}{Q_t^\nu(n)} \right)^2 \leq \alpha (\nu_t(n) - \nu^*). \]

To consolidate the development of the two bounds into one approach, we introduce the extra parameter \( \gamma \in \{0, 1\} \) and replace \( Q_t^\nu(n) \left( \frac{\nu(n) - \nu^*}{Q_t^\nu(n)} \right)^2 \) by \( \gamma \alpha (\nu_t(n) - \nu^*) \) 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 A.1, this is
\[ \frac{\partial}{\partial t} \hat{M}^n(\theta_{\nu}, \theta_Q, t) = \frac{\partial}{\partial t} \mathbb{E} \left[ e^{\theta_Q (\frac{\nu(n) - \nu^*}{\sqrt{n}})} + \theta_Q \left( \frac{Q_t^\nu(n) - \nu^*}{\sqrt{n}} \right) \right] \]
\[ = \mathbb{E} \left[ \nu_t(n) \left( e^{\frac{\nu(n) - \nu^*}{\sqrt{n}}} - 1 \right) e^{\theta_Q (\frac{Q_t^\nu(n) - \nu^*}{\sqrt{n}})} \right] \]
\[ + \mathbb{E} \left[ \mu Q_t^\nu(n) \left( e^{\frac{\theta_Q (\frac{Q_t^\nu(n) - \nu^*}{\sqrt{n}})}{\sqrt{n}}} - 1 \right) e^{\theta_Q (\frac{Q_t^\nu(n) - \nu^*}{\sqrt{n}})} \right] \]
\[ + \mathbb{E} \left[ \frac{\beta \theta_{\nu}}{\sqrt{n}} (\nu^* - \nu_t(n)) e^{\theta_Q (\frac{Q_t^\nu(n) - \nu^*}{\sqrt{n}})} \right]. \]
As a first step, we simplify this expression through the linearity of expectation. Moving deterministic terms outside of the expectation and re-scaling, we have

\[
\frac{\partial}{\partial t} \hat{M}^n(\theta, \theta, t) = \sqrt{n} \left( e^{\frac{\alpha \theta + \theta_Q}{\sqrt{n}}} - 1 \right) E \left[ \frac{\nu_t(n)}{\sqrt{n}} e^{\frac{\nu_t(n) - n\nu^*}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) \right] \\
+ \frac{\mu Q_t^r(n)}{\sqrt{n}} \left( e^{\frac{\nu_t(n) - n\nu^*}{\sqrt{n}}} - \frac{\theta_Q}{\sqrt{n}} - 1 \right) e^{\frac{\nu_t(n) - n\nu^*}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) \\
+ \beta \theta \nu^* \sqrt{n} E \left[ \frac{\nu_t(n) e^{\frac{\nu_t(n) - n\nu^*}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] - \beta \theta \nu E \left[ \frac{\nu_t(n) e^{\frac{\nu_t(n) - n\nu^*}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right].
\]

For the terms on the first and third 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 now yields

\[
\frac{\partial}{\partial t} \hat{M}^n(\theta, \theta, t) = \left( \alpha \theta + \theta_Q + \frac{(\alpha \theta + \theta_Q)^2}{2} + O \left( \frac{1}{n} \right) \right) \left( \frac{\partial}{\partial \theta} \hat{M}^n(\theta, \theta, t) + \nu \sqrt{n} \hat{M}^n(\theta, \theta, t) \right) \\
+ \frac{\mu Q_t^r(n)}{\sqrt{n}} \left( \frac{\nu_t(n) - n\nu^*}{Q_t^r(n)} \right) - \frac{\theta_Q}{\sqrt{n}} + \frac{1}{2n} \left( \frac{\nu_t(n) - n\nu^*}{Q_t^r(n)} + \theta_Q \right)^2 + O \left( n^{-\frac{3}{2}} \right) \\
\cdot e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) \\
+ \beta \theta \nu^* \sqrt{n} E \left[ \frac{\nu_t(n) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] - \beta \theta \nu E \left[ \frac{\nu_t(n) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right].
\]

We now begin distributing and combining like terms through linearity of expectation. Moreover, we distribute within the expectation on the second line and cancel \(Q_t^r(n)\) across the numerator and denominator where possible.

\[
\frac{\partial}{\partial t} \hat{M}^n(\theta, \theta, t) = \left( \alpha \theta + \theta_Q + \frac{(\alpha \theta + \theta_Q)^2}{2} - \beta \theta \nu + O \left( \frac{1}{n} \right) \right) \left( \frac{\partial}{\partial \theta} \hat{M}^n(\theta, \theta, t) + \nu \sqrt{n} \hat{M}^n(\theta, \theta, t) \right) \\
- \mu \theta \nu E \left[ \frac{\nu_t(n) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] + \mu \theta \nu^* \sqrt{n} E \left[ \frac{\nu_t(n) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] \\
+ \mu \theta \nu E \left[ \frac{Q_t^r(n) \left( \frac{\nu_t(n) - n\nu^*}{Q_t^r(n)} \right) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] \\
+ \mu \theta \nu Q_t \left[ \frac{\nu_t(n) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] - \mu \theta \nu \theta Q \nu^* E \left[ \frac{\nu_t(n) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] \\
+ \mu \theta \nu E \left[ \frac{Q_t^r(n) \left( \frac{\nu_t(n) - n\nu^*}{Q_t^r(n)} \right) e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) }{\sqrt{n}} \right] + \frac{1}{2n} \left( \frac{\nu_t(n) - n\nu^*}{Q_t^r(n)} \right)^2 e^{\frac{(\nu_t(n) - n\nu^*)}{\sqrt{n}}} + \theta_Q \left( \frac{Q_t^r(n) - n\nu^*}{\sqrt{n}} \right) \\
+ \beta \theta \nu^* \sqrt{n} \hat{M}^n(\theta, \theta, t).
\]

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 \(Q_t^r(n) \left( \frac{\nu_t(n) - n\nu^*}{Q_t^r(n)} \right)^2\) by \(\gamma \alpha (\nu_t(n) - n\nu^*)\) inside the
We will now solve this limiting partial differential equation through the method of characteristics. To denote that we have now made this replacement and changed the function, we add $\gamma$ as a subscript to the moment generating function, i.e. $\mathcal{M}_\gamma^n(\theta_\nu, \theta_Q, t)$.

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

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.

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

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

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

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

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

\[
\begin{align*}
\frac{dt}{dz}(x, y, z) &= 1, & t(x, y, 0) &= 0, \\
\frac{d\theta_\nu}{dz}(x, y, z) &= (\mu + \beta - \alpha)\theta_\nu - \theta_Q, & \theta_\nu(x, y, 0) &= x, \\
\frac{d\theta_Q}{dz}(x, y, z) &= \mu \theta_Q, & \theta_Q(x, y, 0) &= y, \\
\frac{dg}{dz}(x, y, z) &= \left(\frac{\gamma \alpha \mu \theta_\nu^2}{2} + \mu \theta_\nu \theta_Q\right) (\nu_\infty - \nu^*) + \left(\theta_Q^2 + (\alpha \theta_\nu + \theta_Q) \nu_\infty\right) \frac{\nu_\infty}{2}, & g(x, y, 0) &= x \nu_0 + y Q_0.
\end{align*}
\]

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

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

which we now use to solve the remaining equation. Re-writing the characteristic equation for $g$, we have

\[
\frac{dg}{dz}(x, y, z) = \frac{\gamma \alpha \mu (\nu_\infty - \nu^*)}{2} \left((x - \frac{y}{\beta - \alpha})^2 e^{2(\mu + \beta - \alpha)z} + \frac{2}{\beta - \alpha} (xy - \frac{y^2}{\beta - \alpha}) e^{(2\mu + \beta - \alpha)z} + \frac{y^2}{(\beta - \alpha)^2} e^{2\mu z}\right) + \mu (\nu_\infty - \nu^*) \left((xy - \frac{y^2}{\beta - \alpha}) e^{(2\mu + \beta - \alpha)z} + \frac{y^2}{\beta - \alpha} e^{2\mu z}\right) + \frac{\nu_\infty}{2} \left(\alpha^2 \left(x - \frac{y}{\beta - \alpha}\right)^2 e^{2(\mu + \beta - \alpha)z} + \frac{\beta^2}{(\beta - \alpha)^2} y^2 e^{2\mu z}\right),
\]

and so by grouping coefficients of like exponential functions and then integrating with respect to $z$, this solves to

\[
g(x, y, z) = x \nu_0 + y Q_0 + \left(x - \frac{y}{\beta - \alpha}\right)^2 \left(\frac{\gamma \alpha \mu (\nu_\infty - \nu^*)}{2} + \frac{\alpha^2 \nu_\infty}{2}\right) e^{2(\mu + \beta - \alpha)z} - \frac{1}{2(\mu + \beta - \alpha)}
\]
\[
+ \left(xy - \frac{y^2}{\beta - \alpha}\right) \left(\frac{\gamma \alpha \mu}{\beta - \alpha} + \mu\right) \left(\nu_\infty - \nu^*\right) + \frac{\alpha \beta \nu_\infty}{\beta - \alpha} e^{2\mu - \beta - \alpha} - \frac{1}{2\mu + \beta - \alpha}
\]
\[
+ y^2 \left(\frac{\gamma \alpha \mu (\nu_\infty - \nu^*)}{2(\beta - \alpha)^2} + \frac{\mu (\nu_\infty - \nu^*)}{\beta - \alpha} + \frac{\nu_\infty}{2} + \frac{\nu_\infty \beta^2}{2(\beta - \alpha)^2}\right) e^{2\mu z} - \frac{1}{2\mu}.\]

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:

\[
z = t, \quad y = \theta_Q e^{-\mu t}, \quad x = \theta_\nu e^{-(\mu + \beta - \alpha)t} + \frac{\theta_Q}{\beta - \alpha} \left(e^{-\mu t} - e^{-(\mu + \beta - \alpha)t}\right).
\]
Thus, we can then solve for $G_\gamma(\theta_\nu, \theta_Q, t)$ via
\[
G_\gamma(\theta_\nu, \theta_Q, t) = g \left( \theta_\nu e^{-(\mu + \beta - \alpha)t} + \frac{\theta_Q}{\beta - \alpha} \left( e^{-\mu t} - e^{-(\mu + \beta - \alpha)t} \right), \theta_Q e^{-\mu t} \right)
\]
\[
= \nu_0 \theta_\nu e^{-(\mu + \beta - \alpha)t} + \frac{\nu_0 \theta_Q}{\beta - \alpha} \left( e^{-\mu t} - e^{-(\mu + \beta - \alpha)t} \right) + Q_0 \theta_Q e^{-\mu t}
\]
\[
+ \left( \theta_\nu - \frac{\theta_Q}{\beta - \alpha} \right)^2 \left( \frac{\alpha \mu}{2} - \frac{\alpha^2 \nu_\infty}{2} \right) + e^{-2(\mu + \beta - \alpha)t} - \frac{1}{2(\mu + \beta - \alpha)}
\]
\[
+ \left( \theta_\nu \theta_Q - \frac{\theta_Q^2}{\beta - \alpha} \right) \left( \left( \frac{\gamma \alpha \mu}{\beta - \alpha} \right) \left( \nu_\infty - \nu^* \right) + \frac{\alpha \beta \nu_\infty}{\beta - \alpha} \right) + e^{-2(\mu + \beta - \alpha)t} - \frac{1}{2(\mu + \beta - \alpha)}
\]
\[
+ \left( \frac{\gamma \alpha \mu}{2(\beta - \alpha)^2} \right) \left( \frac{\mu}{2} \left( \nu_\infty - \nu^* \right) + \frac{\nu_\infty}{2} + \frac{\nu_\infty^2}{2(\beta - \alpha)^2} \right) + e^{-2\mu t} - \frac{1}{2\mu}.
\]

By Lemma A.2, we have that $\mathcal{M}_\gamma(\theta_\nu, \theta_Q, t) \leq \mathcal{M}_\gamma^\infty(\theta_\nu, \theta_Q, t) \leq \mathcal{M}_\gamma(\theta_\nu, \theta_Q, t)$ and since $\mathcal{M}_\gamma^\infty(\theta_\nu, \theta_Q, t) = e^{G_\gamma(\theta_\nu, \theta_Q, t)}$, 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
\[
\mathcal{M}_\gamma(\theta_\nu, 0, t) = \exp \left( \nu_0 \theta_\nu e^{-(\mu + \beta - \alpha)t} + \frac{\theta_Q^2}{2} \left( \frac{\gamma \alpha \mu}{\beta - \alpha} \right) \left( \nu_\infty - \nu^* \right) + \frac{\alpha^2 \nu_\infty}{2(\mu + \beta - \alpha)} \right),
\]
whereas for the queue we take $\theta_\nu = 0$ and receive
\[
\mathcal{M}_\gamma(0, \theta_Q, t) = \exp \left( \nu_0 \theta_Q e^{-(\mu + \beta - \alpha)t} \right) + \frac{\theta_Q^2}{(\beta - \alpha)^2} \left( \nu_\infty - \nu^* \right) + \frac{\alpha \beta \nu_\infty}{\beta - \alpha} \left( \nu_\infty - \nu^* \right) + \frac{\alpha \beta \nu_\infty}{\beta - \alpha} \left( \nu_\infty - \nu^* \right) + \frac{\nu_\infty^2}{2(\beta - \alpha)^2} \left( 1 - e^{-2\mu t} \right) - \frac{1}{2\mu}.
\]

Now if $\beta = \alpha$, the solution to the characteristic ODE for $\theta_\nu$ is instead
\[
\theta_\nu = x e^{\mu z} - y z e^{\mu z},
\]
whereas the solutions for $\theta_Q$ and $t$ are unchanged: $\theta_Q = y e^{\mu z}$ and $t = z$. This then implies that ODE for $g$ is given by
\[
\frac{dg}{dz}(x, y, z) = \left( \frac{\gamma \alpha \mu}{2} \right) \left( x^2 e^{2\mu z} - 2xyz e^{2\mu z} + y^2 z^2 e^{2\mu z} \right) + \left( \nu_0 \theta_\nu e^{-(\mu + \beta - \alpha)t} \right) \left( \nu_\infty - \nu^* \right) + \frac{\nu_\infty^2}{2(\beta - \alpha)^2} \left( 1 - e^{-2\mu t} \right) - \frac{1}{2\mu}.
\]

which yields a solution of
\[
g(x, y, z) = x \nu_0 + y Q_0 + \left( \frac{\gamma \alpha \mu}{2} \right) \left( x^2 e^{2\mu z} - 2xyz e^{2\mu z} + y^2 z^2 e^{2\mu z} \right) \mu (xy e^{2\mu z} - y^2 z e^{2\mu z}) \left( \nu_\infty - \nu^* \right) + \frac{\nu_\infty^2}{2(\beta - \alpha)^2} \left( 1 - e^{-2\mu t} \right) - \frac{1}{2\mu}
\]
\[
+ 2(\alpha^2 xy + \alpha y^2) \left( \frac{e^{2\mu z} (2\mu z - 1) + 1}{4\mu} \right) + 2(\alpha^2 xy + \alpha y^2) \left( \frac{e^{2\mu z} (2\mu z - 1) + 1}{4\mu^2} \right) + \alpha^2 y^2 \left( \frac{e^{2\mu z} (2\mu z - 1) + 1}{4\mu^2} \right) + \alpha^2 y^2 \left( \frac{e^{2\mu z} (2\mu z - 1) + 1}{4\mu^3} \right).
\]
In this case the inverse solutions are

\[ z = t, \quad y = \theta_Q e^{-\mu t}, \quad x = \theta_\nu e^{-\mu t} + \theta_Q e^{-\mu t}, \]

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

\[
G_\gamma(\theta_\nu, \theta_Q, t) = g(\theta_\nu e^{-\mu t} + \theta_Q e^{-\mu t}, t, \theta_Q e^{-\mu t} + \theta_\nu e^{-\mu t})
\]

\[
= \nu_0 \theta_\nu e^{-\mu t} + \nu_0 \theta_Q e^{-\mu t} + Q_0 \theta_Q e^{-\mu t} + \Bigg( \left( \frac{\gamma \alpha (\theta_\nu + \theta_Q t)^2}{2} + \theta_\nu \theta_Q + \theta_Q^2 t \right) \frac{1 - e^{-2\mu t}}{2}
\]

\[
- (\gamma \alpha (\theta_\nu \theta_Q + \theta_Q^2 t + \theta_Q^2) \frac{2\mu t - 1 + e^{-2\mu t}}{4\mu} + \gamma \alpha \theta_Q^2 \left( \frac{2\mu (\mu t - 1) + 1 - e^{-2\mu t}}{4\mu^2} \right)) (\nu_\infty - \nu^*)
\]

\[
+ \frac{\nu_\infty}{2} \left( 2\theta_Q^2 + \alpha^2 \theta_\nu^2 + 2\alpha^2 \theta_\nu \theta_Q t + \alpha^2 \theta_Q^2 t^2 + 2\alpha \theta_\nu \theta_Q + 2\alpha \theta_Q^2 t \right) \frac{1 - e^{-2\mu t}}{2\mu}
\]

\[
- 2 (\alpha^2 \theta_\nu \theta_Q + \alpha^2 \theta_Q^2 t + \alpha \theta_Q^2) \left( \frac{2\mu t - 1 + e^{-2\mu t}}{4\mu^2} \right) + \alpha^2 \theta_Q^2 \left( \frac{2\mu (\mu t - 1) + 1 - e^{-2\mu t}}{4\mu^3} \right).
\]

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