Submitted to *Stochastic Systems*
manuscript SSy-2017-533.R3

Queues Driven by Hawkes Processes

Andrew Daw, Jamol Pender
School of Operations Research & Information Engineering, Cornell University, Ithaca, NY 14850
{amd399@cornell.edu, jjp274@cornell.edu}

Many stochastic systems have arrival processes that exhibit clustering behavior. In these systems, arriving entities influence additional arrivals to occur through self-excitation of the arrival process. In this paper, we analyze an infinite server queueing system in which the arrivals are driven by the self-exciting Hawkes process and where service follows a phase-type distribution or is deterministic. In the phase-type setting, we derive differential equations for the moments and a partial differential equation for the moment generating function; we also derive exact expressions for the transient and steady-state mean, variance, and covariances. Furthermore, we also derive exact expressions for the auto-covariance of the queue and provide an expression for the cumulant moment generating function in terms of a single ordinary differential equation. In the deterministic service setting, we provide exact expressions for the first and second moments and the queue auto-covariance. As motivation for our Hawkes queueing model, we demonstrate its usefulness through two novel applications. These applications are trending internet traffic and arrivals to nightclubs. In the web traffic setting, we investigate the impact of a click. In the nightclub or *Club Queue* setting, we design an optimal control problem for the optimal rate to admit club-goers.

Key words: Infinite-Server Queues, Hawkes Processes, Phase-type Distributions, Moments, Social Media, Nightclubs, Cumulant Moment Generating Function, Auto-covariance
MSC2000 subject classification: Primary 60K25; Secondary 90B22, 93E20

1. Introduction

The arrival process is a fundamental component of stochastic queueing models. In most models, these arrival processes are driven by a Poisson process, which is well suited for environments in which arrivals have no influence on one another. If the arrival process is a simple (single jump) random counting process with independent increments, Prékopa (1957) shows that this is equivalent to a non-homogeneous Poisson process. However, this can be unrealistic for many situations. For example, in the trading of financial assets, transactions tend to occur together as traders are often responding to the same information as their peers or to their actions Azizpour et al. (2016).

Additionally, earthquakes and other forms of geological tension frequently occur in quick succession, as aftershocks can continue to affect an area soon after the initial tremors Ogata (1988). Even patterns of violent crime have been known to occur in clusters, as victims may decide to retaliate Mohler et al. (2011). In each of these examples, the occurrence of an event makes the occurrence of the next more likely to happen in quick succession, which means that the sequence of arrivals tends to form clusters. This type of phenomena would be better modeled by variables that are not memoryless, so that the occurrences can have an influence on those that follow soon after and increments are not independent.

One stochastic arrival process that captures clustering of arrivals was introduced in 1971 by Hawkes (1971), and is referred to as the Hawkes process. This stochastic process counts the number of arrivals and, unlike the Poisson process, it self-excites. This means that when one arrival occurs, it increases the likelihood that another arrival will occur soon afterwards. The Hawkes process does so through treating both the counting process and the rate of arrivals as coupled stochastic processes. Because the arrival rate increases, it is treated in a general sense as the arrival "intensity," which can be thought of as a representation of the excitement at that time. The higher the intensity, the more likely it is that an arrival will occur. In this setting, the number of arrivals and the arrival intensity represent the system together as a pair.

Historically, the Hawkes process has been studied predominantly in financial settings. However, it has only recently received a significant amount of attention in broader and more general contexts. For a general overview, a review of the Hawkes process was written by Laub et al. (2015). In our work, we are particularly interested in socially informed queueing systems, and we use these systems as a motivation for both studying the Hawkes process and applying it to queueing models. For example, in situations in which a person does not know the value of competing offers or services, she may decide to pursue the option that has the most other people already waiting for it. When one can't be sure of what is earned by waiting, the willingness of others to wait can often be the best indicator.

As a quick example for the sake of building intuition, consider walking past a street performer. If there is only a handful of other people watching, one may not feel a desire to stop and see the performance. However, if there is a large crowd already watching it is more enticing to join the group and see what is happening. This is the basic motivation of self-exciting and clustering arrival processes. Although this example is simple, the concept itself has powerful implications for service systems. Several naturally occurring examples of these systems were detailed in a recent Chicago Booth Review article, Mordfin (2015). These examples include cellular companies paying employees to join the lines outside stores during product launches and pastry enthusiasts waiting hours in queue to buy baked goods from the famed Dominique Ansel Bakery in New York. (The article even

includes a story of a German man joining a long queue in 1947 without any knowledge of what awaited him, only to find it was for visas to the United States!) Another example of self exciting arrivals in service settings are flight cancellations, discussed in a recent Business Insider article, Zhang (2016). Because of widespread events like inclement weather and information technology infrastructure failures, flights are often cancelled in mass. However, as that article notes, even one plane experiencing mechanical issues can cause a cluster of downstream cancellations throughout its flight legs.

In this paper, we apply our results to two main applications: the viral nature of modern web traffic and the appeal associated with the lengths of queues for nightclubs. In socially informed internet traffic, webpages experience arrivals of users in clusters due to the contagion-like spread of information. If one user shares a webpage, others become more likely to view and share it as well. We demonstrate this through an example from Twitter data and explore the impact of a click. The night club example can be seen as an effect of having to pay a cover fee up front to enter the venue. Because club-goers must pay before ever seeing inside, the number of others already in queue to enter the club gives a sense of the attraction they are awaiting. In this setting we consider the managerial control problem of how quickly to admit customers to maximize earnings. Again, in these examples the occurrence of an event or arrival of a customer increases the likelihood that another will happen soon after.

We model these sort of settings through queueing systems in which the arrivals occur according to a Hawkes process and in which service times follow phase-type distributions. This general type of service allows for accurate and robust modeling while preserving key characteristics for queues, such as the Markov property. Mathematically, this work is most similar to recent work by Gao and Zhu (2016) and Koops et al. (2017). Moreover, transient moments for infinite server queues with Markovian arrivals are also among the findings in Koops et al. (2017), an independent and concurrent work. However the moments in Koops et al. (2017) are only derived for exponential service distributions, whereas we give expressions for any phase-type service distribution. Additionally, we analyze the Hawkes/D/∞ queue and give an explicit analysis for its first two moments. Conceptually, our motivation is most similar to Debo et al. (2012). While the model in Debo et al. (2012) is similar to this one in concept, it is quite different in its probabilistic structure. Rather than using a Hawkes process for the arrivals, the authors model the scenario through a Poisson process with a probability of arrivals joining or balking that increases with the length of the queue. This describes the setting well, but there are a few limitations and room for additional considerations. For example, recency plays no role in the influence of the next arrival. For queues of identical length, that model considers the most recent arrival occurring a minute ago to be equivalent to it occurring an hour ago. Additionally, because events arrive according to a time-homogeneous

Poisson process and then either join or balk, the rate at which arrivals join the queue is bounded by the overall arrival rate, a constant. This prevents any kind of "viral" behavior for the events, so a large influx of arrivals over a short time is unlikely to occur. By contrast, these behaviors are inherent to our model. We will explore these ideas and others after the following descriptions of this paper's composition.

1.1. Main Contributions of Paper

In this paper, we provide exact expressions for the mean, variance, and covariance of the Hawkes process driven queue for all time, in both transient and steady state. These moments are derived for general phase-type service; we also provide examples for hyper-exponential and Erlang service. These results are derived by exploiting linear ordinary differential equations. We also derive expressions for all moments of the queue. We verify these functions via comparisons to simulations. We also derive a partial differential equation for the moment generating function and the cumulant moment generating function for the Hawkes/PH/∞ queue. We are able to show that the solution of the potentially high dimensional PDE for the MGF can be reduced to solving one differential equation, which does not have a closed form expression except in some special cases. Moreover, we analyze the Hawkes/D/∞ queue where the service times are deterministic. We derive exact expressions for the mean, variance, and auto-covariance of the queue length process. Throughout this work we show the relevance of the Hawkes process by direct comparison to the Poisson process and through novel applications. In our applications, we investigate the long run effects of the self-excitement structure, design an optimal control problem, and describe how to solve it numerically.

1.2. Organization of Paper

The remainder of this paper is organized into three main sections. In Section 2, we give an overview of results and properties in the Hawkes process literature that are relevant to this work and we then investigate the infinite server Hawkes process driven queue with deterministic service. In Section 3, we perform the main analysis of this work, which is the investigation of infinite server queues with Hawkes process arrivals and phase-type distributed service. In doing so, we first provide model definitions and technical lemmas, then derive expressions for the moments of the queue, followed by the auto-covariance and moment and cumulant generating functions. In Section 4, we apply this work to two novel settings, trending web traffic and night clubs. To facilitate comprehension of subject-specific notation, we provide the following table of terminology. Listed by order of appearance, these terms are also stated and defined at their first use. Thus, this reference is simply intended as an aid for the reader. In particular, we draw attention to this paper's use of $\mathbf{v}$ to represent the vector of all ones. While such a vector may be more commonly

denoted as $\mathbf{e}$, we avoid that notation as these vectors are frequently used near matrix exponentials since $\mathbf{v}$ is more distinct from e than $\mathbf{e}$ is.

Symbol	Definition
N_t	Hawkes counting process, the self-exciting point process
λ_t	Hawkes process intensity, represents the excitement of the process at time t
α	Hawkes process jump parameter, represents the jump in intensity upon an arrival
β	Hawkes process decay parameter, governs the exponential decrease of λ_t
λ^*	Hawkes process baseline intensity
λ_0	Initial value of λ_t
λ_∞	Equal to $\frac{\beta\lambda^*}{\beta-\alpha}$, represents the limit of the mean intensity as $t\to\infty$
Q_t	Queueing system, where $Q_{t,i}$ is the number in phase i of service at time t
S	Phase-type distribution transient state sub-generator matrix, represents the exponentially distributed rate of transitions of an entity from one phase of service to another with state 0 designated as the absorbing state for the end of the entity's service. Off diagonal elements are μ_{ij} and diagonal elements are $-\mu_i$
μ_{ij}	Transition rate from phase i to phase j where $i\neq j$
μ_i	Overall transition rate out of phase i
θ	Queueing system initial distribution of arrivals over the n phases of service
$\mathbf{v}$	The n-dimensional vector of all ones
$\mathbf{v}_i$	The n-dimensional vector of all zeros other than the value 1 at the i^{th} element
$\mathbf{V}_i$	The $n\times n$ matrix with one at (i,i) and zero otherwise

2. Hawkes Arrival Process

The Hawkes process, introduced in Hawkes (1971), is a self-exciting point process whose arrival intensity is dependent on the point process sample path. This is defined through the following dependence on the intensity process λ_t:

$$\mathbb{P}(N_{t+h}-N_t=1|\mathcal{F}_t)=\lambda_t\cdot h+o(h) \tag{1}$$

$$\mathbb{P}(N_{t+h}-N_t>1|\mathcal{F}_t)=o(h) \tag{2}$$

$$\mathbb{P}(N_{t+h}-N_t=0|\mathcal{F}_t)=1-\lambda_t\cdot h+o(h) \tag{3}$$

where $\mathcal{F}_t$ is a filtration on the underlying probability space $(\Omega,\mathcal{F},\mathbb{P})$ generated by $(N_t)_{t\geq 0}$. The arrival intensity is governed by the following stochastic dynamics:

$$d\lambda_t=\beta(\lambda^*-\lambda_t)dt+\alpha dN_t. \tag{4}$$

Here λ^* represents an underlying stationary arrival rate called the baseline intensity, $\alpha>0$ is the height of the jump in the intensity upon an arrival, and $\beta>0$ describes the decay of the intensity as time passes after an arrival. That is, when the number of arrivals N_t increases by one, the arrival intensity will jump up by amount α, and this increases the probability of another jump occurring.

This is why the Hawkes process is called self-exciting: its prior activity increases the likelihood of its future activity. However, as soon as an arrival occurs the intensity begins to decay exponentially at rate β to the baseline intensity λ^*. Because of the jumps and the decay, the arrivals tend to cluster. If one applies Ito's lemma to the kernel function $e^{-\beta t}\lambda_t$, then one can show that

$$\lambda_t = \lambda^* + e^{-\beta t}(\lambda_0 - \lambda^*) + \alpha \int_0^t e^{-\beta(t-s)} dN_s, \tag{5}$$

as in Da Fonseca and Zaatour (2014), which also discusses the impact of the initial value of the intensity λ_0. This process is known to be stable for $\alpha < \beta$, see Laub et al. (2015). Additionally, it is Markovian when conditioned on the present value of both the counting process and the intensity, which is also given in Laub et al. (2015). For the rest of this study we will restrict our setting to this exponential kernel assumption. When we use the term "Hawkes process" we assume that it has such a kernel. Before proceeding with a review of relevant Hawkes process results from the literature, we motivate the use of this process by showing both its similarities and its differences with the Poisson process.

2.1. Comparison to the Poisson Process

In Equation 5, note that if $\alpha = 0$ and $\lambda_0 = \lambda^*$ then $\lambda_t = \lambda^*$ for all t. In this case, the Hawkes process is equivalent to a stationary Poisson process with rate λ^*. However, if $\alpha = 0$ but $\lambda_0 \neq \lambda^*$ it is equivalent to a non-stationary Poisson process. So, conceptually, a Poisson process is a Hawkes process without excitement. Furthermore, a Hawkes process with $\lambda_0 = \lambda^*$ is in essence a stationary Poisson process until the first arrival occurs. However, once an arrival occurs the intensity process jumps by an amount α from the initial value and then begins to decay towards the baseline rate according to the exponential decay rate β. This is demonstrated in the example in Figure 1 below. This simulation, in addition to all the others throughout this work, is constructed by use of the algorithm described in Ogata (1981).

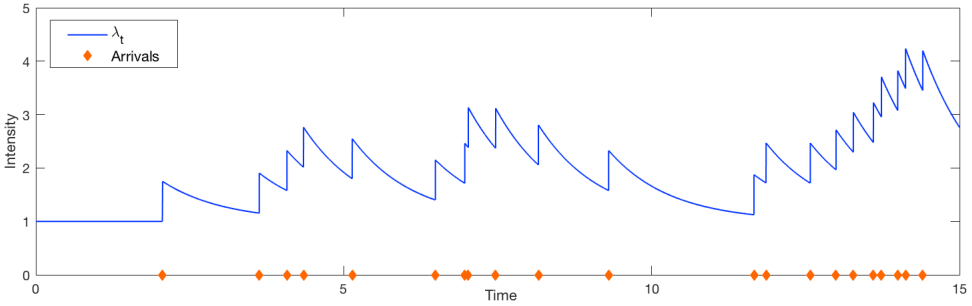

Figure 1 **Simulated λ_t, where $\alpha = \frac{3}{4}$, $\beta = 1$, and $\lambda^* = 1$.**

This example also shows another key difference between the Hawkes and Poisson processes. Because the self-excitation increases the likelihood of an arrival occurring soon after another, the

Hawkes process tends to cluster arrivals together across time. This means that the variance of a Hawkes process will be larger than that of a Poisson process, which is known to be equal to its mean. Below we demonstrate this through simulated limit distributions of the Hawkes process compared with the known Poisson probability mass function (PMF), each with the same mean.

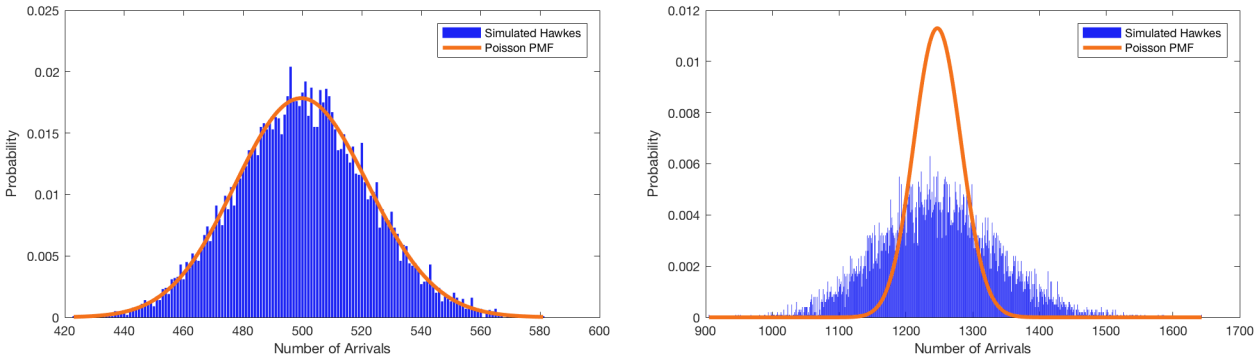

Figure 2 Limit Distributions for $\lambda^* = \beta = 1$ and $\alpha = 0$ (left) and 0.6 (right).

The simulated results are based on 10,000 replications, each with an end time of 500. As described previously, the two processes are equivalent for $\alpha = 0$. However, as α increases, the similarity between the Hawkes process and the Poisson process starts to disappear. Through these examples, we observe that the Hawkes process behaves quite differently from the Poisson process since it has heavier tails and therefore, is more variable. Thus, this provides theoretical motivation for our following investigation.

2.2. Review of Relevant Hawkes Process Literature

We now review a brief selection of Hawkes process results that support our following analysis of Hawkes process driven queueing systems. These results can be found in greater detail in Dassios and Zhao (2011), Da Fonseca and Zaatour (2014, 2015), as discussed specifically after each result statement. This review is primarily focused on the transient and stationary moments of the Hawkes process, and is included both for the sake of completeness and understanding of the problem, but also so that it may be incorporated later in this work. In the first statement, Proposition 1, differential equations for the moments of the Hawkes process are provided.

PROPOSITION 1. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 4, then we have the following differential equations for the moments of N_t and λ_t,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[N_t^m\right] = \sum_{j=0}^{m-1}\binom{m}{j}\mathrm{E}\left[\lambda_t N_t^j\right] \tag{6}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m\right] = m\beta\lambda^*\mathrm{E}\left[\lambda_t^{m-1}\right] - m\beta\mathrm{E}\left[\lambda_t^m\right] + \sum_{j=0}^{m-1}\binom{m}{j}\alpha^{m-j}\mathrm{E}\left[\lambda_t^{j+1}\right] \tag{7}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m N_t^l\right] = m\beta\lambda^*\mathrm{E}\left[\lambda_t^{m-1}N_t^l\right] - m\beta\mathrm{E}\left[\lambda_t^m N_t^l\right] + \sum_{(j,k)\in S}\binom{m}{j}\binom{l}{k}\alpha^{m-j}\mathrm{E}\left[\lambda_t^{j+1}N_t^k\right] \tag{8}$$

where $S = (\{0,\dots,m\} \times \{0,\dots,l\}) \setminus \{(m,l)\}$.

Proof. This follows directly from the approach involving the infinitesimal generator described in Sections 2.1 and 2.2 of Da Fonseca and Zaatour (2014), followed by simplification using the binomial theorem. For the first and second moments of N_t and λ_t and the first product moment, these equations are stated exactly in that work. □

As has been observed in the literature, the differential equations for the moments form a system of linear ordinary differential equations that have explicit solutions. We now provide the exact dynamics of the first two moments of the Hawkes process since this is of particular relevance to our later analysis. We also define notation that will be used throughout the remainder of this work.

PROPOSITION 2. *Given a Hawkes process* $X_t = (\lambda_t, N_t)$ *with dynamics given by Equation 4 with* $\alpha < \beta$*, then the mean, variance, and covariance of* N_t *and* λ_t *are provided by the following equations for all* $t \geq 0$*,*

$$\mathrm{E}[\lambda_t] = \lambda_\infty + (\lambda_0 - \lambda_\infty)\, e^{-(\beta-\alpha)t} \tag{9}$$

$$\mathrm{E}[N_t] = \lambda_\infty t + \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) \tag{10}$$

$$\mathrm{Var}(\lambda_t) = \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)} + \frac{\alpha^2(\lambda_0-\lambda_\infty)}{\beta-\alpha}e^{-(\beta-\alpha)t} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}e^{-2(\beta-\alpha)t} \tag{11}$$

$$\begin{aligned}\mathrm{Var}(N_t) = {} & \frac{\beta^2\lambda_\infty}{(\beta-\alpha)^2}t + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)^3}\left(1 - e^{-2(\beta-\alpha)t}\right) - \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{(\beta-\alpha)^2}te^{-(\beta-\alpha)t} \\ & + \left(\frac{\beta+\alpha}{(\beta-\alpha)^2}(\lambda_0-\lambda_\infty) - \frac{2\alpha\beta}{(\beta-\alpha)^3}\lambda_\infty\right)(1 - e^{-(\beta-\alpha)t})\end{aligned} \tag{12}$$

$$\begin{aligned}\mathrm{Cov}[\lambda_t, N_t] = {} & \left(\frac{\alpha\lambda_\infty}{\beta-\alpha} + \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)^2}\right)\left(1 - e^{-(\beta-\alpha)t}\right) + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)^2}\left(e^{-2(\beta-\alpha)t} - e^{-(\beta-\alpha)t}\right) \\ & + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}te^{-(\beta-\alpha)t}\end{aligned} \tag{13}$$

where

$$\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}.$$

Proof. The proof of this result can be found in Section 3.4 of Dassios and Zhao (2011) (as a particular case where $\rho = 0$) and in Section 3.2 of Da Fonseca and Zaatour (2015), or by solving the corresponding ODE system stated above Proposition 1. □

By further observation of Proposition 2 or simply by further review of the references in this section, the steady-state behavior of various Hawkes process statistics is also available. These expressions are stated in the following corollary.

COROLLARY 1. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 4 with $\alpha < \beta$, then the steady state values of the mean and variance of the intensity and of the covariance between the intensity and the counting process are as follows:*

$$\lim_{t\to\infty} \mathrm{E}\left[\lambda_t\right] = \frac{\beta\lambda^*}{\beta-\alpha} = \lambda_\infty, \tag{14}$$

$$\lim_{t\to\infty} \mathrm{Var}\left(\lambda_t\right) = \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)}, \tag{15}$$

$$\lim_{t\to\infty} \mathrm{Cov}\left[\lambda_t, N_t\right] = \frac{\alpha\lambda_\infty}{\beta-\alpha} + \frac{\alpha^2\lambda_\infty}{2(\beta-\alpha)^2}. \tag{16}$$

In Proposition 2 and Corollary 1, we assume that $\alpha < \beta$, which is a known stability condition in the literature, as detailed in Laub et al. (2015). However, we can also consider the case where $\alpha \geq \beta$ and investigate the behavior of the system through its transient mean values. This is performed in the following corollary.

COROLLARY 2. *Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 4 with $\alpha \geq \beta$, the transient mean intensity and transient mean of the counting process for $t \geq 0$ are*

$$\mathrm{E}\left[\lambda_t\right] = \frac{\beta\lambda^*}{\alpha-\beta}\left(e^{(\alpha-\beta)t} - 1\right) + \lambda_0 e^{(\alpha-\beta)t} \tag{17}$$

$$\mathrm{E}\left[N_t\right] = \left(\frac{\beta\lambda^*}{(\alpha-\beta)^2} + \frac{\lambda_0}{\alpha-\beta}\right)\left(e^{(\alpha-\beta)t} - 1\right) - \frac{\beta\lambda^*}{\alpha-\beta}t \tag{18}$$

when $\alpha > \beta$, and

$$\mathrm{E}\left[\lambda_t\right] = \beta\lambda^* t + \lambda_0 \tag{19}$$

$$\mathrm{E}\left[N_t\right] = \frac{\beta\lambda^*}{2}t^2 + \lambda_0 t \tag{20}$$

when $\alpha = \beta$.

As is stated in the stability condition, we see that the limits of these functions as t goes to infinity diverge for $\alpha \geq \beta$. The effect of the relationship of α and β on the system can be observed in the following graph.

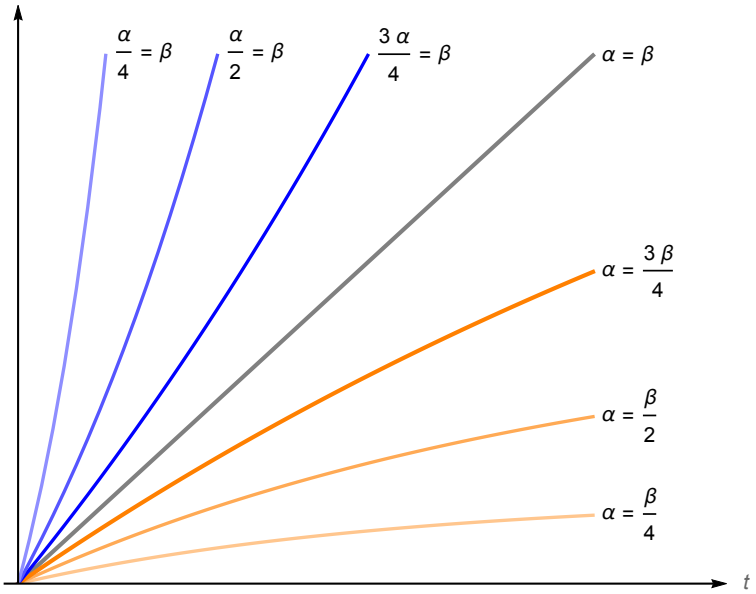

Figure 3 **Transient Mean Intensity for $\alpha < \beta$, $\alpha = \beta$, and $\alpha > \beta$.**

For the majority of this work we will consider settings in which the arrival process is stable and so we will assume $\alpha < \beta$. However, there are settings in which the transient behavior of the unstable arrival process is of interest, and so in our analysis of the queueing system we will also explore the mean behavior of queues under such arrival conditions.

2.3. $Hawkes/D/\infty$ Queue

Before moving on to the phase-type distributed service systems, we will first investigate the deterministic service setting. Since we have a good understanding about the Hawkes process itself, we can leverage our knowledge to analyze the $Hawkes/D/\infty$ queue where D is deterministic and is equal to the exact amount of time each customer spends in service. We exploit the fact that the $Hawkes/D/\infty$ queue can be written as the difference between the Hawkes process evaluated at time t and the Hawkes process evaluated at time $t - D$ i.e.

$$Q_t = N_t - N_{t-D}. \tag{21}$$

This representation of the $Hawkes/D/\infty$ queue leads us to a theorem that provides explicit expressions for the mean, variance, and auto-covariance of the $Hawkes/D/\infty$ queueing process. However, before we state the result, we need a lemma that describes the transient auto-covariance of the Hawkes process. This lemma will be extremely useful for our future calculations of other quantities of interest for the $Hawkes/D/\infty$ queue.

LEMMA 1. *Let N_t be a Hawkes process with dynamics given by Equation 4 with $\alpha < \beta$ and suppose N_t is initialized at zero. If we define $\mathcal{C}(t,\tau)$ as the following function*

$$\mathcal{C}(t,\tau) \equiv \text{Cov}[N_t, N_{t-\tau}], \tag{22}$$

then

$$\begin{aligned}\mathcal{C}(t,\tau) = &\frac{\alpha\left(1-e^{-(\beta-\alpha)\tau}\right)}{2(\beta-\alpha)^3}\left((2\beta-\alpha)\lambda_\infty - 2e^{-(\beta-\alpha)(t-\tau)}\left(\alpha\lambda_0 + \beta(\lambda_\infty-\lambda_0)(\beta-\alpha)(t-\tau) + (\beta-\alpha)\lambda_\infty\right)\right)\\ &+\left(\lambda_\infty + \frac{2\alpha\lambda_\infty}{\beta-\alpha} + \frac{\alpha^2\lambda_\infty}{(\beta-\alpha)^2}\right)(t-\tau) + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)^3}\left(1-e^{-(\beta-\alpha)(2t-\tau)}\right) - \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{(\beta-\alpha)^2}\\ &\cdot(t-\tau)e^{-(\beta-\alpha)(t-\tau)} + \left(\frac{\beta+\alpha}{(\beta-\alpha)^2}(\lambda_0-\lambda_\infty) - \frac{2\alpha\beta}{(\beta-\alpha)^3}\lambda_\infty\right)\left(1-e^{-(\beta-\alpha)(t-\tau)}\right)\end{aligned} \tag{23}$$

for all $t \geq \tau \geq 0$; otherwise $\mathcal{C}(t,\tau) = 0$.

Proof. To see this, we manipulate the definition of the auto-covariance to find an expression in terms of other known functions. Starting from the definition of covariance, we have

$$\text{Cov}\left[N_t, N_{t-\tau}\right] = \text{E}\left[N_t N_{t-\tau}\right] - \text{E}\left[N_t\right]\text{E}\left[N_{t-\tau}\right]$$

and by Proposition 2 we have expressions for $\text{E}[N_t]$ and $\text{E}[N_{t-\tau}]$. Thus, we focus on $\text{E}[N_t N_{t-\tau}]$. However, for brevity's sake we do not yet substitute these known expressions into the equation. By the tower property, we have that

$$\mathcal{C}(t,\tau) = \text{E}\left[\text{E}\left[N_t N_{t-\tau} \mid \mathcal{F}_{t-\tau}\right]\right] - \text{E}\left[N_t\right]\text{E}\left[N_{t-\tau}\right]$$

where $\mathcal{F}_{t-\tau}$ is the filtration of the Hawkes process up to time $t-\tau$. Through this conditioning, $N_{t-\tau}$ is known in the inner expectation, and so we can replace $\text{E}[\text{E}[N_t N_{t-\tau} \mid \mathcal{F}_{t-\tau}]]$ with $\text{E}[\text{E}[N_t \mid \mathcal{F}_{t-\tau}]N_{t-\tau}]$. Then, again by Proposition 2 we have that $\text{E}[N_t \mid \mathcal{F}_{t-\tau}] = \lambda_\infty\tau + \frac{\lambda_{t-\tau}-\lambda_\infty}{\beta-\alpha}\left(1-e^{-(\beta-\alpha)\tau}\right) + N_{t-\tau}$. Making use of this, we now have that

$$\begin{aligned}\mathcal{C}(t,\tau) = &\lambda_\infty\tau\text{E}\left[N_{t-\tau}\right] + \text{E}\left[\lambda_{t-\tau}N_{t-\tau}\right]\frac{1-e^{-(\beta-\alpha)\tau}}{\beta-\alpha} - \frac{\lambda_\infty}{\beta-\alpha}\text{E}\left[N_{t-\tau}\right]\\ &\cdot\left(1-e^{-(\beta-\alpha)\tau}\right) + \text{E}\left[N_{t-\tau}^2\right] - \text{E}\left[N_t\right]\text{E}\left[N_{t-\tau}\right],\end{aligned}$$

and by the definitions of covariance and variance this is equivalent to

$$\begin{aligned}\mathcal{C}(t,\tau) = &\lambda_\infty\tau\text{E}\left[N_{t-\tau}\right] + \frac{\text{Cov}\left[\lambda_{t-\tau}, N_{t-\tau}\right] + \text{E}\left[\lambda_{t-\tau}\right]\text{E}\left[N_{t-\tau}\right]}{\beta-\alpha}\left(1-e^{-(\beta-\alpha)\tau}\right)\\ &-\frac{\lambda_\infty}{\beta-\alpha}\text{E}\left[N_{t-\tau}\right]\left(1-e^{-(\beta-\alpha)\tau}\right) - \text{E}\left[N_t\right]\text{E}\left[N_{t-\tau}\right] + \text{Var}\left(N_{t-\tau}\right) + \text{E}\left[N_{t-\tau}\right]^2.\end{aligned}$$

Here we can recognize that each term in this expression has a known form from Proposition 2. Hence, by substituting these expressions and simplifying, we achieve the stated result. □

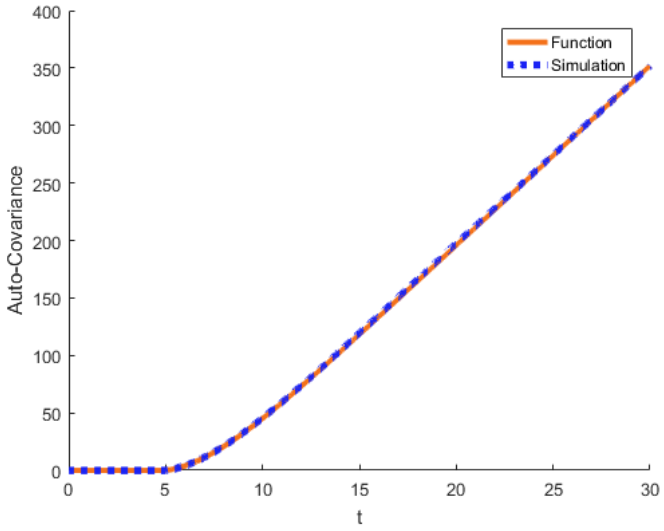

Figure 4 **Auto-covariance of the Hawkes Process with $D = 5$, $\lambda^* = 1$, $\alpha = \frac{3}{4}$, and $\beta = \frac{5}{4}$.**

With the expression for the transient auto-covariance of the Hawkes process in hand, we can now give explicit forms of the mean, variance, and auto-covariance of the $Hawkes/D/\infty$ queue.

THEOREM 1. *The transient mean of the $Hawkes/D/\infty$ when $\alpha < \beta$ is given by the following expression*

$$\mathbb{E}[Q_t] = \begin{cases} \lambda_\infty t + \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) & \text{if } t \le D, \\ \lambda_\infty D + \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha}\left(e^{-(\beta-\alpha)(t-D)} - e^{-(\beta-\alpha)t}\right) & \text{if } t > D. \end{cases} \tag{24}$$

Thus, in steady state the mean queue length is

$$\mathbb{E}[Q_\infty] = \lambda_\infty D. \tag{25}$$

Moreover, the transient variance of the $Hawkes/D/\infty$ queue is given by the following expression

$$\mathrm{Var}[Q_t] = \begin{cases} \mathcal{C}(t,0) & \text{if } t \le D, \\ \mathcal{C}(t,0) + \mathcal{C}(t-D,0) - 2\mathcal{C}(t,D) & \text{if } t > D. \end{cases} \tag{26}$$

Lastly, the transient auto-covariance of the $Hawkes/D/\infty$ queue is given by the following expression when $\tau \ge D$,

$$\mathrm{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \le \tau, \\ \mathcal{C}(t,\tau) - \mathcal{C}(t-D,\tau-D) & \text{if } \tau < t \le \tau + D \\ \mathcal{C}(t,\tau) + \mathcal{C}(t-D,\tau) - \mathcal{C}(t,\tau+D) - \mathcal{C}(t-D,\tau-D) & \text{if } \tau + D < t \end{cases} \tag{27}$$

and when $\tau < D$, then

$$\mathrm{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \le \tau, \\ \mathcal{C}(t,\tau) & \text{if } \tau < t \le D, \\ \mathcal{C}(t,\tau) - \mathcal{C}(t-\tau,D-\tau) & \text{if } D < t \le \tau + D \\ \mathcal{C}(t,\tau) + \mathcal{C}(t-D,\tau) - \mathcal{C}(t,\tau+D) - \mathcal{C}(t-\tau,D-\tau) & \text{if } \tau + D < t. \end{cases} \tag{28}$$

Proof. Throughout this proof we make use of the form of the auto-covariance of N_t given in Lemma 1. The transient mean is straightforward since it follows from the linearity property of expectation and just taking the difference of the two means. Moreover, for the variance we have

$$\begin{aligned}\text{Var}[Q_t] &= \text{Var}[N_t - N_{t-D}] \\ &= \text{Var}[N_t] + \text{Var}[N_{t-D}] - 2\text{Cov}[N_t, N_{t-D}] \\ &= \text{Var}[N_t] + \text{Var}[N_{t-D}] - 2\mathcal{C}(t, D) \\ &= \mathcal{C}(t,0) + \mathcal{C}(t-D,0) - 2\mathcal{C}(t,D).\end{aligned}$$

Finally for the auto-covariance, if $\tau \geq D$ we have that

$$\text{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \leq \tau, \\ \text{Cov}\left[N_t - N_{t-D}, N_{t-\tau}\right] & \text{if } \tau < t \leq \tau + D \\ \text{Cov}\left[N_t - N_{t-D}, N_{t-\tau} - N_{t-\tau-D}\right] & \text{if } \tau + D < t \end{cases}$$

by the definition of the $Hawkes/D/\infty$ queue and from the linearity of covariance. Now, for $\tau < D$, we have that

$$\text{Cov}[Q_t, Q_{t-\tau}] = \begin{cases} 0 & \text{if } t \leq \tau, \\ \text{Cov}\left[N_t, N_{t-\tau}\right] & \text{if } \tau < t \leq D, \\ \text{Cov}\left[N_t - N_{t-D}, N_{t-\tau}\right] & \text{if } D < t \leq \tau + D \\ \text{Cov}\left[N_t - N_{t-D}, N_{t-\tau} - N_{t-\tau-D}\right] & \text{if } \tau + D < t. \end{cases}$$

Again by the definition of the deterministic, Hawkes-driven, infinite server queue and the linearity of covariance, we achieve the stated result. □

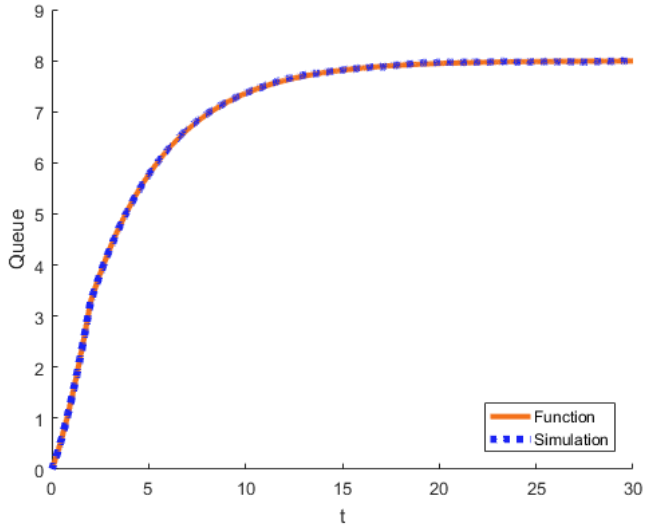

Figure 5 **Mean of the $Hawkes/D/\infty$ Queue with $D = 5$, $\lambda^* = 1$, $\alpha = \frac{3}{4}$, and $\beta = \frac{5}{4}$.**

3. $Hawkes/PH/\infty$ Queue

In this section, we will explore queueing systems in which arrivals occur according to a Hawkes process. This section is organized in the following manner. In Subsection 3.1, we provide key model definitions such as the phase-type distribution and we detail technical lemmas that support our analysis. Next, in Subsection 3.2, we derive differential equations for all moments of the queueing system and solve for exact expressions for the first and second moments. In Subsection 3.3, we consider the stationary limits of queues with stable arrival processes and investigate the transient behavior of those with unstable arrivals. Afterwards, we consider the auto-covariance of the queue in Subsection 3.4. Finally, in Subsection 3.5 we derive partial differential equations for the moment generating function and the cumulant moment generating function for this system.

3.1. Model Definitions and Technical Lemmas

To begin, we define the phase-type distribution. This form of service, formally defined below, can be thought of as a sequence of sub-services that have independent and exponentially distributed durations. We use this primarily for two factors. The first is that this is more general than just exponential service, and it can be shown that phase-type distributions can approximate any non-negative continuous distribution, see Cox (1955). Secondly, because the phase-type distribution is comprised of independent exponential service times, a queueing system with such service distributions is Markovian. Thus, these two properties together give us a system that is both flexible in application and practical in terms of analysis. A phase-type distribution with n phases represents the time taken from an initial state to an absorbing state of a continuous time Markov chain (CTMC) with the following infinitesimal generator matrix,

$$\Gamma = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{s} & \mathbf{S} \end{bmatrix}.$$

Here $\mathbf{0}$ is a $1 \times n$ zero vector, $\mathbf{s}$ is an $n \times 1$ vector, and $\mathbf{S}$ is an $n \times n$ matrix. Note $\mathbf{s} = -\mathbf{S}\mathbf{v}$ where $\mathbf{v}$ is an $n \times 1$ vector of ones. The matrix $\mathbf{S}$ and the initial distribution θ, which is a $1 \times n$ vector, identify the phase-type distributions. The number of phases in $\mathbf{S}$ is n. The matrix $\mathbf{S}$ and vector $\mathbf{s}$ can be expressed as:

$$\mathbf{S} = \begin{bmatrix} -\mu_1 & \cdots & \mu_{1,n} \\ \vdots & \ddots & \vdots \\ \mu_{n,1} & \cdots & -\mu_n \end{bmatrix}, \quad \mathbf{s} = (\mu_{1,0}, \ldots, \mu_{n,0})^{\mathrm{T}}, \tag{29}$$

where the μ_{ij}'s agree with the definition of the infinitesimal generator matrix $\mathbf{\Gamma}$. For notational consistency, we use a term *phase* to indicate the state of CTMC of the phase-type distributions throughout this paper. Additionally, we now note that in all following use of the matrix S we will not use a bold notation as in those settings additional emphasis that it is a matrix is not necessary.

With the phase-type distributions as described above, we build a Markovian queueing model referred to as the $Hawkes/PH/\infty$ queue. We assume that the system starts with no customers and that there are infinitely many servers. Further, we suppose that there are n phases of service and the transition rate between two distinct phases i and j is μ_{ij}. Let $\theta \in [0,1]^n$ be a distribution over the phases such that the probability that an arriving entity joins the i^{th} phase is θ_i, with $\sum_{i=1}^n \theta_i = 1$. An entity departs the system at rate μ_{i0}, where i is the entity's phase of service before leaving. For brevity of notation, define $\mu_i \equiv \mu_{i0} + \mu_{i1} + \cdots + \mu_{i,i-1} + \mu_{i,i+1} + \mu_{i,n}$. Let $Q_t \in \mathbb{N}^n$ represent the number in the queueing system, with $Q_{t,i}$ representing the number in phase i of service i.e.

$$Q_t = \sum_{i=1}^n Q_{t,i}\mathbf{v}_i \tag{30}$$

where $\mathbf{v}_i$ is the unit column vector in the i^{th} coordinate. We let (λ_t, N_t) represent a Hawkes process as described in Equation 4. We will now find the infinitesimal generator for real valued functions of the state space, $f : \mathbb{R}^+ \times \mathbb{N} \times \mathbb{N}^n \to \mathbb{R}$. For simplicity of notation, when describing the difference in values of f for changed arguments we will only list the variables that change, rather than listing all n queueing phase variables. This generator is shown below.

$$\begin{aligned}\mathcal{L}f(x) = &\underbrace{\beta(\lambda^* - \lambda_t)\frac{\partial f(x)}{\partial \lambda_t}}_{\text{Excitation Decay}} + \underbrace{\sum_{i=1}^n \lambda_t \theta_i \left(f(\lambda_t + \alpha, N_t + 1, Q_{t,i} + 1) - f(x)\right)}_{\text{Arrivals}} \\ &\underbrace{+ \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \mu_{ij} Q_{t,i} \left(f(\lambda_t, N_t, Q_{t,i} - 1, Q_{t,j} + 1) - f(x)\right)}_{\text{Transfers}} + \underbrace{\sum_{i=1}^n \mu_{i0} Q_{t,i} \left(f(\lambda_t, N_t, Q_{t,i} - 1) - f(x)\right)}_{\text{Departures}}\end{aligned} \tag{31}$$

Here, x is an element of the state space $(\mathbb{R}^+ \times \mathbb{N} \times \mathbb{N}^n)$. We can use this to obtain Dynkin's formula for the full $Hawkes/PH/\infty$ queueing system. We have that

$$\mathrm{E}_t\left[f(X_s)\right] = f(X_t) + E_t\left[\int_t^s \mathcal{L}f(X_u)du\right], \tag{32}$$

where $X_t = (\lambda_t, N_t, Q_t)$. This gives rise to the following lemma.

Lemma 2. *Let f be a function such that Equation 32 holds. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(X_t)\right] = \mathrm{E}\left[\mathcal{L}f(X_t)\right]$$

for all $t \geq 0$.

Proof. This is achieved through use of Fubini's theorem and the fundamental theorem of calculus. Using Equation 32 we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(X_t)\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\left(f(X_0) + \mathrm{E}\left[\int_0^t \mathcal{L}f(X_u)\mathrm{d}u\right]\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\int_0^t \mathcal{L}f(X_u)\mathrm{d}u\right] = \frac{\mathrm{d}}{\mathrm{d}t}\int_0^t \mathrm{E}\left[\mathcal{L}f(X_u)\right]\mathrm{d}u = \mathrm{E}\left[\mathcal{L}f(X_t)\right]\end{aligned}$$

and this completes this proof. □

REMARK 1. It is important for the reader to recognize that this is equivalent to Dynkin's theorem. In most textbooks, Dynkin's theorem is proved for sufficiently differentiable and more importantly bounded functions. However, this assumption of boundedness can often be relaxed. In fact this relaxation of the boundedness is very common for extending results like Ito's lemma and the Feynman Kac formula for unbounded, but polynomial bounded functions. This is often extended by stopping the process when it hits a certain level by using stopping times. Then one applies the previous results for bounded functions and takes limits as the bound tends to infinity. For the interested reader, see Lemma 2 of Oelschlager (1984) for a proof.

Now, before using these differential equations to find explicit functions as we did previously, we will first introduce a series of technical lemmas to aid our analysis. These lemmas are presented without proof as they follow from standard approaches for matrix exponentials and integration. First, we give a form for the indefinite integral of the exponential of a non-singular matrix.

LEMMA 3. *Let $L \in \mathbb{R}^{n\times n}$ be invertible. Then, if the integral of e^{Lt} exists it can be expressed*

$$\int e^{Lt}\,\mathrm{d}t = L^{-1}e^{Lt} + c$$

where c is some constant of integration.

Proof. The proof follows from standard approaches. □

The second lemma now provides explicit forms for the definite integral from 0 to t of the product of an exponential of an invertible matrix, a vector, a scalar power of the variable of integration, and a scalar exponential function of the variable of integration.

LEMMA 4. *Let $L \in \mathbb{R}^{n\times n}$ be invertible, let $\nu \in \mathbb{R}^n$, let $\eta \in \mathbb{N}$, and let $\gamma \in \mathbb{R}$. Then, if $L+\gamma I$ is invertible,*

$$\int_0^t e^{Ls}\nu s^\eta e^{\gamma s}\,\mathrm{d}s = \sum_{k=0}^{\eta} \frac{\eta!}{(\eta-k)!}(-1)^k (L+\gamma I)^{-(k+1)} \left(e^{Lt}\nu t^{\eta-k}e^{\gamma t}\right) - \eta!(-1)^\eta (L+\gamma I)^{-(\eta+1)}\nu$$

for $t > 0$.

Proof. The proof follows from the preceding lemma, induction, and integration by parts. □

The next lemma is a quick demonstration of commutativity of the inverse of a matrix exponential and an inverse of the same matrix shifted in the direction of the identity.

LEMMA 5. *Let $A \in \mathbb{R}^{n\times n}$ be invertible and let $b, c \in \mathbb{R}$ be such that $cA + bI$ is also invertible. Then,*

$$e^{-A}(cA+bI)^{-1} = (cA+bI)^{-1} e^{-A}.$$

Proof. The proof follows from the definition of the matrix exponential. □

These lemmas now come together to give the general solution to differential equations of a certain form.

LEMMA 6. *Let $g(t) \in \mathbb{R}^n$ be a function described by the dynamics*

$$\dot{g}(t) = -Lg(t) + \sum_{i\in\mathcal{S}} \nu_i t^{\eta_i} e^{\gamma_i t}$$

with an initial condition of $g(0) = g_0$, where $L \in \mathbb{R}^{n\times n}$ is invertible and $\mathcal{S}$ is a finite index set such that $\nu_i \in \mathbb{R}^n$, $\eta_i \in \mathbb{N}$, and $\gamma_i \in \mathbb{R}$ for each $i \in \mathcal{S}$. Then, if $L + \gamma_i I$ is invertible for all $i \in \mathcal{S}$ the explicit function for $g(t)$ is given by

$$g(t) = \sum_{i\in\mathcal{S}} \sum_{k=0}^{\eta_i} \frac{\eta_i!(-1)^k}{(\eta_i-k)!} (L+\gamma_i I)^{-(k+1)} \left(\nu_i t^{\eta_i-k} e^{\gamma_i t}\right) - \eta_i!(-1)^{\eta_i} (L+\gamma_i I)^{-(\eta_i+1)} e^{-Lt}\nu_i + e^{-Lt} g_0$$

for all $t \geq 0$.

Proof. The proof follows from standard differential equation techniques and the three preceding lemmas. □

Now, before introducing one final lemma we first define a useful matrix. For $\gamma, c \in \mathbb{R}$, $\nu \in \mathbb{R}^n$, and $L \in \mathbb{R}^{n\times n}$, let $M_{\gamma,\nu,L}(t) \in \mathbb{R}^{n\times n}$ be such that

$$M_{\gamma,\nu,L}(t) = \int_0^t e^{(\gamma I - L^{\mathrm{T}})s} \nu\nu^{\mathrm{T}} e^{-Ls}\, \mathrm{d}s \tag{33}$$

for all $t \geq 0$. Element-wise, we can express this matrix after integration as

$$\left(M_{\gamma,\nu,L}(t)\right)_{i,j} = \begin{cases} \sum_{k=1}^n \sum_{l=1}^n \nu_k\nu_l \sum_{r=0}^\infty \sum_{w=0}^\infty \frac{(L^r)_{k,i}(L^w)_{l,j}}{\gamma^{r+w+1}} \binom{r+w}{r} \left(e^{\gamma t} \sum_{z=0}^{r+w} \frac{(-\gamma t)^z}{z!} - 1\right) & \text{if } \gamma \neq 0, \\ \sum_{k=1}^n \sum_{l=1}^n \nu_k\nu_l \sum_{r=0}^\infty \sum_{w=0}^\infty \frac{(L^r)_{k,i}(L^w)_{l,j} t^{r+w+1}}{r!w!(r+w+1)} & \text{if } \gamma = 0. \end{cases}$$

This function provides shorthand when integrating a particular function that otherwise does not produce a nice linear algebraic form. The difficulty of expressing this integral in matrix form stems from the fact that L and $\nu\nu^{\mathrm{T}}$ need not commute. With defining $M_{\gamma,\nu,L}(t)$ we circumvent this issue by integrating on the element-level, but if L and $\nu\nu^{\mathrm{T}}$ were to commute we could avoid this function entirely, as we will later see. For now, this definition leads us to our next lemma.

LEMMA 7. *Let $\eta, \gamma, c \in \mathbb{R}$, $\nu \in \mathbb{R}^n$, $L \in \mathbb{R}^{n\times n}$ be such that L, $\gamma I + L$, and $(\eta+1)\gamma I - L$ are each invertible. Then,*

$$\begin{aligned}
&\int_0^t \Bigg(\left((\eta+1)\gamma I - L^{\mathrm{T}}\right)^{-1} \left(e^{(\eta\gamma I - L^{\mathrm{T}})s} - e^{-\gamma I s}\right) \nu\nu^{\mathrm{T}} c e^{-Ls} + e^{-L^{\mathrm{T}}s}\nu\nu^{\mathrm{T}} c \left(e^{(\eta\gamma I - L)s} - e^{-\gamma I s}\right) \\
&\quad \cdot \left((\eta+1)\gamma I - L\right)^{-1} \Bigg) \mathrm{d}s \\
=& c\left((\eta+1)\gamma I - L^{\mathrm{T}}\right)^{-1} \Bigg((\eta+2)\gamma M_{\eta\gamma,\nu,L}(t) + e^{(\eta\gamma I - L^{\mathrm{T}})t}\nu\nu^{\mathrm{T}} e^{-Lt} - \nu\nu^{\mathrm{T}} + \nu\nu^{\mathrm{T}} \left(e^{-(\gamma I + L)t} - I\right)(\gamma I + L)^{-1} \\
&\quad \cdot \left((\eta+1)\gamma I - L\right) + \left((\eta+1)\gamma I - L^{\mathrm{T}}\right)(\gamma I + L^{\mathrm{T}})^{-1} \left(e^{-(\gamma I + L^{\mathrm{T}})t} - I\right)\nu\nu^{\mathrm{T}} \Bigg) \left((\eta+1)\gamma I - L\right)^{-1}
\end{aligned}$$

for all $t \geq 0$.

Proof. The proof follows from the given definition of $M_{\gamma,\nu,L}(t)$, the product rule, and the preceding lemma. □

With these lemmas and definitions now in hand we can proceed to our analysis of the $Hawkes/PH/\infty$ queueing system. These results, stated in the following theorem, make use of the form of the infinitesimal generator in Lemma 2, with simplification through linearity of expectation and the binomial theorem.

3.2. Mean Dynamics of the $Hawkes/PH/\infty$ Queue

To begin investigation of the $Hawkes/PH/\infty$ queueing system, we first derive differential equations for the moments of the number in each phase of service and the intensity.

THEOREM 2. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 and phase-type distributed service. Then we have differential equations for the moments of $Q_{t,i}$ given by*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_{t,i}^m\right] =& \theta_i \sum_{g=0}^{m-1} \binom{m}{g} \mathrm{E}\left[\lambda_t Q_{t,i}^g\right] + \sum_{g=0}^{m-1}\sum_{\substack{j=1\\ j\neq i}}^{n} \binom{m}{g} \mu_{ji} \mathrm{E}\left[Q_{t,j} Q_{t,i}^g\right] \\
&+ \sum_{g=1}^{m} \binom{m}{g-1} \mu_i (-1)^{m-g+1} \mathrm{E}\left[Q_{t,i}^g\right],
\end{aligned} \tag{34}$$

for the products of $Q_{t,i}$ and $Q_{t,j}$ where $i \neq j$ given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_{t,i}^m Q_{t,j}^l\right] =& \theta_i \sum_{g=0}^{m-1} \binom{m}{g} \mathrm{E}\left[\lambda_t Q_{t,j}^l Q_{t,i}^g\right] + \theta_j \sum_{h=0}^{l-1} \binom{l}{h} \mathrm{E}\left[\lambda_t Q_{t,i}^m Q_{t,j}^h\right] \\
&+ \sum_{\substack{k=1\\ i\neq k\neq j}}^{n} \sum_{g=0}^{m-1} \binom{m}{g} \mu_{ki} \mathrm{E}\left[Q_{t,k} Q_{t,i}^g Q_{t,j}^l\right] + \sum_{\substack{k=1\\ j\neq k\neq i}}^{n} \sum_{h=0}^{l-1} \binom{l}{h} \mu_{kj} \mathrm{E}\left[Q_{t,k} Q_{t,i}^m Q_{t,j}^h\right]
\end{aligned} \tag{35}$$

$$+\mu_i\sum_{g=0}^{m-1}\binom{m}{g}(-1)^{m-g}\mathrm{E}\left[Q_{t,j}^{l}Q_{t,i}^{g+1}\right]+\mu_{ij}\sum_{g=0}^{m}\sum_{h=0}^{l-1}\binom{m}{g}\binom{l}{h}(-1)^{m-g}\mathrm{E}\left[Q_{t,i}^{g+1}Q_{t,j}^{h}\right]$$
$$+\mu_j\sum_{h=0}^{l-1}\binom{l}{h}(-1)^{l-h}\mathrm{E}\left[Q_{t,i}^{m}Q_{t,j}^{h+1}\right]+\mu_{ji}\sum_{h=0}^{l}\sum_{g=0}^{m-1}\binom{l}{h}\binom{m}{g}(-1)^{l-h}\mathrm{E}\left[Q_{t,j}^{h+1}Q_{t,i}^{g}\right],$$

and for the products of λ_t *and* $Q_{t,i}$ *given by*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l\right]=\beta\lambda^* m\mathrm{E}\left[\lambda_t^{m-1}Q_{t,i}^l\right]-\beta m\mathrm{E}\left[\lambda_t^m Q_{t,i}^l\right]+\theta_i\sum_{g=0}^{m}\sum_{h=0}^{l-1}\binom{m}{g}\binom{l}{h} \tag{36}$$
$$\cdot\alpha^{m-g}\mathrm{E}\left[\lambda_t^{g+1}Q_{t,i}^h\right]+\sum_{g=0}^{m-1}\binom{m}{g}\alpha^{m-g}\mathrm{E}\left[\lambda_t^{g+1}Q_{t,i}^l\right]+\mu_i\sum_{h=0}^{l-1}\binom{l}{h}$$
$$\cdot(-1)^{l-h}\mathrm{E}\left[\lambda_t^m Q_{t,i}^{h+1}\right]+\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{h=0}^{l-1}\binom{l}{h}\mu_{ji}\mathrm{E}\left[\lambda_t^m Q_{t,j}Q_{t,i}^h\right],$$

where $t\geq 0$.

Proof. We can first observe that each of these moments can be generalized to $\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right]$. From Lemma 2 we see that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right]=\mathrm{E}\Bigg[&\beta(\lambda^*-\lambda_t)m\lambda_t^{m-1}Q_{t,i}^l Q_{t,j}^k+\lambda_t\theta_i\left((\lambda_t+\alpha)^m(Q_{t,i}+1)^l Q_{t,j}^k-\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right)\\
&+\lambda_t\theta_j\left((\lambda_t+\alpha)^m Q_{t,i}^l(Q_{t,j}+1)^k-\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right)+\sum_{\substack{x=1\\j\neq x\neq i}}^{n}\lambda_t\theta_x Q_{t,i}^l Q_{t,j}^k\left((\lambda_t+\alpha)^m-\lambda_t^m\right)\\
&+\sum_{\substack{x=1\\i\neq x\neq j}}^{n}\mu_{xi}Q_{t,x}\lambda_t^m Q_{t,j}^k\left((Q_{t,i}+1)^l-Q_{t,i}^l\right)+\sum_{\substack{x=1\\j\neq k\neq i}}^{n}\mu_{xj}Q_{t,x}\lambda_t^m Q_{t,i}^l\left((Q_{t,j}+1)^k-Q_{t,j}^k\right)\\
&+\sum_{\substack{x=0\\i\neq x\neq j}}^{n}\mu_{ix}Q_{t,i}\lambda_t^m Q_{t,j}^k\left((Q_{t,i}-1)^l-Q_{t,i}^l\right)+\sum_{\substack{x=0\\j\neq x\neq i}}^{n}\mu_{jx}Q_{t,j}\lambda_t^m Q_{t,i}^l\left((Q_{t,j}-1)^k-Q_{t,j}^k\right)\\
&+\mu_{ij}Q_{t,i}\lambda_t^m\left((Q_{t,i}-1)^l(Q_{t,j}+1)^k-Q_{t,i}^l Q_{t,j}^k\right)+\mu_{ji}Q_{t,j}\lambda_t^m\left((Q_{t,j}-1)^k(Q_{t,i}+1)^l-Q_{t,i}^l Q_{t,j}^k\right)\Bigg]\end{aligned}$$

where we have combined the transfers from one phase to another and departures from that phase into the same summation by starting the index at 0. Using the binomial theorem and linearity of expectation, we have the following:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right]=\beta\lambda^* m\mathrm{E}\left[\lambda_t^{m-1}Q_{t,i}^l Q_{t,j}^k\right]-\beta m\mathrm{E}\left[\lambda_t^m Q_{t,i}^l Q_{t,j}^k\right]+\sum_{\substack{x=1\\j\neq x\neq i}}^{n}\sum_{y=0}^{m-1}\binom{m}{y}\theta_x\alpha^{m-y}$$
$$\cdot\mathrm{E}\left[\lambda_t^{y+1}Q_{t,i}^l Q_{t,j}^k\right]+\theta_i\left(\sum_{x=0}^{m}\sum_{y=0}^{l}\binom{m}{x}\binom{l}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1}Q_{t,i}^y Q_{t,j}^k\right]-\mathrm{E}\left[\lambda_t^{m+1}Q_{t,i}^l Q_{t,j}^k\right]\right)$$

$$+\theta_j\left(\sum_{x=0}^{m}\sum_{y=0}^{k}\binom{m}{x}\binom{k}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1}Q_{t,i}^{l}Q_{t,j}^{y}\right]-\mathrm{E}\left[\lambda_t^{m+1}Q_{t,i}^{l}Q_{t,j}^{k}\right]\right)+\sum_{\substack{x=1\\i\neq x\neq j}}^{n}\sum_{y=0}^{l-1}\binom{l}{y}\mu_{xi}$$
$$\cdot\mathrm{E}\left[\lambda_t^{m}Q_{t,x}Q_{t,i}^{y}Q_{t,j}^{k}\right]+\sum_{\substack{x=1\\i\neq x\neq j}}^{n}\sum_{y=0}^{k-1}\binom{k}{y}\mu_{xj}\mathrm{E}\left[\lambda_t^{m}Q_{t,x}Q_{t,i}^{l}Q_{t,j}^{y}\right]+\sum_{\substack{x=0\\i\neq x\neq j}}^{n}\sum_{y=0}^{l-1}\binom{l}{y}(-1)^{l-y}\mu_{ix}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{y+1}Q_{t,j}^{k}\right]$$
$$+\sum_{\substack{x=0\\i\neq x\neq j}}^{n}\sum_{y=0}^{k-1}\binom{k}{y}(-1)^{k-y}\mu_{jx}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{l}Q_{t,j}^{y+1}\right]+\mu_{ij}\left(\sum_{x=0}^{l}\sum_{y=0}^{k}\binom{l}{x}\binom{k}{y}(-1)^{l-x}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{x+1}Q_{t,j}^{y}\right]\right.$$
$$\left.-\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{l+1}Q_{t,j}^{k}\right]\right)+\mu_{ji}\left(\sum_{x=0}^{l}\sum_{y=0}^{k}\binom{l}{x}\binom{k}{y}(-1)^{k-y}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{x}Q_{t,j}^{y+1}\right]-\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{l}Q_{t,j}^{k+1}\right]\right).$$

Now we simplify by recognizing that $\sum_{x\neq j}\mu_{ix}=\mu_i-\mu_{ij}$ and $\sum_{i\neq x\neq j}\theta_x=1-\theta_i-\theta_j$. This leaves us with

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{l}Q_{t,j}^{k}\right]=\beta\lambda^{*}m\mathrm{E}\left[\lambda_t^{m-1}Q_{t,i}^{l}Q_{t,j}^{k}\right]-\beta m\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{l}Q_{t,j}^{k}\right]+\sum_{y=0}^{m-1}\binom{m}{y}\alpha^{m-y}\mathrm{E}\left[\lambda_t^{y+1}Q_{t,i}^{l}Q_{t,j}^{k}\right]$$
$$+\theta_i\sum_{x=0}^{m}\sum_{y=0}^{l-1}\binom{m}{x}\binom{l}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1}Q_{t,i}^{y}Q_{t,j}^{k}\right]+\theta_j\sum_{x=0}^{m}\sum_{y=0}^{k-1}\binom{m}{x}\binom{k}{y}\alpha^{m-x}\mathrm{E}\left[\lambda_t^{x+1}Q_{t,i}^{l}Q_{t,j}^{y}\right]$$
$$+\sum_{\substack{x=1\\i\neq x\neq j}}^{n}\sum_{y=0}^{l-1}\binom{l}{y}\mu_{xi}\mathrm{E}\left[\lambda_t^{m}Q_{t,x}Q_{t,i}^{y}Q_{t,j}^{k}\right]+\sum_{\substack{x=1\\i\neq x\neq j}}^{n}\sum_{y=0}^{k-1}\binom{k}{y}\mu_{xj}\mathrm{E}\left[\lambda_t^{m}Q_{t,x}Q_{t,i}^{l}Q_{t,j}^{y}\right]$$
$$+\mu_i\sum_{y=0}^{l-1}\binom{l}{y}(-1)^{l-y}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{y+1}Q_{t,j}^{k}\right]+\mu_{ij}\sum_{x=0}^{l}\sum_{y=0}^{k-1}\binom{l}{x}\binom{k}{y}(-1)^{l-x}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{x+1}Q_{t,j}^{y}\right]$$
$$+\mu_j\sum_{y=0}^{k-1}\binom{k}{y}(-1)^{k-y}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{l}Q_{t,j}^{y+1}\right]+\mu_{ji}\sum_{y=0}^{k}\sum_{x=0}^{l-1}\binom{l}{x}\binom{k}{y}(-1)^{k-y}\mathrm{E}\left[\lambda_t^{m}Q_{t,i}^{x}Q_{t,j}^{y+1}\right]$$

which is equivalent to each stated result when $m=k=0$, $k=0$, and $m=0$, respectively. □

We can now observe that we can form closed systems of linear ordinary differential equations from these equations. To do so, we restrict our focus to the equations for moments of combined power at most $m\in\mathbb{Z}^{+}$. Of course, the collection of equations that is of most practical interest is found by setting $m=2$, as this yields a system for the means and variances. This now gives rise to Corollary 3, which states the differential equations for the mean, variance, and covariances of queues driven by Hawkes processes.

Corollary 3. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 and phase-type distributed service. Then, we*

have the following differential equations for the mean, variance, and covariances of the number of entities in each phase and in the system as a whole:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_{t,i}\right] = \theta_i \mathrm{E}\left[\lambda_t\right] + \sum_{\substack{j=1 \\ j\neq i}}^{n} \mu_{ji}\mathrm{E}\left[Q_{t,j}\right] - \mu_i \mathrm{E}\left[Q_{t,i}\right] \tag{37}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Var}\left(Q_{t,i}\right) = {} & \theta_i \mathrm{E}\left[\lambda_t\right] + 2\theta_i \mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] + 2\sum_{\substack{j=1 \\ j\neq i}}^{n} \mu_{ji}\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] + \mu_i \mathrm{E}\left[Q_{t,i}\right] \\ & + \sum_{\substack{j=1 \\ j\neq i}}^{n} \mu_{ji}\mathrm{E}\left[Q_{t,j}\right] - 2\mu_i \mathrm{Var}\left(Q_{t,i}\right)\end{aligned} \tag{38}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] = {} & (\alpha - \beta - \mu_i)\mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] + \alpha\theta_i \mathrm{E}\left[\lambda_t\right] + \sum_{\substack{j=1 \\ j\neq i}}^{n} \mu_{ji}\mathrm{Cov}\left[\lambda_t, Q_{t,j}\right] \\ & + \theta_i \mathrm{Var}\left(\lambda_t\right)\end{aligned} \tag{39}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] = {} & \theta_i \mathrm{Cov}\left[\lambda_t, Q_{t,j}\right] + \theta_j \mathrm{Cov}\left[\lambda_t, Q_{t,i}\right] - (\mu_i + \mu_j)\mathrm{Cov}\left[Q_{t,i}, Q_{t,j}\right] \\ & + \sum_{\substack{k=1 \\ k\neq i}}^{n} \mu_{ki}\mathrm{Cov}\left[Q_{t,k}, Q_{t,j}\right] + \sum_{\substack{k=1 \\ k\neq i}}^{n} \mu_{kj}\mathrm{Cov}\left[Q_{t,k}, Q_{t,i}\right] - \mu_{ij}\mathrm{E}\left[Q_{t,i}\right] - \mu_{ji}\mathrm{E}\left[Q_{t,j}\right].\end{aligned} \tag{40}$$

We will find that it is quite useful to also be able to state the equations in Corollary 3 in linear algebraic form. Recall that the vector of the number in each phase of service is $Q_t \in \mathbb{N}^n$, the distribution of arrivals into phases is $\theta \in [0,1]^n$, and the sub-generator-matrix for the n phases of service is $S \in \mathbb{R}^{n\times n}$ so that $S_{i,i} = -\mu_i$ for each $i \in \{1,\ldots,n\}$ and $S_{i,j} = \mu_{i,j}$ for all $j \neq i$. We now also incorporate the notation $\mathrm{diag}\left(x\right) \in \mathbb{R}^{n\times n}$ for $x \in \mathbb{R}^n$ as $\mathrm{diag}\left(x\right) \equiv \sum_{i=1}^{n} \mathbf{V}_i x \mathbf{v}_i^{\mathrm{T}}$, where $\mathbf{v}_i \in \mathbb{R}^n$ is the unit column vector in the direction of the i^{th} coordinate and $\mathbf{V}_i = \mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}$, meaning that the i^{th} diagonal element is 1 and the rest are 0. Together, we have that the vector form of Equation 37 is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = \theta \mathrm{E}\left[\lambda_t\right] + S^{\mathrm{T}}\mathrm{E}\left[Q_t\right],$$

the vector form of Equation 39 is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[\lambda_t, Q_t\right] = \left(S^{\mathrm{T}} - (\beta - \alpha)I\right)\mathrm{Cov}\left[\lambda_t, Q_t\right] + \alpha\theta\mathrm{E}\left[\lambda_t\right] + \theta\mathrm{Var}\left(\lambda_t\right),$$

and the matrix form of Equations 38 and 40 is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}\left[Q_t, Q_t\right] = {} & S^{\mathrm{T}}\mathrm{Cov}\left[Q_t, Q_t\right] + \mathrm{Cov}\left[Q_t, Q_t\right]S + \theta\mathrm{Cov}\left[\lambda_t, Q_t\right]^{\mathrm{T}} + \mathrm{Cov}\left[\lambda_t, Q_t\right]\theta^{\mathrm{T}} \\ & + \mathrm{diag}\left(\theta\mathrm{E}\left[\lambda_t\right] + S^{\mathrm{T}}\mathrm{E}\left[Q_t\right]\right) - S^{\mathrm{T}}\mathrm{diag}\left(\mathrm{E}\left[Q_t\right]\right) - \mathrm{diag}\left(\mathrm{E}\left[Q_t\right]\right)S\end{aligned}$$

where the diagonal elements of the matrix $\mathrm{Cov}\left[Q_t, Q_t\right]$ correspond to the variance of the number in each phase of service and the off-diagonal elements represent the covariance between two phases of service. We can now use the technical lemmas in Subsection 3.1 to find explicit linear algebraic solutions to the closed system of differential equations in Corollary 3.

THEOREM 3. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 with $\alpha < \beta$ and phase-type distributed service. Let $S \in \mathbb{R}^{n\times n}$ be the sub-generator matrix for the transient states in the phase-distribution CTMC and let $\theta \in [0,1]^n$ be the initial distribution for arrivals to these states. If $S + (\beta-\alpha)I$ is invertible, then the vector of the mean number in service in each phase of service is*

$$\mathrm{E}\left[Q_t\right] = \lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta - \left(\lambda_0 - \lambda_\infty\right)\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta \tag{41}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$. Further, the vector of covariances between the intensity and each phase of service is

$$\begin{aligned}\mathrm{Cov}\left[\lambda_t, Q_t\right] &= \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\left(I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)t}\right)\theta - \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\\ &\cdot\left(S^{\mathrm{T}}\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)t}\right)\theta + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\\ &\cdot\left(e^{-2(\beta-\alpha)t}I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)t}\right)\theta\,.\end{aligned} \tag{42}$$

Finally, the matrix of covariances between phases of service is given by

$$\begin{aligned}\mathrm{Cov}\left[Q_t, Q_t\right] &= \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\Bigg(2(\beta-\alpha)e^{S^{\mathrm{T}}t}M_{0,\theta,S}(t)e^{St} + \theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{St}\\ &+ e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I - e^{St}\right)\left((\beta-\alpha)I + S\right)^{-1}\left((\beta-\alpha)I - S\right) + \left((\beta-\alpha)I - S^{\mathrm{T}}\right)\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1}\\ &\cdot\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\theta^{\mathrm{T}}e^{St}\Bigg)\left((\beta-\alpha)I - S\right)^{-1} + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\left(S^{\mathrm{T}}\right)^{-1}\Bigg((\beta-\alpha)e^{S^{\mathrm{T}}t}M_{-(\beta-\alpha),\theta,S}(t)e^{St}\\ &+ e^{-(\beta-\alpha)t}\theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{St} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I - e^{St}\right)\left((\beta-\alpha)I + S\right)^{-1}S - S^{\mathrm{T}}\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1}\\ &\cdot\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\theta^{\mathrm{T}}e^{St}\Bigg)S^{-1} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left((\beta-\alpha)I + S^{\mathrm{T}}\right)^{-1}\Bigg(e^{-2(\beta-\alpha)t}\theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{St}\\ &- e^{S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I - e^{St}\right) - \left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\theta^{\mathrm{T}}e^{St}\Bigg)\left((\beta-\alpha)I + S\right)^{-1} - \lambda_\infty\mathrm{diag}\Big(\left(S^{\mathrm{T}}\right)^{-1}\\ &\cdot\left(I - e^{S^{\mathrm{T}}t}\right)\theta\Big) - (\lambda_0-\lambda_\infty)\mathrm{diag}\left(\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\right)\end{aligned} \tag{43}$$

where all $t \geq 0$.

Proof. Throughout this proof we use the fact that a matrix being invertible implies that its transpose is invertible as well. To begin, we can see from Corollary 3 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = S^{\mathrm{T}}\mathrm{E}\left[Q_t\right] + \theta\mathrm{E}\left[\lambda_t\right] = S^{\mathrm{T}}\mathrm{E}\left[Q_t\right] + \theta\left(\lambda_\infty + \left(\lambda_0-\lambda_\infty\right)e^{-(\beta-\alpha)t}\right)$$

and so we apply Lemma 6. Let $\nu_1 = \theta\lambda_\infty$ and $\eta_1 = \gamma_1 = 0$, and let $\nu_2 = \theta(\lambda_0-\lambda_\infty)$, $\eta_2 = 0$, and $\gamma_2 = -(\beta-\alpha)$. We assume that the queue starts empty. Then, we have

$$\begin{aligned}\mathrm{E}\left[Q_t\right] = &-\left(S^{\mathrm{T}}\right)^{-1}\theta\lambda_\infty + \left(S^{\mathrm{T}}\right)^{-1}e^{-S^{\mathrm{T}}t}\theta\lambda_\infty - \left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\theta(\lambda_0-\lambda_\infty)e^{-(\beta-\alpha)t}\\ &+ \left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}e^{-S^{\mathrm{T}}t}\theta(\lambda_0-\lambda_\infty)\end{aligned}$$

which now simplifies to the stated result. Note that S is invertible because it is diagonally dominant by definition and we have assumed the invertibility of $S+(\beta-\alpha)I$, which implies non-singularity of the respective transposes. We find the stated result for $\text{Cov}\left[\lambda_t, Q_t\right]$ through repeating the same technique to the corresponding differential equation systems, where again we make use of the linear algebraic representation. Thus, we are left to solve for the covariance matrix. Note that from Corollary 3, the variance of each phase and the covariance between phases can form one linear algebraic form as the covariance matrix, as shown below.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\text{Cov}\left[Q_t, Q_t\right] = S^{\mathrm{T}}\text{Cov}\left[Q_t, Q_t\right] + \text{Cov}\left[Q_t, Q_t\right]S + \theta\text{Cov}\left[\lambda_t, Q_t\right]^{\mathrm{T}} + \text{Cov}\left[\lambda_t, Q_t\right]\theta^{\mathrm{T}} \\ + \text{diag}\left(\theta\text{E}\left[\lambda_t\right] + S^{\mathrm{T}}\text{E}\left[Q_t\right]\right) - S^{\mathrm{T}}\text{diag}\left(\text{E}\left[Q_t\right]\right) - \text{diag}\left(\text{E}\left[Q_t\right]\right)S\end{aligned}$$

Using the product rule and multiplying through by matrix exponentials on the right and left, we can also express this as below:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-S^{\mathrm{T}}t}\text{Cov}\left[Q_t, Q_t\right]e^{-St}\right) = e^{-S^{\mathrm{T}}t}\theta\text{Cov}\left[\lambda_t, Q_t\right]^{\mathrm{T}}e^{-St} + e^{-S^{\mathrm{T}}t}\text{Cov}\left[\lambda_t, Q_t\right]\theta^{\mathrm{T}}e^{-St} \\ + e^{-S^{\mathrm{T}}t}\text{diag}\left(\theta\text{E}\left[\lambda_t\right] + S^{\mathrm{T}}\text{E}\left[Q_t\right]\right)e^{-St} - e^{-S^{\mathrm{T}}t}S^{\mathrm{T}}\text{diag}\left(\text{E}\left[Q_t\right]\right)e^{-St} \\ - e^{-S^{\mathrm{T}}t}\text{diag}\left(\text{E}\left[Q_t\right]\right)Se^{-St}.\end{aligned}$$

For the pair of $\text{Cov}\left[\lambda_t, Q_t\right]$ terms, we use Lemma 7 in conjunction with the explicit function for $\text{Cov}\left[\lambda_t, Q_t\right]$ to find

$$\begin{aligned}&\int_0^t \left(e^{-S^{\mathrm{T}}s}\theta\text{Cov}\left[\lambda_s, Q_s\right]^{\mathrm{T}}e^{-Ss} + e^{-S^{\mathrm{T}}s}\text{Cov}\left[\lambda_s, Q_s\right]\theta^{\mathrm{T}}e^{-Ss}\right)\mathrm{d}s \\ &= \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\Bigg(2(\beta-\alpha)M_{0,\theta,S}(t) + e^{-S^{\mathrm{T}}t}\theta\theta^{\mathrm{T}}e^{-St} - \theta\theta^{\mathrm{T}} + \theta\theta^{\mathrm{T}}\left(e^{-((\beta-\alpha)I+S)t} - I\right) \\ &\cdot\left((\beta-\alpha)I+S\right)^{-1}\left((\beta-\alpha)I-S\right) + \left((\beta-\alpha)I - S^{\mathrm{T}}\right)\left((\beta-\alpha)I+S^{\mathrm{T}}\right)^{-1}\left(e^{-((\beta-\alpha)I+S^{\mathrm{T}})t} - I\right)\theta\theta^{\mathrm{T}}\Bigg) \\ &\cdot\left((\beta-\alpha)I-S\right)^{-1} + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\left(S^{\mathrm{T}}\right)^{-1}\Bigg((\beta-\alpha)M_{-(\beta-\alpha),\theta,S}(t) + e^{-((\beta-\alpha)I+S^{\mathrm{T}})t}\theta\theta^{\mathrm{T}}e^{-St} - \theta\theta^{\mathrm{T}} \\ &- \theta\theta^{\mathrm{T}}\left(e^{-((\beta-\alpha)I+S)t} - I\right)\left((\beta-\alpha)I+S\right)^{-1}S - S^{\mathrm{T}}\left((\beta-\alpha)I+S^{\mathrm{T}}\right)^{-1}\left(e^{-((\beta-\alpha)I+S^{\mathrm{T}})t} - I\right)\theta\theta^{\mathrm{T}}\Bigg)S^{-1} \\ &- \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left((\beta-\alpha)I+S^{\mathrm{T}}\right)^{-1}\Bigg(e^{-(2(\beta-\alpha)I+S^{\mathrm{T}})t}\theta\theta^{\mathrm{T}}e^{-St} - \theta\theta^{\mathrm{T}} - \theta\theta^{\mathrm{T}}\left(e^{-((\beta-\alpha)I+S)t} - I\right) \\ &- \left(e^{-((\beta-\alpha)I+S^{\mathrm{T}})t} - I\right)\theta\theta^{\mathrm{T}}\Bigg)\left((\beta-\alpha)I+S\right)^{-1}\end{aligned}$$

and so we now integrate the remaining terms in the covariance matrix differential equations. Note that the product rule for three terms is $(fgh)' = f'gh + fg'h + fgh'$. We have already used this

in concatenating the covariance matrix terms in the differential equation, and we can now make use of it again. Recall that $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] = S^{\mathrm{T}}\mathrm{E}\left[Q_t\right] + \theta\mathrm{E}\left[\lambda_t\right]$. Using this realization, the integral of the remaining three terms is

$$\begin{aligned}
&\int_0^t \left(e^{-S^{\mathrm{T}}s}\mathrm{diag}\left(\theta\mathrm{E}\left[\lambda_s\right] + S^{\mathrm{T}}\mathrm{E}\left[Q_s\right]\right)e^{-Ss} - e^{-S^{\mathrm{T}}s}S^{\mathrm{T}}\mathrm{diag}\left(\mathrm{E}\left[Q_s\right]\right)e^{-Ss} - e^{-S^{\mathrm{T}}s}\mathrm{diag}\left(\mathrm{E}\left[Q_s\right]\right)Se^{-Ss}\right)\mathrm{d}s \\
&= e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\mathrm{E}\left[Q_t\right]\right)e^{-St} \\
&= -e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\left(S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta\right)e^{-St}\lambda_\infty - e^{-S^{\mathrm{T}}t}\mathrm{diag}\left(\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\right)\theta\right) \\
&\qquad \cdot e^{-St}(\lambda_0 - \lambda_\infty)
\end{aligned}$$

where we are justified in moving the differentiation through the diagonalization and distributing it across sums via the definition of diagonalization as a linear combination. Combining this with the integral for the covariance between the queue and intensity and multiplying each side by the corresponding exponentials, we achieve the stated result. □

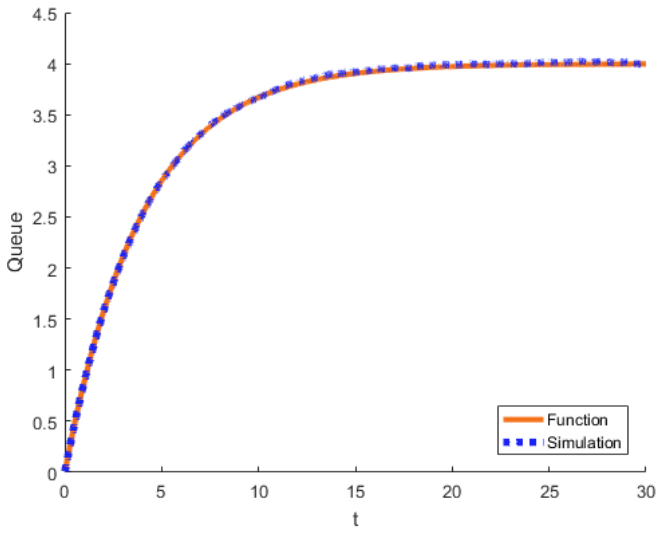

Figure 6 **Example Mean of the $Hawkes/PH/\infty$ Queue with Sub-Generator Matrix S_{Cox} as in Equation 44.**

As a brief example, consider a Hawkes process driven queueing system with infinite servers and suppose that the service is phase-type distributed with initial distribution $\theta = \mathbf{v}_1$ and the following sub-generator matrix:

$$S_{\mathrm{Cox}} = \begin{bmatrix} -4 & 3 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -3 & 2 & 0 \\ 0 & 0 & 0 & -5 & 4 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{44}$$

This is referred to as a Coxian distribution. It is characterized by each phase of service having an associated probability of either system departure or advancement to the next phase upon service completion. In this example, $\lambda^* = 1$, $\alpha = \frac{3}{4}$, and $\beta = 1$. The simulation is based on 100,000 replications.

REMARK 2. We now note that the assumed nonsingularity of $S + (\beta - \alpha)I$ is necessary to implement the technical lemmas, but need not hold in order for a closed form solution to exist. If these conditions do not hold, one can instead make use of the structure of invertibility that is implied by a specific phase-type distribution. In Corollaries 4 and 5, we demonstrate this for Erlang and hyper-exponential service, respectively. Like we have seen in Theorem 3, these expressions can be found through solving systems of differential equations provided by Corollary 3.

We start with the case of service times following a Erlang distribution. In this case, we define $N \in \mathbb{R}^{n\times n}$ as the matrix of all ones on the first lower diagonal and zeros otherwise. Then, $S^{\mathrm{T}} = n\mu(N - I)$ for this phase-type distribution. Observe that N is a nilpotent matrix of a particular structure: for $k \in \mathbb{N}$, N^k is the matrix of all ones on the k^{th} lower diagonal if $k \leq n-1$ and is the zero matrix otherwise. Additionally, in this case $\theta = \mathbf{v}_1$ as all arrivals occur in the first phase. With this in hand, we see that

$$\begin{aligned}\left(M_{\gamma,\mathbf{v}_1,n\mu(I-N^{\mathrm{T}})}(t)\right)_{i,j} &= \left(M_{\gamma+2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)\right)_{i,j} \\ &= \begin{cases} \binom{i+j-2}{i-1}(n\mu)^{i+j-2}\frac{e^{(\gamma+2n\mu)t}\sum_{k=0}^{i+j-2}\frac{(-(\gamma+2n\mu)t)^k}{k!}-1}{(\gamma+2n\mu)^{i+j-1}} & \text{if } \gamma+2n\mu \neq 0 \\ \frac{(tn\mu)^{i+j-1}}{n\mu(i-1)!(j-1)!(i+j-1)} & \text{if } \gamma+2n\mu = 0 \end{cases}\end{aligned}$$

and we make use of this in the following corollary.

COROLLARY 4. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 with $\alpha < \beta$ and Erlang distributed service with n phases and mean $\frac{1}{\mu}$. Then, when $n\mu \neq \beta - \alpha$, the vector of mean number in each phase of service is given by*

$$\mathrm{E}\left[Q_t\right] = \frac{\lambda_\infty}{n\mu}\left(I - e^{n\mu(N-I)t}\right)\mathbf{v} - \left(\lambda_0 - \lambda_\infty\right)\left(n\mu N - (n\mu - \beta + \alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)t}I - e^{n\mu(N-I)t}\right)\mathbf{v}_1, \quad (45)$$

and when $n\mu = \beta - \alpha$, this vector is

$$\mathrm{E}\left[Q_t\right] = \frac{\lambda_\infty}{n\mu}\left(I - e^{n\mu(N-I)t}\right)\mathbf{v} + \left(\lambda_0 - \lambda_\infty\right)e^{n\mu(N-I)t}x(t), \quad (46)$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$ and $x : \mathbb{R}^+ \to \mathbb{R}^n$ is such that $x_i(t) = \frac{(-n\mu)^{i-1}t^i}{i!}$. Further, when $n\mu \neq \beta - \alpha$ the vector of covariances between the number in each phase of service and the intensity is

$$\mathrm{Cov}\left[\lambda_t, Q_t\right] = \lambda_\infty\left(\alpha + \frac{\alpha^2}{2(\beta-\alpha)}\right)\left((n\mu + \beta - \alpha)I - n\mu N\right)^{-1}\left(I - e^{(n\mu N - (n\mu+\beta-\alpha)I)t}\right)\mathbf{v}_1$$

$$+\frac{\alpha\beta(\lambda_0-\lambda_\infty)}{n\mu(\beta-\alpha)}\left(e^{-(\beta-\alpha)t}I-e^{(n\mu N-(n\mu+\beta-\alpha)I)t}\right)\mathbf{v}+\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left(n\mu N-(n\mu-\beta+\alpha)I\right)^{-1}$$
$$\cdot\left(e^{-2(\beta-\alpha)t}I-e^{(n\mu N-(n\mu+\beta-\alpha)I)t}\right)\mathbf{v}_1, \tag{47}$$

and when $n\mu=\beta-\alpha$*, this is*

$$\mathrm{Cov}\left[\lambda_t,Q_t\right]=\lambda_\infty\left(\frac{\alpha}{n\mu}+\frac{\alpha^2}{2(n\mu)^2}\right)(2I-N)^{-1}\left(I-e^{n\mu(N-2I)t}\right)\mathbf{v}_1+(\lambda_0-\lambda_\infty)\left(\frac{\alpha}{n\mu}+\frac{\alpha^2}{(n\mu)^2}\right)$$
$$\cdot\left(e^{-n\mu t}I-e^{n\mu(N-2I)t}\right)\mathbf{v}-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2n\mu}e^{n\mu(N-2I)t}x(t). \tag{48}$$

Finally, when $n\mu\neq\beta-\alpha$*, the matrix of the covariance between the number in the phases of service is given by*

$$\mathrm{Cov}\left[Q_t,Q_t\right]=\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((n\mu+\beta-\alpha)I-n\mu N\right)^{-1}\Bigg(2(\beta-\alpha)e^{n\mu(N-I)t}M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)e^{n\mu(N^{\mathrm{T}}-I)t}$$
$$+\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}-e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}e^{n\mu(N^{\mathrm{T}}-I)t}+e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}\Big(e^{-(\beta-\alpha)t}I-e^{n\mu(N^{\mathrm{T}}-I)t}\Big)(n\mu N^{\mathrm{T}}-(n\mu-\beta+\alpha)I)^{-1}$$
$$\cdot\left((n\mu+\beta-\alpha)I-n\mu N^{\mathrm{T}}\right)+\left((n\mu+\beta-\alpha)I-n\mu N\right)(n\mu N-(n\mu-\beta+\alpha)I)^{-1}\left(e^{-(\beta-\alpha)t}I-e^{n\mu(N-I)t}\right)$$
$$\cdot\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}e^{n\mu(N^{\mathrm{T}}-I)t}\Bigg)\left((n\mu+\beta-\alpha)I-n\mu N^{\mathrm{T}}\right)^{-1}+\frac{\alpha\beta(\lambda_0-\lambda_\infty)}{(n\mu)^2(\beta-\alpha)}\left(N-I\right)^{-1}\Bigg((\beta-\alpha)e^{n\mu(N-I)t}$$
$$\cdot M_{2n\mu-\beta+\alpha,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)e^{n\mu(N^{\mathrm{T}}-I)t}+e^{-(\beta-\alpha)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}-e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}e^{n\mu(N^{\mathrm{T}}-I)t}-n\mu e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}$$
$$\cdot\Big(e^{-(\beta-\alpha)t}I-e^{n\mu(N^{\mathrm{T}}-I)t}\Big)\left(n\mu N^{\mathrm{T}}-(n\mu-\beta+\alpha)I\right)^{-1}(N^{\mathrm{T}}-I)-n\mu(N-I)(n\mu N-(n\mu-\beta+\alpha)I)^{-1}$$
$$\cdot\Big(e^{-(\beta-\alpha)t}I-e^{n\mu(N-I)t}\Big)\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}e^{n\mu(N^{\mathrm{T}}-I)t}\Bigg)(N^{\mathrm{T}}-I)^{-1}-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}(n\mu N-(n\mu-\beta+\alpha)I)^{-1}$$
$$\cdot\Bigg(e^{-2(\beta-\alpha)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}-e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}e^{n\mu(N^{\mathrm{T}}-I)t}-e^{n\mu(N-I)t}\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}\left(e^{-(\beta-\alpha)t}I-e^{n\mu(N^{\mathrm{T}}-I)t}\right)$$
$$-\left(e^{-(\beta-\alpha)t}I-e^{n\mu(N-I)t}\right)\mathbf{v}_1\mathbf{v}_1{}^{\mathrm{T}}e^{n\mu(N^{\mathrm{T}}-I)t}\Bigg)(n\mu N^{\mathrm{T}}-(n\mu-\beta+\alpha)I)^{-1}+\frac{\lambda_\infty}{n\mu}\mathrm{diag}\Big(\left(I-e^{n\mu(N-I)t}\right)\mathbf{v}\Big)$$
$$-(\lambda_0-\lambda_\infty)\mathrm{diag}\left((n\mu N-(n\mu-\beta+\alpha)I)^{-1}\left(e^{-(\beta-\alpha)t}I-e^{n\mu(N-I)t}\right)\mathbf{v}_1\right), \tag{49}$$

whereas when $n\mu=\beta-\alpha$*, this matrix is*

$$\mathrm{Cov}\left[Q_t,Q_t\right]=\mathrm{diag}\left(\frac{\lambda_\infty}{n\mu}\left(I-e^{n\mu(N-I)t}\right)\mathbf{v}+(\lambda_0-\lambda_\infty)\,e^{n\mu(N-I)t}x(t)\right)+e^{n\mu(N-I)t}\Bigg(\lambda_\infty\left(\frac{\alpha}{n\mu}+\frac{\alpha^2}{2(n\mu)^2}\right)$$
$$\cdot\Bigg(\left(M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)-x(t)\mathbf{v}_1{}^{\mathrm{T}}\right)\left(2I-N^{\mathrm{T}}\right)^{-1}+(2I-N)^{-1}\left(M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)-\mathbf{v}_1x^{\mathrm{T}}(t)\right)\Bigg)+(\lambda_0-\lambda_\infty)$$
$$\cdot\left(\frac{\alpha}{n\mu}+\frac{\alpha^2}{(n\mu)^2}\right)\Bigg(M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)\left(I-N^{\mathrm{T}}\right)^{-1}+(I-N)^{-1}M_{2n\mu,\mathbf{v}_1,n\mu N^{\mathrm{T}}}(t)-x(t)\mathbf{v}^{\mathrm{T}}-\mathbf{v}x^{\mathrm{T}}(t)\Bigg)$$
$$-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2n\mu}\left(X(t)+X^{\mathrm{T}}(t)\right)\Bigg)e^{n\mu(N^{\mathrm{T}}-I)t}, \tag{50}$$

where all $t\geq 0$ *and* $X:\mathbb{R}^+\to\mathbb{R}^{n\times n}$ *is such that* $X_{i,j}(t)=\frac{(-n\mu)^{i+j-2}t^{i+j-1}}{(i-1)!j!(i+j)}$.

As with the Erlang, we also provide explicit formulas for the hyper-exponential distribution. In this case we have that $S=-D$ where D is a diagonal matrix of the rates of service in each phase. This allows it to commute with the symmetric $\theta\theta^{\mathrm{T}}$, giving us

$$M_{\gamma,\theta,-D}(t)=\int_0^t e^{(\gamma I+D)s}\theta\theta^{\mathrm{T}}e^{Ds}\,\mathrm{d}s=\int_0^t e^{(\gamma I+2D)s}\,\mathrm{d}s\theta\theta^{\mathrm{T}}=(\gamma I+2D)^{-1}\left(e^{(\gamma I+2D)t}-I\right)\theta\theta^{\mathrm{T}}$$

as long as $\gamma I+2D$ is invertible. However, we also seek to address the case where $(\beta-\alpha)I+S=(\beta-\alpha)I-D$ is not invertible. In the hyper-exponential service setting, $(\beta-\alpha)I-D$ being singular implies that some $\mu_i=\beta-\alpha$, but it is not clear which or for how many μ_i this is the case. So, we instead use the element-level equations in Corollary 3 to solve for the explicit expressions. This method is preferable to the linear algebra approach for hyper-exponential service since in this setting $\mu_{ij}=0$ for every i and j.

Corollary 5. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t,N_t) with dynamics given in Equation 4 with $\alpha<\beta$ and hyper-exponential distributed service with n phases and distinct service rates $\mu_1,\ldots,\mu_n$. Then, the mean number in phase $i\in\{1,\ldots,n\}$ of service is*

$$\mathrm{E}\left[Q_{t,i}\right]=\begin{cases}\frac{\lambda_\infty}{\mu_i}\left(1-e^{-\mu_i t}\right)\theta_i+\frac{\lambda_0-\lambda_\infty}{\mu_i-\beta+\alpha}\left(e^{-(\beta-\alpha)t}-e^{-\mu_i t}\right)\theta_i & \text{if } \mu_i\neq\beta-\alpha,\\ \frac{\lambda_\infty}{\mu_i}\left(1-e^{-\mu_i t}\right)\theta_i+\left(\lambda_0-\lambda_\infty\right)\theta_i t e^{-\mu_i t} & \text{if } \mu_i=\beta-\alpha,\end{cases} \tag{51}$$

where $\lambda_\infty=\frac{\beta\lambda^*}{\beta-\alpha}$. Furthermore the covariance between the number in phase i of service and the intensity is

$$\mathrm{Cov}\left[\lambda_t,Q_{t,i}\right]=\begin{cases}\frac{\alpha\theta_i(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(1-e^{-(\mu_i+\beta-\alpha)t}\right)+\frac{\alpha\beta\theta_i(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\left(e^{-(\beta-\alpha)t}\right.\\ \quad\left.-e^{-(\mu_i+\beta-\alpha)t}\right)-\frac{\alpha^2\theta_i(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)}\left(e^{-2(\beta-\alpha)t}-e^{-(\mu_i+\beta-\alpha)t}\right) & \text{if } \mu_i\neq\beta-\alpha,\\ \frac{\alpha\theta_i(2\mu_i+\alpha)\lambda_\infty}{4\mu_i^2}\left(1-e^{-2\mu_i t}\right)+\frac{\alpha\beta\theta_i(\lambda_0-\lambda_\infty)}{\mu_i^2}\left(e^{-\mu_i t}-e^{-2\mu_i t}\right)\\ \quad-\frac{\alpha^2\theta_i(2\lambda_0-\lambda_\infty)}{2\mu_i}te^{-2\mu_i t} & \text{if } \mu_i=\beta-\alpha.\end{cases} \tag{52}$$

Then, the covariance between the number in phase i of service and the number in phase j of service where $i,j \in \{1,\dots,\}$ and $i \neq j$ is

$$\mathrm{Cov}\left[Q_{t,i},Q_{t,j}\right] = \begin{cases} \frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu_j+\beta-\alpha)}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j}-\frac{e^{-(\mu_j+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\beta+\alpha}\right) \\ \quad+\frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_j(\beta-\alpha)}\left(\frac{e^{-(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-\beta+\alpha}-\frac{e^{-(\mu_j+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\beta+\alpha}\right) \\ \quad-\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_j-\beta+\alpha)}\left(\frac{e^{2(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-2\beta+2\alpha}-\frac{e^{-(\mu_j+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\beta+\alpha}\right) \\ \quad+\frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j}-\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_j-\beta+\alpha}\right) \\ \quad+\frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\left(\frac{e^{-(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-\beta+\alpha}-\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_j-\beta+\alpha}\right) \\ \quad-\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)}\left(\frac{e^{2(\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j-2\beta+2\alpha}-\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-(\mu_i+\mu_j)t}}{\mu_j-\beta+\alpha}\right) & \text{if } \mu_i \neq \beta-\alpha \neq \mu_j, \\ \frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{4\mu_j^2}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j}-\frac{e^{-2\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\mu_j}\right)+\frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_j^2} \\ \quad\cdot\left(\frac{e^{-\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i}-\frac{e^{-2\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\mu_j}\right)-\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2\mu_j} \\ \quad\cdot\left(\frac{te^{-2\mu_i t}}{\mu_j-\mu_i}+\frac{e^{-(\mu_i+\mu_j)t}-e^{-2\mu_i t}}{(\mu_j-\mu_i)^2}\right)+\frac{\alpha\theta_i\theta_j(2\beta-\alpha)\lambda_\infty}{2\mu_j(\mu_i+\mu_j)}\left(\frac{1-e^{-(\mu_i+\mu_j)t}}{\mu_i+\mu_j}\right. \\ \quad\left.-te^{-(\mu_i+\mu_j)t}\right)+\frac{\alpha\beta\theta_i\theta_j(\lambda_0-\lambda_\infty)}{\mu_i\mu_j}\left(\frac{e^{-\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i}-te^{-(\mu_i+\mu_j)t}\right) \\ \quad-\frac{\alpha^2\theta_i\theta_j(2\lambda_0-\lambda_\infty)}{2\mu_j(\mu_i-\mu_j)}\left(\frac{e^{2\mu_j t}-e^{-(\mu_i+\mu_j)t}}{\mu_i-\mu_j}-te^{-(\mu_i+\mu_j)t}\right) & \text{if } \mu_i \neq \beta-\alpha=\mu_j, \end{cases} \tag{53}$$

Finally, the variance of the number in phase $i \in \{1,\dots,n\}$ of service is given by

$$\mathrm{Var}\left(Q_{t,i}\right) = \begin{cases} \frac{\lambda_\infty\theta_i}{\mu_i}\left(1-e^{-\mu_i t}\right)+\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{2\mu_i(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(1-e^{-2\mu_i t}\right)-\left(\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu_i+\beta-\alpha)}\right. \\ \quad\left.+\frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}-\frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu_i-\beta+\alpha)}\right)\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-2\mu_i t}}{\mu_i-\beta+\alpha}+\left((\lambda_0-\lambda_\infty)\theta_i\right. \\ \quad\left.+\frac{\mu_i(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha}+\frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\right)\frac{e^{-(\beta-\alpha)t}-e^{-2\mu_i t}}{2\mu_i-\beta+\alpha}-\frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)^2} \\ \quad\cdot\left(e^{-2(\beta-\alpha)t}-e^{-2\mu_i t}\right)-\frac{(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha}\left(e^{-\mu_i t}-e^{-2\mu_i t}\right) & \text{if } \mu_i \neq \beta-\alpha \neq 2\mu_i, \\ \frac{\lambda_\infty\theta_i}{\mu_i}\left(1-e^{-\mu_i t}\right)+\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{2\mu_i(\beta-\alpha)(\mu_i+\beta-\alpha)}\left(1-e^{-2\mu_i t}\right)-\left(\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu_i+\beta-\alpha)}\right. \\ \quad\left.+\frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}-\frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu_i-\beta+\alpha)}\right)\frac{e^{-(\mu_i+\beta-\alpha)t}-e^{-2\mu_i t}}{\mu_i-\beta+\alpha}+\left((\lambda_0-\lambda_\infty)\theta_i\right. \\ \quad\left.+\frac{\mu_i(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha}+\frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i(\beta-\alpha)}\right)te^{-2\mu_i t}-\frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu_i-\beta+\alpha)^2}\left(e^{-2(\beta-\alpha)t}\right. \\ \quad\left.-e^{-2\mu_i t}\right)-\frac{(\lambda_0-\lambda_\infty)\theta_i}{\mu_i-\beta+\alpha}\left(e^{-\mu_i t}-e^{-2\mu_i t}\right) & \text{if } 2\mu_i=\beta-\alpha, \\ \frac{\lambda_\infty\theta_i}{\mu_i}\left(1-e^{-\mu_i t}\right)+\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{4\mu_i^3}\left(1-e^{-2\mu_i t}\right)-\left(\frac{\alpha\theta_i^2(2\beta-\alpha)\lambda_\infty}{2\mu_i^2}\right. \\ \quad\left.+\frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i^2}\right)te^{-2\mu_i t}+\left((\lambda_0-\lambda_\infty)\theta_i+\frac{2\alpha\beta\theta_i^2(\lambda_0-\lambda_\infty)}{\mu_i^2}\right) \\ \quad\cdot\frac{e^{-\mu_i t}-e^{-2\mu_i t}}{\mu_i}-\frac{\alpha^2\theta_i^2(2\lambda_0-\lambda_\infty)}{2\mu_i}t^2e^{-2\mu_i t}+(\lambda_0-\lambda_\infty)\theta_i\left(\frac{te^{-\mu_i t}}{\mu_i}\right. \\ \quad\left.+\frac{e^{-2\mu_i t}-e^{-\mu_i t}}{\mu_i^2}\right) & \text{if } \mu_i=\beta-\alpha, \end{cases} \tag{54}$$

where all $t \geq 0$.

We now note that in both Corollary 4 and Corollary 5, taking $n=1$ reduces the setting to exponential service. We demonstrate the simplification and use of the singe-phase expressions in

finding the auto-covariance of the $Hawkes/M/\infty$ queue, shown in Proposition 3. We also note that these findings compare quite nicely to simulations in numerical demonstrations. In Subsection 3.6, we provide several example figures of these equations and their simulated counterparts.

Now that we have investigated the transient behavior of the $Hawkes/PH/\infty$ queue for a variety of settings it is natural to consider the behavior of the system in steady-state. This, along with the behavior of the system with an unstable arrival process, is the focus of the next subsection.

3.3. Limiting Behavior of the $Hawkes/PH/\infty$ Queue

In many situations, the steady-state behavior of a queueing system may be of particular interest. With that in mind, we now investigate the mean and variance of the $Hawkes/PH/\infty$ queue as time goes to infinity.

COROLLARY 6. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 and phase-type distributed service. Let $S \in \mathbb{R}^{n\times n}$ be the sub-generator matrix for the transient states in the phase-distribution CTMC and let $\theta \in [0,1]^n$ be the initial distribution for arrivals to these states. Then, the steady-state mean number in each phase of service is given by the vector*

$$\mathcal{Q}_\infty \equiv \lim_{t\to\infty} \mathrm{E}\left[Q_t\right] = \lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}\theta \tag{55}$$

where $\lambda_\infty = \frac{\beta\lambda^*}{\beta-\alpha}$. Further, the vector of steady-state covariances between the number in each phase of service and the intensity is

$$\mathcal{C}_\infty \equiv \lim_{t\to\infty} \mathrm{Cov}\left[\lambda_t, Q_t\right] = \lambda_\infty \frac{\alpha(2\beta-\alpha)}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\theta\,. \tag{56}$$

Finally, the matrix of steady-state covariances between each phase of service $\lim_{t\to\infty}\mathrm{Cov}\left[Q_t, Q_t\right]$, denoted $\mathcal{V}_\infty$, is given by the solution to the Lyapunov equation

$$S^{\mathrm{T}}\mathcal{V}_\infty + \mathcal{V}_\infty S + \mathcal{M} = 0 \tag{57}$$

where $\mathcal{M} = \theta\mathcal{C}_\infty^{\mathrm{T}} + \mathcal{C}_\infty\theta^{\mathrm{T}} - S^{\mathrm{T}}\mathrm{diag}\left(\mathcal{Q}_\infty\right) - \mathrm{diag}\left(\mathcal{Q}_\infty\right)S$. If S is symmetric, then $\mathcal{V}_\infty = -\frac{1}{2}S^{-1}\mathcal{M}$.

Proof. The proof follows by either taking the limit of the equations in Theorem 3 or setting the corresponding differential equations to 0 and finding the equilibrium solution. □

REMARK 3. We note that in steady-state the invertibility conditions from Theorem 3 are no longer necessary. We can further observe that these equations reveal an interesting relationship

among these steady-state values for the case of single phase service. For μ as the rate of exponential service, Corollary 6 yields

$$\mathcal{V}_\infty = \mathcal{Q}_\infty + \frac{1}{\mu}\mathcal{C}_\infty = \frac{\lambda_\infty}{\mu}\left(1 + \frac{\alpha(2\beta-\alpha)}{2(\beta-\alpha)(\mu+\beta-\alpha)}\right). \tag{58}$$

Thus, we have that the steady-state variance of the number in system for the $Hawkes/M/\infty$ queue is equal to the mean number in system plus the expected service duration times the steady-state covariance between the number in system and the intensity. Thus this provides an explicit contrast with Poisson-driven queues, as the steady-state distribution of a $M/M/\infty$ system is known to be Poisson distributed with rate equal to the steady-state mean number in system. This implies that the steady-state variance for such a queue is equal to its steady-state mean, unlike the relationship we observe for the $Hawkes/M/\infty$ system in Equation 58.

However, as we have noted, if $\alpha \geq \beta$ the Hawkes process is unstable and so steady-state analysis of the queue will not apply. Thus, in this scenario we instead investigate the transient behavior of the mean of the queue under the unstable arrival process.

COROLLARY 7. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 with $\alpha \geq \beta$ and phase-type distributed service. Let $S \in \mathbb{R}^{n\times n}$ be the sub-generator matrix for the transient states in the phase-distribution CTMC and let $\theta \in [0,1]^n$ be the initial distribution for arrivals to these states. Then the vector of mean number in service in each phase of service is given by*

$$\begin{aligned}\mathrm{E}\left[Q_t\right] = &\left((\alpha-\beta)I - S^{\mathrm{T}}\right)^{-1}\left(e^{(\alpha-\beta)t}I - e^{S^{\mathrm{T}}t}\right)\theta\left(\frac{\beta\lambda^*}{\alpha-\beta} + \lambda_0\right)\\ &+ (S^{\mathrm{T}})^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta\frac{\beta\lambda^*}{\alpha-\beta}\end{aligned} \tag{59}$$

when $\alpha > \beta$ and

$$\mathrm{E}\left[Q_t\right] = -(S^{\mathrm{T}})^{-1}\left(I - e^{S^{\mathrm{T}}t}\right)\theta(\lambda_0 - \beta\lambda^*) - (S^{\mathrm{T}})^{-1}\theta\beta\lambda^* t \tag{60}$$

when $\alpha = \beta$.

3.4. Auto-covariance of the $Hawkes/PH/\infty$ Queue

We now consider the auto-covariance of the number in this queueing system, $Q_t \in \mathbb{R}^n$. Analogous to the auto-covariance for the number of arrivals from the Hawkes process discussed in Subsection 2.3, this matrix quantity is defined as

$$\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] = \mathrm{E}\left[Q_t Q_{t-\tau}^{\mathrm{T}}\right] - \mathrm{E}\left[Q_t\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}$$

where $t \geq \tau \geq 0$ and otherwise the covariance is equal to 0. For an infinite server queue with Hawkes process arrivals and phase-type distributed service, the findings in Subsection 3.2 give us expressions for $\mathrm{E}[Q_t]$ and $\mathrm{E}[Q_{t-\tau}]$. Let $\mathcal{F}_s$ be the filtration of the queueing system, the Hawkes process, and the intensity at time $s \geq 0$. Then, assuming $S + (\beta - \alpha)I$ is invertible, conditional expectation yields

$$\begin{aligned}
\mathrm{E}\left[Q_t Q_{t-\tau}^{\mathrm{T}}\right] &= \mathrm{E}\left[\mathrm{E}\left[Q_t \mid F_{t-\tau}\right] Q_{t-\tau}^{\mathrm{T}}\right] \\
&= \mathrm{E}\Big[\Big(\lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}\tau}\right)\theta - \left(\lambda_{t-\tau} - \lambda_\infty\right)\left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)\tau}I \\
&\quad - e^{S^{\mathrm{T}}\tau}\right)\theta + e^{S^{\mathrm{T}}\tau} Q_{t-\tau}\Big) Q_{t-\tau}^{\mathrm{T}}\Big] \\
&= \lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}\tau}\right)\theta \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} - \left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)\tau}I - e^{S^{\mathrm{T}}\tau}\right)\theta \\
&\quad \cdot \left(\mathrm{E}\left[\lambda_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right] - \lambda_\infty \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}\right) + e^{S^{\mathrm{T}}\tau}\mathrm{E}\left[Q_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right]
\end{aligned}$$

by application of the expression for the vector of the mean number in each phase given in Theorem 3, modified to start at time $t - \tau$. Upon recognizing that $\mathrm{E}\left[\lambda_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right] = \mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right] + \mathrm{E}\left[\lambda_{t-\tau}\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}$ and $\mathrm{E}\left[Q_{t-\tau} Q_{t-\tau}^{\mathrm{T}}\right] = \mathrm{Cov}\left[Q_{t-\tau}, Q_{t-\tau}^{\mathrm{T}}\right] + \mathrm{E}\left[Q_{t-\tau}\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}$, we have that

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}\tau}\right)\theta \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} - \left(S^{\mathrm{T}} + (\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)\tau}I - e^{S^{\mathrm{T}}\tau}\right) \\
&\quad \cdot \theta\left(\mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right]^{\mathrm{T}} + \mathrm{E}\left[\lambda_{t-\tau}\right]\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} - \lambda_\infty \mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}}\right) + e^{S^{\mathrm{T}}\tau}\mathrm{Cov}\left[Q_{t-\tau}, Q_{t-\tau}\right] \\
&\quad + \left(e^{S^{\mathrm{T}}\tau}\mathrm{E}\left[Q_{t-\tau}\right] - \mathrm{E}\left[Q_t\right]\right)\mathrm{E}\left[Q_{t-\tau}\right]^{\mathrm{T}} \qquad (61)
\end{aligned}$$

and that each term in this expression can be calculated by applying Theorem 3. An explicit expression for the transient auto-covariance using this approach is given in the Appendix. In this section we give an explicit expression for the auto-covariance of the $Hawkes/M/\infty$ queue. In this setting with service rate μ, the same approach as above yields

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu\tau}\right)\mathrm{E}\left[Q_{t-\tau}\right] + e^{-\mu\tau}\mathrm{Var}\left(Q_{t-\tau}\right) + \mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right]\frac{e^{-(\beta-\alpha)\tau} - e^{-\mu\tau}}{\mu - \beta + \alpha} \\
&\quad + \left(\mathrm{E}\left[\lambda_{t-\tau}\right] - \lambda_\infty\right)\mathrm{E}\left[Q_{t-\tau}\right]\frac{e^{-(\beta-\alpha)\tau} - e^{-\mu\tau}}{\mu - \beta + \alpha} + e^{-\mu\tau}\mathrm{E}\left[Q_{t-\tau}\right]^2 - \mathrm{E}\left[Q_t\right]\mathrm{E}\left[Q_{t-\tau}\right] \qquad (62)
\end{aligned}$$

when $\mu \neq \beta - \alpha$ and

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu\tau}\right)\mathrm{E}\left[Q_{t-\tau}\right] + e^{-\mu\tau}\mathrm{Var}\left(Q_{t-\tau}\right) + \mathrm{Cov}\left[\lambda_{t-\tau}, Q_{t-\tau}\right]\tau e^{-\mu\tau} \\
&\quad + \left(\mathrm{E}\left[\lambda_{t-\tau}\right] - \lambda_\infty\right)\mathrm{E}\left[Q_{t-\tau}\right]\tau e^{-\mu\tau} + e^{-\mu\tau}\mathrm{E}\left[Q_{t-\tau}\right]^2 - \mathrm{E}\left[Q_t\right]\mathrm{E}\left[Q_{t-\tau}\right] \qquad (63)
\end{aligned}$$

when $\mu = \beta - \alpha$, where each of these makes use of Corollary 5 with $n = 1$, $\theta_1 = 1$, and $\mu_i = \mu$. These expressions are made explicit in the following proposition.

PROPOSITION 3. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 with $\alpha < \beta$ and exponentially distributed service with rate μ. Then, for $t \geq \tau \geq 0$ the auto-covariance of the number in system is*

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \frac{\lambda_\infty}{\mu}\left(1-e^{-\mu\tau}\right)\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+\frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-(\beta-\alpha)(t-\tau)}-e^{-\mu(t-\tau)}\right)\right)+\frac{\lambda_\infty}{\mu} \\
&\cdot\left(e^{-\mu\tau}-e^{-\mu t}\right)+\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2\mu(\beta-\alpha)(\mu+\beta-\alpha)}\left(e^{-\mu\tau}-e^{-\mu(2t-\tau)}\right)-\left(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu+\beta-\alpha)}+\frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\right. \\
&\left.-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu-\beta+\alpha)}\right)\frac{e^{-(\mu+\beta-\alpha)t+(\beta-\alpha)\tau}-e^{-\mu(2t-\tau)}}{\mu-\beta+\alpha}+\left(\lambda_0-\lambda_\infty+\frac{\mu(\lambda_0-\lambda_\infty)}{\mu-\beta+\alpha}+\frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\right) \\
&\cdot h(t-\tau)e^{-\mu\tau}-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu-\beta+\alpha)^2}\left(e^{-2(\beta-\alpha)t-(\mu-2\beta+2\alpha)\tau}-e^{-\mu(2t-\tau)}\right)-\frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-\mu t}\right. \\
&\left.-e^{-\mu(2t-\tau)}\right)+e^{-\mu\tau}\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+\frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-(\beta-\alpha)(t-\tau)}-e^{-\mu(t-\tau)}\right)\right)^2+\frac{e^{-(\beta-\alpha)\tau}-e^{-\mu\tau}}{\mu-\beta+\alpha} \\
&\cdot\left(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu+\beta-\alpha)}\left(1-e^{-(\mu+\beta-\alpha)(t-\tau)}\right)+\frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\left(e^{-(\beta-\alpha)(t-\tau)}-e^{-(\mu+\beta-\alpha)(t-\tau)}\right)\right. \\
&\left.-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu-\beta+\alpha)}\left(e^{-2(\beta-\alpha)(t-\tau)}-e^{-(\mu+\beta-\alpha)(t-\tau)}\right)\right)+(\lambda_0-\lambda_\infty)\frac{e^{-(\beta-\alpha)\tau}-e^{-\mu\tau}}{\mu-\beta+\alpha}\left(\frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\right. \\
&\cdot\left(e^{-2(\beta-\alpha)(t-\tau)}-e^{-(\mu+\beta-\alpha)(t-\tau)}\right)+\left.\frac{\lambda_\infty}{\mu}\left(e^{-(\beta-\alpha)(t-\tau)}-e^{-(\mu+\beta-\alpha)(t-\tau)}\right)\right)-\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu t}\right)+\frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\right. \\
&\left.\cdot\left(e^{-(\beta-\alpha)t}-e^{-\mu t}\right)\right)\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+\frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-(\beta-\alpha)(t-\tau)}-e^{-\mu(t-\tau)}\right)\right) \qquad (64)
\end{aligned}$$

when $\mu \neq \beta-\alpha$ and

$$\begin{aligned}
\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] &= \frac{\lambda_\infty}{\mu}\left(1-e^{-\mu\tau}\right)\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+(\lambda_0-\lambda_\infty)\left(t-\tau\right)e^{-\mu(t-\tau)}\right)+\frac{\lambda_\infty}{\mu}\left(e^{-\mu\tau}-e^{-\mu t}\right) \\
&+\frac{\alpha(2\beta-\alpha)\lambda_\infty}{4\mu^3}\left(e^{-\mu\tau}-e^{-\mu(2t-\tau)}\right)-\left(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2\mu^2}+\frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu^2}\right)(t-\tau)e^{-\mu(2t-\tau)}+\left(\lambda_0-\lambda_\infty\right. \\
&\left.+\frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu^2}\right)\frac{e^{-(\beta-\alpha)t-(\mu-\beta+\alpha)\tau}-e^{-\mu(2t-\tau)}}{\mu}-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2\mu}(t-\tau)^2e^{-\mu(2t-\tau)}+(\lambda_0-\lambda_\infty) \\
&\cdot\left(\frac{(t-\tau)e^{-\mu t}}{\mu}+\frac{e^{-\mu(2t-\tau)}-e^{-\mu t}}{\mu^2}\right)+e^{-\mu\tau}\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+(\lambda_0-\lambda_\infty)\left(t-\tau\right)e^{-\mu(t-\tau)}\right)^2 \\
&+\left(\frac{\alpha(2\mu+\alpha)\lambda_\infty}{4\mu^2}\left(1-e^{-2\mu(t-\tau)}\right)+\frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\mu^2}\left(e^{-\mu(t-\tau)}-e^{-2\mu(t-\tau)}\right)-\frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2\mu}(t-\tau)e^{-2\mu(t-\tau)}\right) \\
&\cdot\tau e^{-\mu\tau}+\tau(\lambda_0-\lambda_\infty)e^{-\mu t}\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+(\lambda_0-\lambda_\infty)\left(t-\tau\right)e^{-\mu(t-\tau)}\right)-\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu t}\right)\right. \\
&\left.+(\lambda_0-\lambda_\infty)\,te^{-\mu t}\right)\left(\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu(t-\tau)}\right)+(\lambda_0-\lambda_\infty)\left(t-\tau\right)e^{-\mu(t-\tau)}\right) \qquad (65)
\end{aligned}$$

when $\mu=\beta-\alpha$, where $h(s)=se^{-2\mu s}$ if $2\mu=\beta-\alpha$ and $h(s)=\frac{e^{-(\beta-\alpha)s}-e^{-2\mu s}}{2\mu-\beta+\alpha}$ if $2\mu\neq\beta-\alpha$ for all $s\geq 0$.

Proof. The stated forms follow by simplification of the expressions in Corollary 5, yielding

$$\mathrm{E}\left[Q_t\right] = \frac{\lambda_\infty}{\mu}\left(1-e^{-\mu t}\right) + \frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-(\beta-\alpha)t}-e^{-\mu t}\right)$$

for the mean of the $Hawkes/M/\infty$ queue,

$$\begin{aligned}\mathrm{Cov}\left[\lambda_t, Q_t\right] = &\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)(\mu+\beta-\alpha)}\left(1-e^{-(\mu+\beta-\alpha)t}\right) + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\left(e^{-(\beta-\alpha)t}-e^{-(\mu+\beta-\alpha)t}\right)\\ &- \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu-\beta+\alpha)}\left(e^{-2(\beta-\alpha)t}-e^{-(\mu+\beta-\alpha)t}\right)\end{aligned}$$

for the covariance between the queue and the intensity, and

$$\begin{aligned}\mathrm{Var}\left(Q_t\right) = &\frac{\lambda_\infty}{\mu}\left(1-e^{-\mu t}\right) + \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2\mu(\beta-\alpha)(\mu+\beta-\alpha)}\left(1-e^{-2\mu t}\right) - \left(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{(\beta-\alpha)(\mu+\beta-\alpha)} + \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\right.\\ &\left. - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{(\beta-\alpha)(\mu-\beta+\alpha)}\right)\frac{e^{-(\mu+\beta-\alpha)t}-e^{-2\mu t}}{\mu-\beta+\alpha} + \left(\lambda_0-\lambda_\infty + \frac{\mu(\lambda_0-\lambda_\infty)}{\mu-\beta+\alpha} + \frac{2\alpha\beta(\lambda_0-\lambda_\infty)}{\mu(\beta-\alpha)}\right)\\ &\cdot\frac{e^{-(\beta-\alpha)t}-e^{-2\mu t}}{2\mu-\beta+\alpha} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)(\mu-\beta+\alpha)^2}\left(e^{-2(\beta-\alpha)t}-e^{-2\mu t}\right) - \frac{\lambda_0-\lambda_\infty}{\mu-\beta+\alpha}\left(e^{-\mu t}-e^{-2\mu t}\right)\end{aligned}$$

for the variance of the queue, all in the case where $\mu \neq \beta-\alpha$. The remaining derivation follows directly from substitution of these functions and the corresponding expressions for remaining cases and epochs into Equations 62 and 63. □

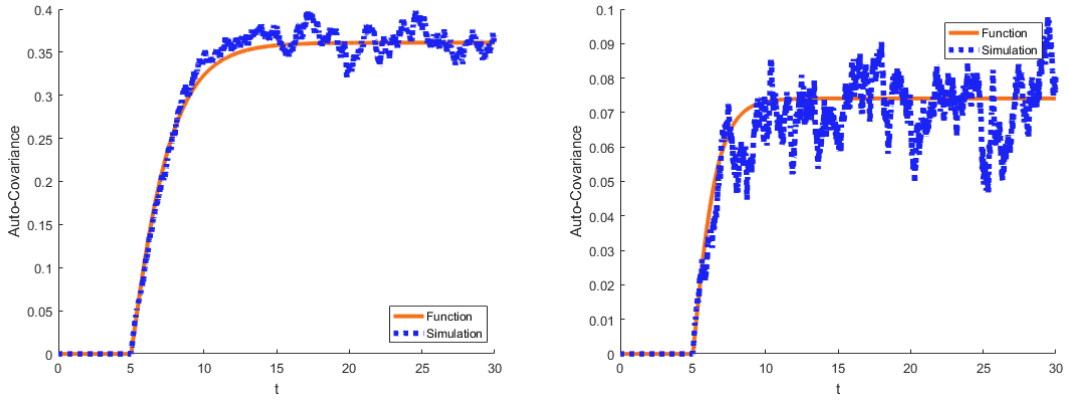

Figure 7 **Auto-covariance of the $Hawkes/M/\infty$ Queue for $\tau = 5$, where $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, $\lambda^* = \mu = 1$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = \mu = 1$ (right).**

In Figure 7 the expressions in Proposition 3 are compared to simulations, based on 100,000 replications.

3.5. Generating Functions for the $Hawkes/PH/\infty$ Queue

To complement these findings, we also derive a form for the moment generating function for a general queueing system driven by a Hawkes process.

THEOREM 4. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 with $\alpha < \beta$ and phase-type distributed service. Let $\delta \in \mathbb{R}_+^{n+1}$ and let $M(\delta,t) = M(\delta_0,\ldots,\delta_n,t) = \mathrm{E}\left[e^{\delta_0\lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}\right]$. Then, the moment generating function for the queueing system $M(\delta,t)$ is given by the solution to the following partial differential equation,*

$$\frac{\partial M(\delta,t)}{\partial t} = \delta_0\beta\lambda^* M(\delta,t) + \left(\sum_{i=1}^n \theta_i(e^{\delta_0\alpha+\delta_i}-1) - \delta_0\beta\right)\frac{\partial M(\delta,t)}{\partial \delta_0} + \sum_{i=1}^n\left(\mu_{i0}(e^{-\delta_i}-1) + \sum_{k\neq i}\mu_{ik}(e^{\delta_k-\delta_i}-1)\right)\frac{\partial M(\delta,t)}{\partial \delta_i}. \tag{66}$$

Proof. This proof makes use of techniques similar to the prior theorems, and so we omit the preceding infinitesimal generator steps. Note that $\frac{\partial M(\delta,t)}{\partial t} = \frac{\partial}{\partial t}\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right]$. From this, we start with the following.

$$\begin{aligned}\frac{\partial M(\delta,t)}{\partial t} = \mathrm{E}\Bigg[&\delta_0\beta(\lambda^*-\lambda_t)e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}} + \sum_{j=1}^n \lambda_t\theta_j\left(e^{\delta_0(\lambda_t+\alpha)+\sum_{k\neq j}\delta_k Q_{t,k}+\delta_j(Q_{t,j}+1)} - e^{\delta\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right)\\ &+\sum_{k=1}^n\sum_{j\neq k}\mu_{jk}Q_{t,j}\left(e^{\delta_0\lambda_t+\sum_{l\neq j\wedge l\neq k}\delta_l Q_{t,l}+\delta_j(Q_{t,j}-1)+\delta_k(Q_{t,k}+1)} - e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right)\\ &+\sum_{j=1}^n\mu_{j0}Q_{t,j}\left(e^{\delta_0\lambda_t+\sum_{k\neq j}\delta_k Q_{t,k}+\delta_j(Q_{t,k}-1)} - e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right)\Bigg]\end{aligned}$$

Now, we distribute terms and notice that the difference of exponentials here can be expressed as the following products.

$$\begin{aligned}\frac{\partial M(\delta,t)}{\partial t} = \mathrm{E}\Bigg[&\delta_0\beta\lambda^* e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}} - \delta_0\beta\lambda_t e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}} + \sum_{j=1}^n\lambda_t\theta_j e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\left(e^{\delta_0\alpha+\delta_j}-1\right)\\ &+\sum_{k=1}^n\sum_{j\neq k}\mu_{jk}Q_{t,j}e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\left(e^{\delta_k-\delta_j}-1\right) + \sum_{j=1}^n\mu_{j0}Q_{t,j}e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\left(e^{-\delta_j}-1\right)\Bigg]\end{aligned}$$

Here, we can now use linearity of expectation and group like terms.

$$\begin{aligned}\frac{\partial M(\delta,t)}{\partial t} = &\delta_0\beta\lambda^*\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right] + \left(\sum_{j=1}^n\theta_j(e^{\delta_0\alpha+\delta_j}-1)-\delta_0\beta\right)\mathrm{E}\left[\lambda_t e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right]\\ &+\sum_{j=1}^n\left(\mu_{j0}(e^{-\delta_j}-1)+\sum_{k\neq j}\mu_{jk}(e^{\delta_k-\delta_j}-1)\right)\mathrm{E}\left[Q_{t,j}e^{\delta_0\lambda_t+\sum_{i=1}^n\delta_i Q_{t,i}}\right]\end{aligned}$$

Finally, here we recognize the form of partial derivatives of $M(\delta,t)$ in each expectation, and so we simplify to the desired result. $\square$

We can use this to also find a partial differential equation for the natural logarithm of the moment generating function. This is called the cumulant moment generating function, as the derivative of this function yields the cumulant moments.

COROLLARY 8. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process (λ_t, N_t) with dynamics given in Equation 4 and phase-type distributed service. Let $\delta \in \mathbb{R}_+^{n+1}$ and let $G(\delta,t) = G(\delta_0,\ldots,\delta_n,t) = \log\left(\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]\right)$. Then, the cumulant moment generating function for the queueing system $G(\delta,t)$ is given by the solution to the following partial differential equation,*

$$\frac{\partial G(\delta,t)}{\partial t} = \delta_0\beta\lambda^* + \left(\sum_{i=1}^n \theta_i(e^{\delta_0\alpha+\delta_i}-1) - \delta_0\beta\right)\frac{\partial G(\delta,t)}{\partial \delta_0} + \sum_{i=1}^n \left(\mu_{i0}(e^{-\delta_i}-1) + \sum_{k\neq i}\mu_{ik}(e^{\delta_k-\delta_i}-1)\right)\frac{\partial G(\delta,t)}{\partial \delta_i}. \tag{67}$$

Proof. To begin, we see from the derivative of the logarithm and the chain rule that

$$\frac{\partial G(\delta,t)}{\partial t} = \frac{\partial}{\partial t}\log\left(\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]\right) = \frac{\frac{\partial}{\partial t}\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]}$$

and here we can recognize that these expectations are the moment generating function. Using Theorem 4, we have

$$\frac{\partial G(\delta,t)}{\partial t} = \delta_0\beta\lambda^* + \left(\sum_{i=1}^n \theta_i(e^{\delta_0\alpha+\delta_i}-1) - \delta_0\beta\right)\frac{\frac{\partial}{\partial \delta_0}\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]} + \sum_{i=1}^n \left(\mu_{i0}(e^{-\delta_i}-1) + \sum_{k\neq i}\mu_{ik}(e^{\delta_k-\delta_i}-1)\right)\frac{\frac{\partial}{\partial \delta_i}\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]}.$$

Now we recognize that $\frac{\frac{\partial}{\partial \delta_i}\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]}{\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]} = \frac{\partial G(\delta,t)}{\partial \delta_i}$, and so we have the stated result. $\square$

Comparing these two partial differential equations, we see that the expression for the cumulant moment generating function only depends on the partial derivatives, not on the function itself. In some cases the cumulant moment generating function is better since it directly will compute the variance, skewness, and higher order cumulants directly without having to know the relationships between cumulants and moments. Moreover, the cumulant moments have shift and scale invariance properties, which are often desired. The PDE in Corollary 8 produces a form that provides insight to the solution through use of the method of characteristics, which we now show in the following theorem.

THEOREM 5. *Consider a queueing system with arrivals occurring in accordance to a Hawkes process* (λ_t, N_t) *with dynamics given in Equation 4 and phase-type distributed service with transient state sub-generator matrix* $S \in \mathbb{R}^{n\times n}$. *Let* $\delta \in \mathbb{R}_+^{n+1}$ *and let* $G(\delta,t) = G(\delta_0,\ldots,\delta_n,t) = \log\left(\mathrm{E}\left[e^{\delta_0\lambda_t+\sum_{i=1}^n \delta_i Q_{t,i}}\right]\right)$. *Then, the cumulant moment generating function for the queueing system* $G(\delta,t)$ *is given by*

$$G(\delta,t) = \beta\lambda^* \int_0^t h(z)\mathrm{d}z + h(0)\lambda_0 \tag{68}$$

where $h(z)$ *is the solution to the ordinary differential equation*

$$\overset{\bullet}{h}(z) = 1 - e^{\alpha h(z)}\theta^{\mathrm{T}}\left(\mathbf{v} + e^{-S(z-t)}\left(e^{\mathrm{diag}(\delta)} - I\right)\mathbf{v}\right) + \beta h(z)$$

with initial value $h(t) = \delta_0$.

Proof. We proceed by the method of characteristics for the PDE given in Corollary 8. To do so, let z be a parametrization variable and let $\Delta_0, \Delta_1, \ldots, \Delta_n$ be characteristics variables. From recognizing the linearity of the PDE, we see that we can implement the method of characteristics by setting $\overset{\bullet}{\Delta}_i(z) := \frac{\mathrm{d}\Delta_i(z)}{\mathrm{d}z}$ equal to the function serving as coefficient of $\frac{\partial G(\delta,t)}{\partial \delta_i}$ in the PDE for each $i \in \{0,\ldots,n\}$, each with initial condition that $\Delta_i(t) = \delta_i$. This yields the following system of characteristic ODE's:

$$\begin{aligned}
\overset{\bullet}{\Delta}_0(z) &= 1 - e^{\Delta_0\alpha}\sum_{j\neq i}\theta_j e^{\Delta_j} + \Delta_0\beta, \\
\overset{\bullet}{\Delta}_i(z) &= \mu_i - \mu_{i0}e^{-\Delta_i} - \sum_{j\neq i}\mu_{ij}e^{\Delta_j-\Delta_i} \qquad\qquad \forall i \in \{1,\ldots,n\}.
\end{aligned}$$

We now let $x \in \mathbb{R}^n$ be such that $x_i = e^{\Delta_i}$. Note that this substitution can also be expressed $x = e^{\mathrm{diag}(\Delta)}\mathbf{v}$, as this will be of use in solving the system. Then, we have that $\overset{\bullet}{x}_i(z) = x_i(z)\overset{\bullet}{\Delta}_i(z)$. In this form, the last n characteristic ODE's can be expressed as

$$\overset{\bullet}{x}(z) = -Sx(z) + S\mathbf{v}$$

which means that

$$x(z) = \mathbf{v} + e^{-S(z-t)}\left(e^{\mathrm{diag}(\delta)} - I\right)\mathbf{v}$$

where we have used the initial condition $x(t) = e^{\mathrm{diag}(\Delta(t))} = e^{\mathrm{diag}(\delta)}$. We now note that to follow the method of characteristics fully and receive a closed form solution to the PDE we would want to solve the remaining characteristic ODE

$$\overset{\bullet}{\Delta}_0(z) = 1 - e^{\Delta_0\alpha}\sum_{j\neq i}\theta_j e^{\Delta_j} + \Delta_0\beta = 1 - e^{\Delta_0\alpha}\theta^{\mathrm{T}}x + \Delta_i\beta$$

which has initial condition that $\Delta_0(t) = \delta_0$. Because this form of ODE is not known to have a closed form solution in terms of standard math functions, we let $h(z)$ be defined as the solution to this initial value problem. Then, we now complete the method of characteristics by solving

$$\dot{g}(z) = \beta\lambda^* \Delta_0(z) = \beta\lambda^* h(z)$$

with the initial condition that $g(0) = G(\Delta(0), 0) = \Delta_0(0)\lambda_0 = h(0)\lambda_0$. Since this ODE is already separated, we have

$$g(z) - h(0)\lambda_0 = g(z) - g(0) = \int_0^z \dot{g}(\xi)\mathrm{d}\xi = \beta\lambda^* \int_0^z h(\xi)\mathrm{d}\xi.$$

Thus, we now have

$$G(\delta, t) = g(t) = \beta\lambda^* \int_0^t h(\xi)\mathrm{d}\xi + h(0)\lambda_0$$

and this is the stated result. □

While the ODE in this statement may not be able to be solved for a closed form expression outside of special cases, this reduction of the PDE to an ODE simplifies numerical implementations. We now note that this of course extends to the moment generating function as well by simply taking the exponential of the cumulant generating function.

3.6. Simulation Study

To conclude Section 3 we provide a collection of simulation examples that verify the accuracy of our expressions for the moments in a variety of settings. In each example we derive the simulated functions via 100,000 replications of the procedure described in Ogata (1981). We start with the mean and variance of a single phase system, as shown in the pair of plots below in Figure 8.

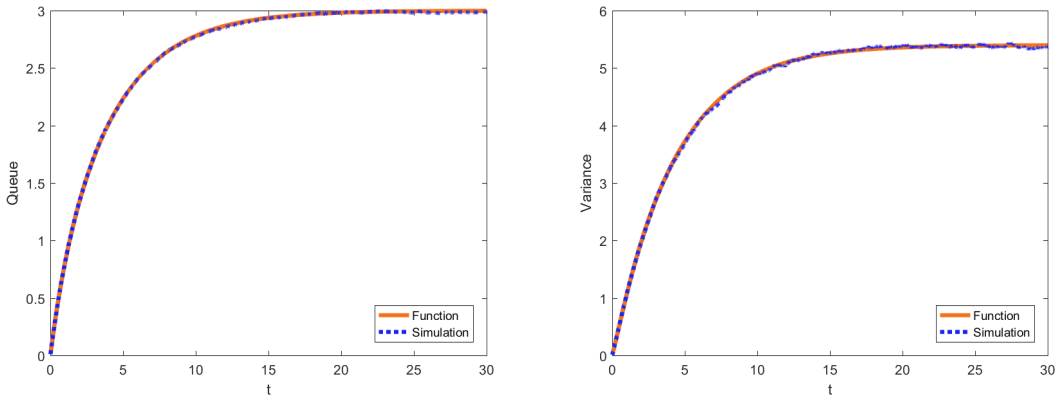

Figure 8 **Mean (left) and Variance (right) of Q_t in $Hawkes/M/\infty$, $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = \mu = 1$.**

As a second example, we also consider a three-phase Erlang distributed service. We use two different parameter settings, one in which the mean service duration is 1 and another in which the mean service length is 6. In the first case, $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, and $\lambda^* = 1$. In the latter, $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, and $\lambda^* = 1$. The mean is shown in Figure 9, the variance in Figure 10, the covariance of the queue and the intensity in Figure 11, and the covariance of the phases of the queue in Figure 12.

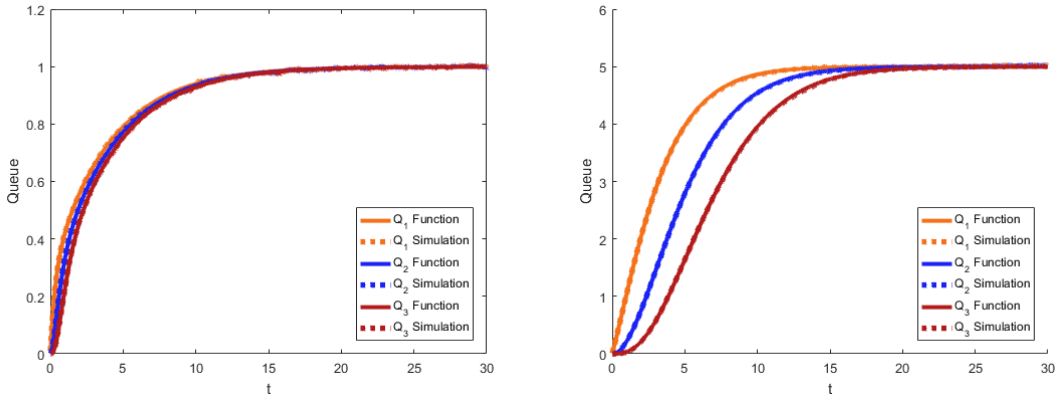

Figure 9 **Mean of the $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 6$ (right).**

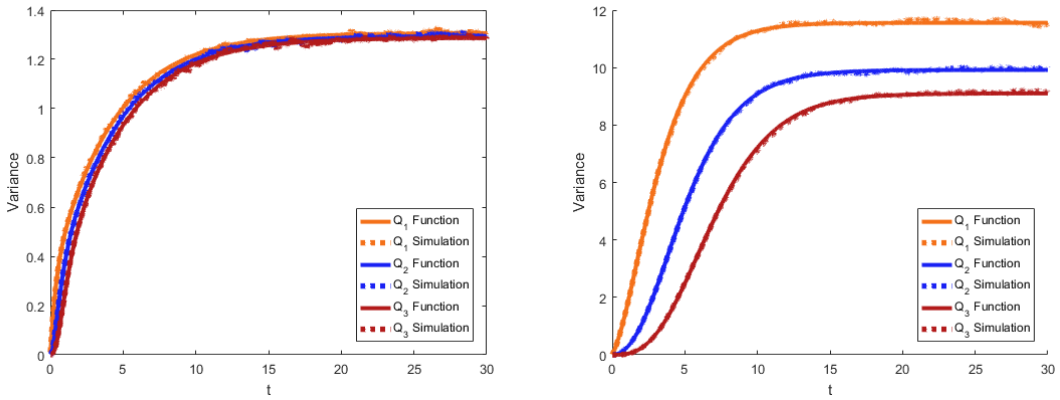

Figure 10 **Variance of the $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 6$ (right).**

In addition to the Erlang setting, we also verify the performance of the hyper-exponential service equations. We again consider a three phase distributed service and display a pair of scenarios. In both parameter groups $\theta = [.15, .4, .45]^{\mathrm{T}}$ and $\mu = [1, 4, 6]^{\mathrm{T}}$. In the first setting we consider $\alpha = \frac{1}{2}$, $\beta = 1$, and $\lambda^* = 2$, whereas in the second setting $\alpha = 1$, $\beta = 2$, and $\lambda^* = 2$. These are displayed in the same order as the Erlang examples are: mean in Figure 13, variance in Figure 14, covariance with the intensity in Figure 15, and covariance of the queues in Figure 16.

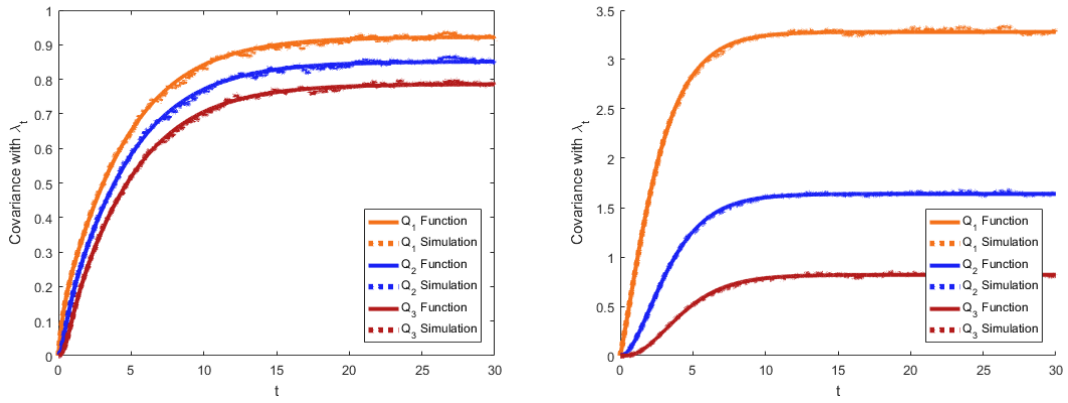

Figure 11 Covariance of $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 6$ (right).

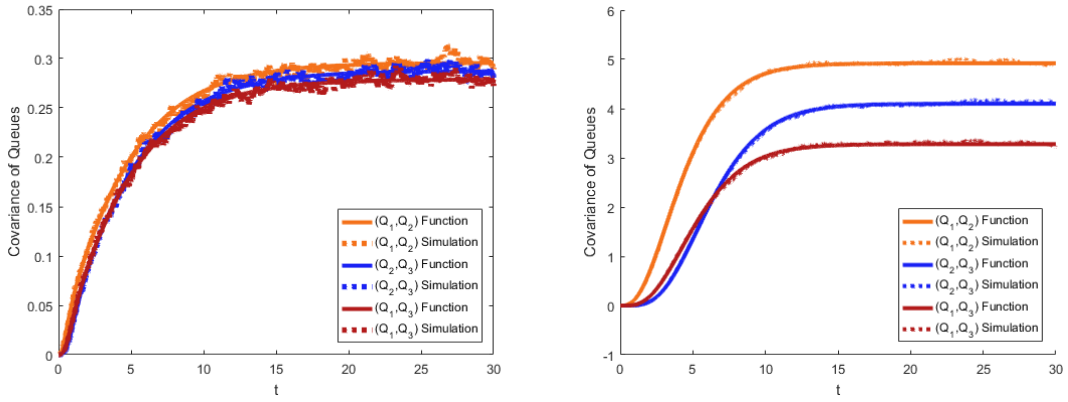

Figure 12 Covariance between Phases in the $Hawkes/E_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 1$ (left) and $\alpha = \frac{3}{4}$, $\beta = \frac{5}{4}$, $\lambda^* = 1$, $\frac{1}{\mu} = 6$ (right).

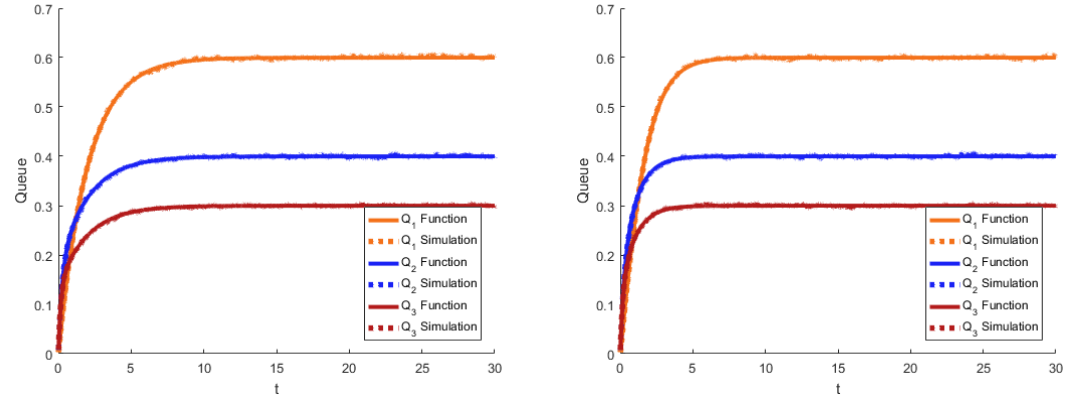

Figure 13 Mean of the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).

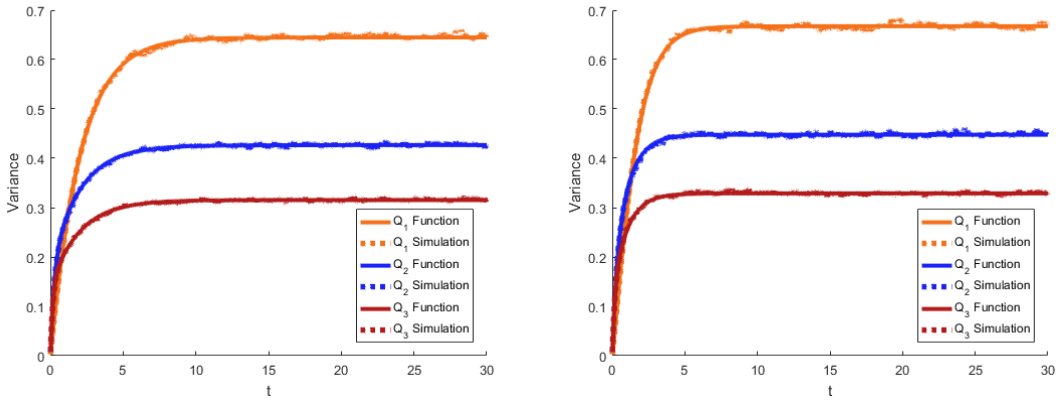

Figure 14 **Variance of the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).**

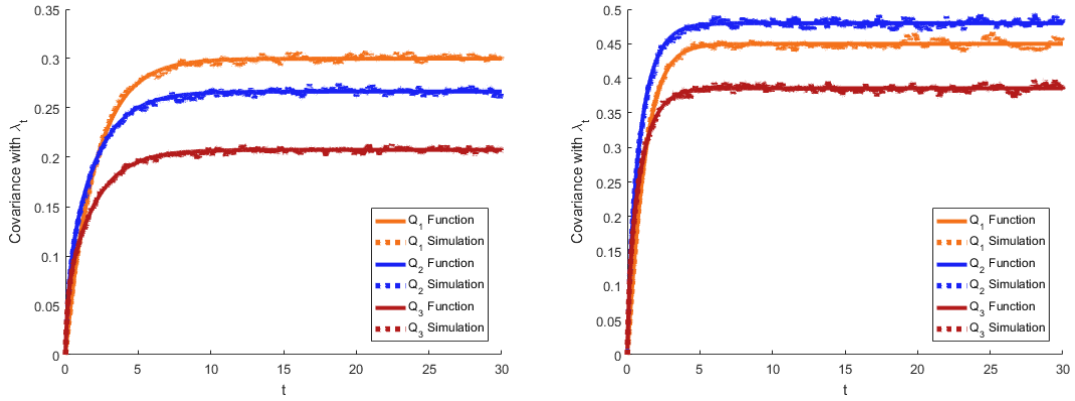

Figure 15 **Covariance of λ_t and the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).**

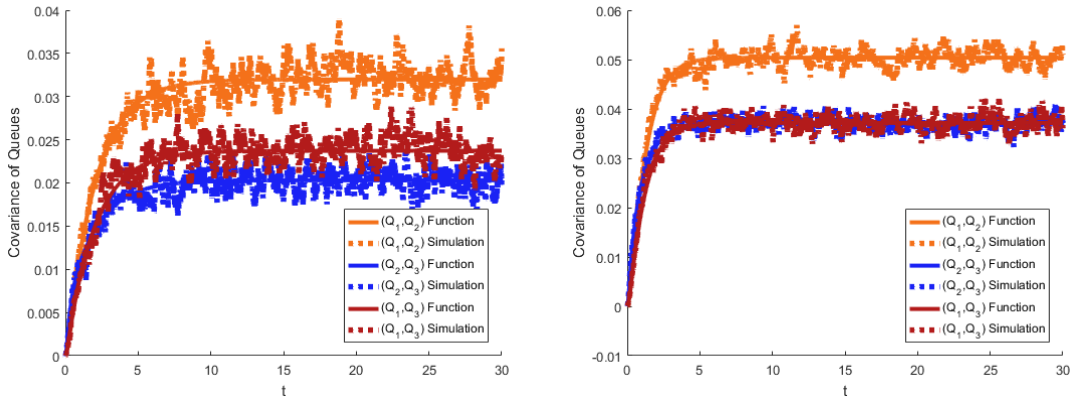

Figure 16 **Covariance between Phases in the $Hawkes/H_3/\infty$ Queue, where $\alpha = \frac{1}{2}$, $\beta = 1$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (left) and $\alpha = 1$, $\beta = 2$, $\lambda^* = 2$, $\theta = [.15, .4, .45]^{\mathrm{T}}$, $\mu = [1, 4, 6]^{\mathrm{T}}$ (right).**

In conducting these simulation experiments we have made an interesting observation. Consider the following example: let $\lambda^* = 1$, $\alpha = 1$, and $\beta = 2$. Then, let $D = 1$ be the fixed service length in a $Hawkes/D/\infty$ system and let $\mu = 1$ be the parameter of the exponential distribution in a $Hawkes/M/\infty$ system. We plot the simulated variances of these two systems in Figure 17 based on 10,000 replications, in which we find that the variance is larger in the deterministic service setting.

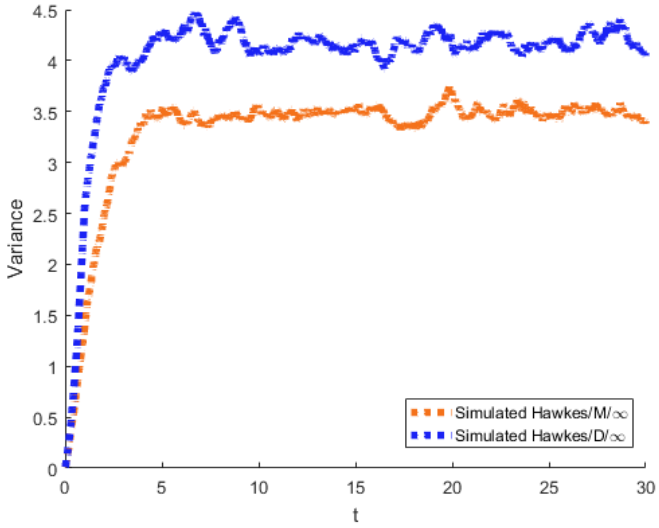

Figure 17 Comparison of Variances in $Hawkes/M/\infty$ and $Hawkes/D/\infty$ Queues when $\frac{1}{\mu} = D = 1$, with $\lambda^* = 1$, $\alpha = 1$, and $\beta = 2$.

While this relationship may seem unexpected, there is an intuitive explanation for it. Because the Hawkes process exhibits clustering behavior in the arrival times, a service system with fixed service length will also experience clusters of departures times. By comparison, a system with random service durations has the opportunity to counteract the clustering behavior and disperse the departure times. In Proposition 4 we show that the steady-state variance in the deterministic service setting is greater than that of the exponential service setting.

PROPOSITION 4. *For equal Hawkes process parameters λ^*, α, and β and equivalent service parameters $D = \frac{1}{\mu} > 0$, the steady-state variance of the $Hawkes/D/\infty$ queue is greater than the steady-state variance of the $Hawkes/M/\infty$ queue.*

Proof. Let $\beta > \alpha > 0$ and let $\lambda^* > 0$. Further, let $D = \frac{1}{\mu} > 0$. By Theorem 1, the steady-state variance of the $Hawkes/D/\infty$ queue is

$$\mathcal{V}_{\mathrm{D}} \equiv \lambda_\infty D \left(1 + \frac{2\alpha\beta - \alpha^2}{(\beta-\alpha)^2}\right) - \lambda_\infty (1 - e^{-(\beta-\alpha)D}) \frac{2\alpha\beta - \alpha^2}{(\beta-\alpha)^3}.$$

Likewise, Corollary 6 gives the steady-state variance in the exponential service case as

$$\mathcal{V}_{\mathrm{M}} \equiv \frac{\lambda_\infty}{\mu} \left(1 + \frac{2\alpha\beta - \alpha^2}{2(\beta-\alpha)(\mu+\beta-\alpha)}\right),$$

as noted in Remark 3. Then, the difference between these terms is

$$\mathcal{V}_{\mathrm{D}} - \mathcal{V}_{\mathrm{M}} = \frac{\lambda_\infty}{\mu}\left(\frac{2\alpha\beta-\alpha^2}{(\beta-\alpha)^2}\right)\left(1-\frac{\beta-\alpha}{2(\mu+\beta-\alpha)}-\frac{\mu-\mu e^{-(\beta-\alpha)\frac{1}{\mu}}}{\beta-\alpha}\right),$$

where we have substituted $\frac{1}{\mu}$ for D. Because of the assumed relationships among the parameters, $\mathcal{V}_{\mathrm{D}} - \mathcal{V}_{\mathrm{M}}$ is positive if and only if the expression inside the lattermost parenthesis is. Multiplying this expression by $\frac{2}{\mu^2}(\mu+\beta-\alpha)(\beta-\alpha) > 0$ and simplifying yields

$$\Upsilon\left(\frac{\beta-\alpha}{\mu}\right) \equiv \left(\frac{\beta-\alpha}{\mu}\right)^2 - 2\left(1-e^{-\frac{\beta-\alpha}{\mu}}\right) + 2\left(\frac{\beta-\alpha}{\mu}\right)e^{-\frac{\beta-\alpha}{\mu}}.$$

We can re-parameterize this expression as $\Upsilon(x)$ for $x \equiv \frac{\beta-\alpha}{\mu}$. By checking the first derivative of $\Upsilon(x)$, we see that it is strictly increasing for $x \geq 0$. Since $\Upsilon(0) = 0$ and $\frac{\beta-\alpha}{\mu} > 0$ for any valid α, β, and μ, we have that $\mathcal{V}_{\mathrm{D}} - \mathcal{V}_{\mathrm{M}} > 0$. □

In Figure 18 we observe that this behavior can also occur in non-Markovian service settings, shown here for lognormal distributions based on 10,000 simulation replications. In this experiment each lognormal distribution has a mean of 1 and the variances increase from 0 to 5 with a step size of 0.5. Note that all the mean queue lengths appear to be converging to 1 in steady-state. Further, we see that the means of systems with higher variance in the lognormal service distribution are converging more slowly than those of lower lognormal variance. However, the opposite relationship appears to hold in terms of the variances of the queues: the higher the variance of the lognormal, the lower the variance of the queue.

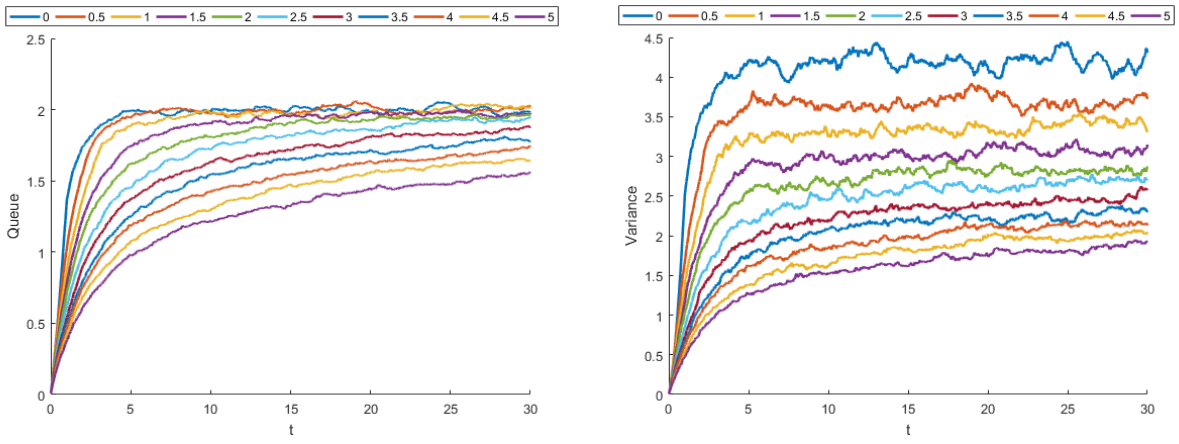

Figure 18 **Mean (left) and Variance (right) of the $Hawkes/Lognormal/\infty$ with $\lambda^* = 1$, $\alpha = 1$, and $\beta = 2$ where Mean Service Durations is 1 and Service Variance Increases from 0 to 5.**

4. Applications

To motivate this study and demonstrate its findings, we now briefly discuss two applications of this work, one concerned with viral internet traffic and one covering night clubs. Each is inspired by the self-excitement behavior of the Hawkes process, and in these settings we consider the impact and influence one arrival can have on a system and how managers of such systems might try to harness that influence for some kind of benefit.

4.1. Trending Web Traffic

In May 2017 website rankings for the United States, Youtube, Facebook, and Reddit each ranked among the top 5 most visited websites, with Twitter in the top 10 and LinkedIn and Instagram both in the top 15, per Alexa the Web Information Company (2017). For Facebook, Reddit, and Twitter in particular, users' interactions with the sites frequently involve viewing links to external media like videos, articles, and shopping sales. A user's exposure to a webpage and her likelihood to share it herself is directly influenced by whether she sees the link from other users. As users choose to visit and potentially re-share links posted by other users, the link may start trending or become "viral." This means that it is receiving high levels of traffic and arrivals to the site, and this may lead to even more arrivals while the users continue to share it on various social platforms. For a business or organization, going viral can lead to significant jumps in exposure, interest, and revenue.

As a basic example, we analyzed publicly available Twitter data McKelvey and Menczer (2013). This data set covers all tweets featuring both a URL and a hashtag from November 2012 and includes the tweet timestamp, the hashtags used, and the URL's linked, as well as an anonymous user ID. Perhaps the most notable event captured among the reactions in this data set is the 2012 U.S. Presidential election, which was held on November 6. Among the bountiful election-related tweets are 106 posts of the music video for Young Jeezy's 2008 song *My President* from the start of November 5 to midday on November 7. A plot of the timestamps of these tweets along with the total number of tweets occurring by that time is below. Note the flurry of posts once the election results were announced; 60 of the data's 106 postings of the video occur within an hour's time. A quick numerical investigation suggests that this type of extreme viral reaction may be more likely in certain parameter settings. In 100,000 simulation replications of a system with $\lambda^* = 0.5$, $\alpha = 19.5$, and $\beta = 20$, 82.4% of the trials had a majority of arrivals occur within one time quartile. By comparison, in the same number of replications for a system with $\lambda^* = 1$, $\alpha = 0.5$, and $\beta = 1$, this only occurred for 18.0% of the experiments. However, even outside of the main spike in this data, users seem to be posting the video in clustered time segments, approximately at the 6, 20, 45, 48, and 52 hour marks. These clusters suggest that these arrivals could be appropriately modeled by a Hawkes process, particularly when compared to a Poisson process.

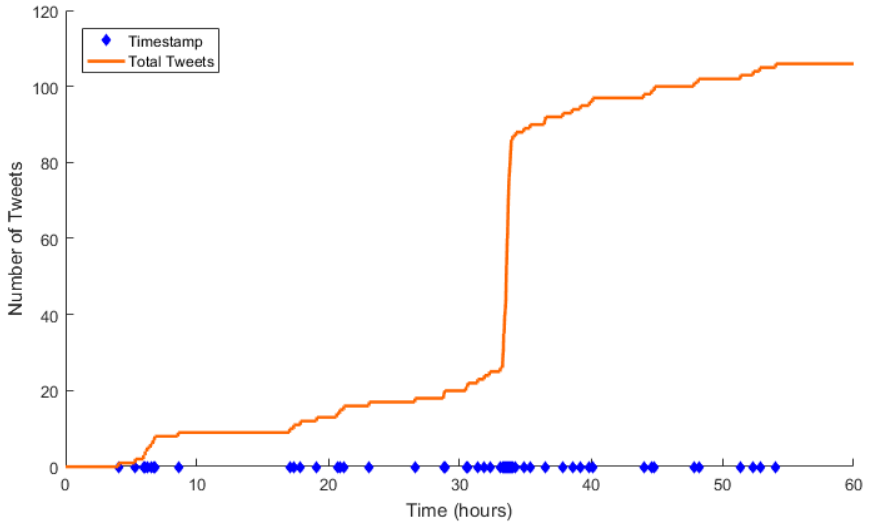

Figure 19 **Tweets of Young Jeezy - My President music video from November 5 - 7, 2012.**

Using what we have observed from this data as inspiration, we now model users arriving to a webpage as a Hawkes process. Because of the viral behavior we have seen in this type of arrivals, we will investigate the impact of a click. Consider a Hawkes Process N_t with baseline intensity λ^*, initial intensity λ_0, jump size α, and decay parameter β. Now, let $\hat{N}_t$ represent an independent Hawkes process that is identical to N_t in terms of parameters with the exception that it experienced an arrival at time 0, whereas N_t starts empty. This means that the baseline intensity, jump size, and decay parameter are the same for $\hat{N}_t$ as they were for N_t, but the initial intensity is $\lambda_0 + \alpha$ and $\hat{N}_0 = 1$. Then, by Proposition 1,

$$\begin{aligned}\mathrm{E}\left[\hat{N}_t\right] - \mathrm{E}\left[N_t\right] &= \lambda_\infty t + \frac{\lambda_0 + \alpha - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) + 1 - \lambda_\infty t - \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha}\left(1 - e^{-(\beta-\alpha)t}\right) \\ &= \frac{\beta}{\beta - \alpha} - \frac{\alpha}{\beta - \alpha} e^{-(\beta-\alpha)t} \longrightarrow \frac{\beta}{\beta - \alpha} \text{ as } t \to \infty\end{aligned}$$

which shows that the gap between the two expectations is positive and grows throughout time. However, this is simply tracking the number of visitors; it does not account for the time the users spend on the site. To capture this, we can extend this arrival model to a queueing model in which the service represents the time the user spends on the webpage. Provided the website is well hosted, this can be modeled as an infinite server queue as any user can visit the webpage that chooses to do so. If the time each user spends on the page is independently and exponentially distributed with rate μ, we see that the expected number of users on the page at time t is $\mathrm{E}\left[Q_t\right]$. Then, from time 0 to time T the expected total time spent on the page across all users $\sigma(T)$ is

$$\begin{aligned}\sigma(T) = \int_0^T \mathrm{E}\left[Q_t\right] \mathrm{d}t &= \int_0^T \left(\frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu t}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(e^{-(\beta-\alpha)t} - e^{-\mu t}\right) \right) \mathrm{d}t \\ &= \frac{\lambda_\infty}{\mu}\left(T - \frac{1 - e^{-\mu T}}{\mu}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha}\left(\frac{1 - e^{-(\beta-\alpha)T}}{\beta - \alpha} - \frac{1 - e^{-\mu T}}{\mu}\right)\end{aligned}$$

where we have applied the results of Corollary 5 for hyper-exponential service with $n = 1$ and $\mu \neq \beta - \alpha$, thus yielding exponential service. Now, suppose that a website earns m dollars per unit of time in advertising revenue for each user on the site. Then, the expected earnings by time T is $A(T) = m\sigma(T)$. We can now repeat the value of a click experiment when also considering service. Let Q_t be a queueing system with exponential service at rate μ, infinite servers, and Hawkes process arrivals with parameters λ^*, α, and β and assume the queue starts empty. Then, let $\hat{Q}_t$ be the analogous adaptation of Q_t that $\hat{N}_t$ is to N_t. Let $A(T)$ and $\hat{A}(T)$ be the corresponding expected dwell time revenues, each with earning rate m. Note that the expected time the initial customer has spent in the system by time T is $\min\{S, T\}$ where S is the duration of her service. Hence the revenue associated with her visit to the page by time T is $m\frac{1-e^{-\mu T}}{\mu}$. Then,

$$\begin{aligned}\hat{A}(T) - A(T) &= m\frac{1-e^{-\mu T}}{\mu} + m\frac{\alpha}{\mu-\beta+\alpha}\left(\frac{1-e^{-(\beta-\alpha)T}}{\beta-\alpha} - \frac{1-e^{-\mu T}}{\mu}\right) \\ &= \frac{m}{\mu}\left(1+\frac{1}{\beta-\alpha}\right) - m\frac{\alpha e^{-(\beta-\alpha)T}}{(\beta-\alpha)(\mu-\beta+\alpha)} - m\frac{(\mu-\beta)e^{-\mu T}}{\mu(\mu-\beta+\alpha)},\end{aligned}$$

which can be shown to also always grow with T via its first derivative. We can also further observe that as $\alpha \to \beta$ each of these gaps grows towards infinity, and thus so grows the impact of a click in viral settings.

Note that this model can also be used for internet-inspired applications other than users arriving to internet pages. For example, as mobile carriers continue to add cloud storage based services and allow customers to upload pictures from their smart phones as soon as they are taken, the $Hawkes/M/\infty$ queue can be used to describe the number of pictures being uploaded at once. For further reading on the Hawkes process and its use in internet traffic applications see Rizoiu et al. (2017), in which the authors develop a novel Hawkes-process-based model for the popularity of online content in great detail.

4.2. Club Queue

From our Hawkes driven infinite server queue with phase-type service distributions, we can construct what we refer to as the *Club Queue*. This stems from an application perhaps uncommon to queueing systems, a nightclub. This setting features a key characteristic: the best club has the most people waiting for it. Because of this, the Hawkes process naturally represents the excitation exhibited by club-goers joining a queue as many club-goers might call their friends to join them. With this application in mind, it is important to understand the characteristics of nightclubs. Many nightclubs have waiting spaces for potential customers outside the club. Moreover, inside the club is where much of the activity happens. Thus, using phase-type distributions we can model the inside and outside of the club as two phases of services or a two dimensional phase-type queue.

The first phase of service can be considered "admittance" to the service with the second step being the service itself. Because the clubs' bouncers have the ability to admit customers into the venue from any position in the external queue and because each customer determines how long she stays in the club, we model this scenario as an infinite server queue. This process is visualized below, where μ_O and μ_I are the rates of each step of service.

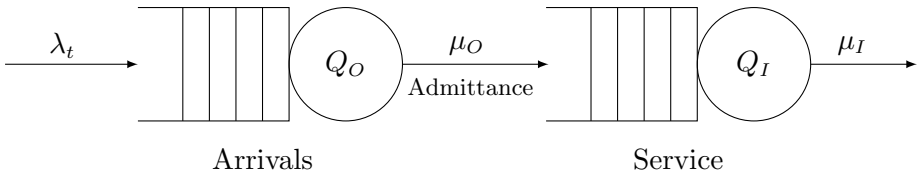

Figure 20 Club Queue Process Diagram.

We can represent the Club Queue using the two dimensional vector of queue lengths $Q(t)$ for $t \geq 0$, with coordinates $Q_I(t)$ and $Q_O(t)$ representing the service systems inside and outside the club, respectively. A fundamental managerial task is to figure out at what rate to admit club-goers into the club to maximize profitability while making the club attractive from the outside. This is non-trivial as a short line outside the club might signal to others that the club is not interesting and make them choose to not go inside the club. However, if the line is too long, there are many customers not actively generating revenue for the club and becoming frustrated with the wait outside. With this in mind, we construct the following objective function that maximizes the rate at which the bouncer of the club should let club-goers inside the club over the finite time horizon $[0, T]$, where $T > 0$.

$$\zeta(\mu_O(t)) = r_O \mu_O \mathrm{E}\left[Q_O(t)\right] + r_I \mathrm{E}\left[Q_I(t)\right] - c(\mu_O \mathrm{E}\left[Q_O(t)\right] - k)^2 - w\mu_O^2 \tag{69}$$

Here $r_O \geq 0$ and $r_I \geq 0$ are revenues generated from the cover outside and inside the club respectively. We also have that c is a penalty for having the overall admittance rate be too slow or too fast and finally, w is a penalty for admitting each individual customer too quickly. A complete formulation of this optimal control problem is presented next.

Problem 1 (Unconstrained Club Profit Model).

$$\max_{\{\mu_O \geq 0\}} \int_0^T \left[r_O \mu_O(t) \mathrm{E}\left[Q_O(t)\right] + r_I \mathrm{E}\left[Q_I(t)\right] - c(\mu_O(t) \mathrm{E}\left[Q_O(t)\right] - k)^2 - w\mu_O(t)^2\right] \mathrm{d}t$$

subject to

$$\dot{\mathrm{E}}[\lambda(t)] = \beta \cdot (\lambda^* - \mathrm{E}[\lambda(t)]) + \alpha \cdot \mathrm{E}[\lambda(t)]$$

$$\dot{\mathrm{E}}[Q_O(t)] = \mathrm{E}[\lambda(t)] - \mu_O(t) \cdot \mathrm{E}[Q_O(t)]$$

$$\dot{\mathrm{E}}[Q_I(t)] = \mu_O(t) \cdot \mathrm{E}[Q_O(t)] - \mu_I \cdot \mathrm{E}[Q_I(t)]$$

The solution to this problem gives the optimal rate to admit club-goers across time in order to maximize the difference between club revenue and the queue length and admittance rate penalties. This is characterized by the following theorem.

THEOREM 6. *The optimal solution to Problem 1 is given by $\mu_O^*(t)$, where*

$$\mu_O^*(t) = \frac{(r_O + 2ck - \gamma_1 + \gamma_2)\mathrm{E}\left[Q_O(t)\right]}{2w + 2c\mathrm{E}\left[Q_O(t)\right]^2} \tag{70}$$

for all $t \in [0,T]$.

Proof. We start by transforming the optimization model into a single Hamiltonian equation, which can be thought of as an unconstrained version of the Lagrangian. For this problem, we have the Hamiltonian $\mathcal{H}$ as

$$\begin{aligned}\mathcal{H}(t,\gamma) = \zeta(\mu_O(t)) - \gamma_1\left(\dot{\mathrm{E}}[Q_O(t)] - \mathrm{E}[\lambda(t)] + \mu_O\mathrm{E}[Q_O(t)]\right) - \gamma_2\Big(\dot{\mathrm{E}}[Q_I(t)] - \mu_O\mathrm{E}[Q_O(t)] \\ + \mu_I\mathrm{E}[Q_I(t)]\Big) - \gamma_3\left(\dot{\mathrm{E}}[\lambda(t)] - \beta\cdot(\lambda^*(t) - \mathrm{E}[\lambda(t)]) - \alpha\cdot\mathrm{E}[\lambda(t)]\right)\end{aligned}$$

where each $\gamma_i \in \mathbb{R}$ for $i \in \{1,2,3\}$. To achieve optimality in the control problem, the method ensures that $\mu_O(t)$ is such that $\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\mu_O(t)} = 0$ for all $t \in [0,T]$. We see that the derivative of the Hamiltonian with respect to $\mu_O(t)$ is

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\mu_O(t)} = r_O\mathrm{E}\left[Q_O(t)\right] - 2c\mu_O(t)\mathrm{E}\left[Q_O(t)\right]^2 + 2ck\mathrm{E}\left[Q_O(t)\right] - 2w\mu_O(t) - \gamma_1\mathrm{E}\left[Q_O(t)\right] + \gamma_2\mathrm{E}\left[Q_O(t)\right].$$

Thus, the optimal $\mu_O^*(t)$ is found by solving

$$0 = \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}\mu_O(t)} = (r_O + 2ck - \gamma_1 + \gamma_2)\mathrm{E}\left[Q_O(t)\right] - (2c\mathrm{E}\left[Q_O(t)\right]^2 + 2w)\mu_O^*(t)$$

for $\mu_O^*(t)$, which yields the expression in Equation 70. Because the objective function is concave in $\mu_O(t)$ at every t, we have that this solution corresponds to a maximum. □

Using the differential equations shown in Section 3, this optimization problem can be solved numerically by the Forward Backward sweep method as in Niyirora and Pender (2016), Qin and Pender (2017), Lenhart and Workman (2007). We now give two example outputs of this method below.

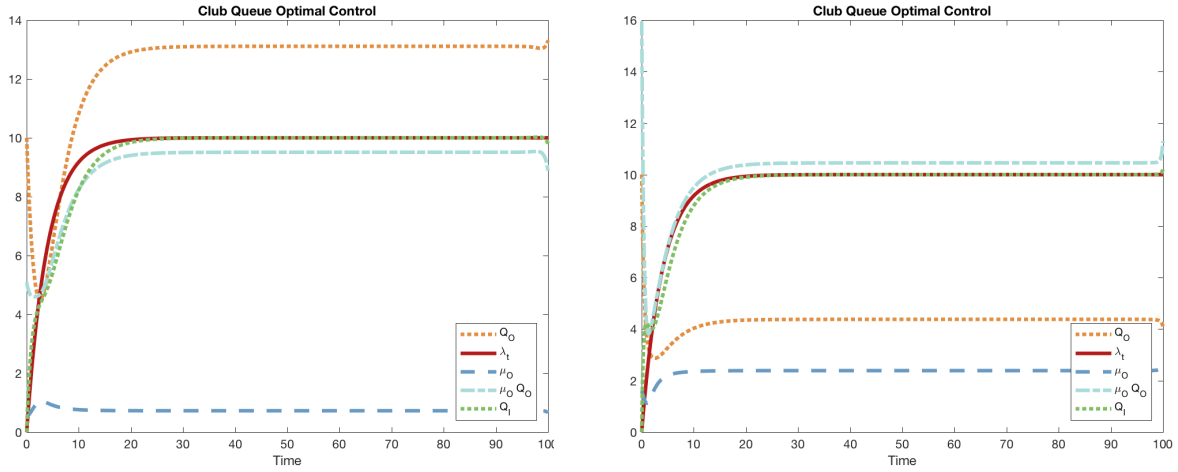

Figure 21 Example Forward Backward Sweep Implementation.

In the scenario on the left, the parameters are as follows: r_O, the external entrance revenue rate, is equal to 100 units of currency per units of time. The revenue per person inside, r_I, is equal to 100 units of currency per person. The cost of deviating from the desired admittance rate k, c, is also 100, whereas $k = 8$. Finally, the penalty for admitting individuals too quickly, $w = 150$. On the right, w is instead 100 and $k = 12$. These changes have significant impacts on the resulting solution. On the left the outside queue is allowed to grow roughly three times as large whereas on the right μ_O is approximately twice the size of that on the left.

5. Conclusion and Final Remarks

In this paper, we analyze a new infinite server stochastic queueing model that is driven by a Hawkes arrival process and phase-type distributed service. We are able to derive the exact moments and moment generating function for the Hawkes driven queue as well as the Hawkes process itself.

Although we have analyzed this queueing model in great detail, there are many extensions that are worthy of future study. One extension that we intend to explore is the impact of a non-stationary baseline intensity in the spirit of Massey and Pender (2013), Pender (2014a), Engblom and Pender (2014), Pender (2016a, 2015a,b, 2016b). In one simple example, we could set the baseline be $\lambda^*(t) = \lambda^* + \rho \cdot \sin(t)$. This analysis of a non-stationary baseline intensity is important not only because arrival rates of customers are not constant over time, but also because it is important to know how to distinguish and separate the impact of the time varying arrival rate from the impact of the stochastic dynamics of the self-excitation. The extension of one periodic function such as $\sin(t)$ seems analytically tractable, however, additional functions may require Fourier analysis.

Other extensions include the modeling of different types of queueing models other than the infinite server model. For example, it would be interesting to apply our analysis to the Erlang-A queueing model with abandonments. With regard to obtaining analytical expressions for the

Erlang-A model, this is a non-trivial problem because even the Erlang-A queueing model with a Poisson arrival process is analytically somewhat intractable. This presents new challenges for deriving analytical formulas and approximations for the moment behavior of this type of queueing model. Work by Massey and Pender (2011), Pender (2014c,b, 2015a, 2016c), Daw and Pender (2017) shows that simple closure approximations or spectral expansions can be effective at approximating the dynamics of the Erlang-A model and variants. Thus, a natural extension is to apply these techniques to the Erlang-A setting when it is driven by a Hawkes process. Not only do these approximations have the potential to describe the moment dynamics, but they can be used to stabilize performance measures like in Pender and Massey (2017). A detailed analysis of these extensions will provide a better understanding how the information that operations managers provide to their customers will affect the dynamics of these real world systems like in Pender et al. (2017a, 2018, 2017b). We plan to explore these extensions in subsequent work.

Acknowledgements

This work is supported by the National Science Foundation under grant DGE-1650441.

References

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

David R Cox. A use of complex probabilities in the theory of stochastic processes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 313–319. Cambridge University Press, 1955.

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.

José Da Fonseca and Riadh Zaatour. Clustering and mean reversion in a Hawkes microstructure model. *Journal of Futures Markets*, 35(9):813–838, 2015.

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

Andrew Daw and Jamol Pender. New perspectives on the Erlang-A queue. *arXiv preprint arXiv:1712.08445*, 2017.

Laurens G Debo, Christine Parlour, and Uday Rajan. Signaling quality via queues. *Management Science*, 58(5):876–891, 2012.

Stefan Engblom and Jamol Pender. Approximations for the moments of nonstationary and state dependent birth-death queues. *arXiv preprint arXiv:1406.6164*, 2014.

Xuefeng Gao and Lingjiong Zhu. A functional central limit theorem for stationary Hawkes processes and its application to infinite-server queues. *arXiv preprint arXiv:1607.06624*, 2016.

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

David Koops, Mayank Saxena, Onno Boxma, and Michel Mandjes. Infinite-server queues with Hawkes arrival processes. *arXiv preprint arXiv:1707.02196*, 2017.

Patrick J. Laub, Thomas Taimre, and Philip K. Pollett. Hawkes processes. *arXiv preprint*, 07 2015. URL `https://arxiv.org/abs/1507.02822`.

Suzanne Lenhart and John T Workman. *Optimal control applied to biological models*. CRC Press, 2007.

William A Massey and Jamol Pender. Poster: skewness variance approximation for dynamic rate multiserver queues with abandonment. *ACM SIGMETRICS Performance Evaluation Review*, 39(2):74–74, 2011.

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.

Karissa Rae McKelvey and Filippo Menczer. Truthy: Enabling the study of online social networks. In *Proceedings of the 2013 conference on Computer supported cooperative work companion*, pages 23–26. ACM, 2013.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106 (493):100–108, 2011.

Robin Mordfin. Why long lines can be good for shoppers, and business, 2015.

Jerome Niyirora and Jamol Pender. Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)*, 63(8):615–630, 2016.

Karl Oelschlager. A martingale approach to the law of large numbers for weakly interacting stochastic processes. *The Annals of Probability*, pages 458–479, 1984.

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

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

Jamol Pender. Gram Charlier expansion for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics*, 74(4):1238–1265, 2014a.

Jamol Pender. Laguerre polynomial expansions for time varying multiserver queues with abandonment, 2014b.

Jamol Pender. A Poisson–Charlier approximation for nonstationary queues. *Operations Research Letters*, 42(4):293–298, 2014c.

Jamol Pender. Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Informational Sciences*, 29(01):27–49, 2015a.

Jamol Pender. The truncated normal distribution: Applications to queues with impatient customers. *Operations Research Letters*, 43(1):40–45, 2015b.

Jamol Pender. An analysis of nonstationary coupled queues. *Telecommunication Systems*, 61(4):823–838, 2016a.

Jamol Pender. Risk measures and their application to staffing nonstationary service systems. *European Journal of Operational Research*, 254(1):113–126, 2016b.

Jamol Pender. Sampling the functional Kolmogorov forward equations for nonstationary queueing networks. *INFORMS Journal on Computing*, 29(1):1–17, 2016c.

Jamol Pender and William A Massey. Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probability in the Engineering and Informational Sciences*, 31(1):1–42, 2017.

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

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

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.

András Prékopa. On Poisson and composed Poisson stochastic set functions. *Stud. Math*, 16:142–155, 1957.

Ziyuan Qin and Jamol Pender. Dynamic control for nonstationary queueing networks. 2017.

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.

Alexa the Web Information Company. Top sites in United States, 05 2017. URL `http://www.alexa.com/topsites/countries/US`.

Benjamin Zhang. Here's what happens when an airline suffers a catastrophic shutdown, 08 2016. URL `http://www.businessinsider.com/what-happens-airline-mass-canceled-flight-2016-8`.

Appendix

A.1. Auto-covariance of the $Hawkes/PH/\infty$ Queue

PROPOSITION 5. *Consider the $Hawkes/PH/\infty$ queue described in Section 3 with sub-generator matrix $S \in \mathbb{R}^{n\times n}$ such that $S+(\beta-\alpha)I$ is nonsingular. Then, for $t \geq \tau \geq 0$,*

$$\mathrm{Cov}\left[Q_t, Q_{t-\tau}\right] = \lambda_\infty \left(-S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}\tau}\right)\theta\Bigg(\lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}\left(I - e^{S^{\mathrm{T}}(t-\tau)}\right)\theta - (\lambda_0-\lambda_\infty)\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}$$

$$\cdot\left(e^{-(\beta-\alpha)(t-\tau)}I - e^{S^{\mathrm{T}}(t-\tau)}\right)\theta\Bigg)^{\mathrm{T}} - \left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}\left(e^{-(\beta-\alpha)\tau}I - e^{S^{\mathrm{T}}\tau}\right)\theta\Bigg(\frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}$$

$$\cdot \left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}\left(I - e^{(S^{\mathrm{T}}-(\beta-\alpha)I)(t-\tau)}\right)\theta - \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\left(S^{\mathrm{T}}\right)^{-1}\Big(e^{-(\beta-\alpha)(t-\tau)}I$$
$$- e^{(S^{\mathrm{T}}-(\beta-\alpha)I)(t-\tau)}\Big)\theta + \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}\Big(e^{-2(\beta-\alpha)(t-\tau)}I$$
$$- e^{(S^{\mathrm{T}}-(\beta-\alpha)I)(t-\tau)}\Big)\theta + \left(\lambda_\infty + (\lambda_0-\lambda_\infty)e^{-(\beta-\alpha)(t-\tau)}\right)\Big(\lambda_\infty\big(-S^{\mathrm{T}}\big)^{-1}\big(I - e^{S^{\mathrm{T}}(t-\tau)}\big)\theta$$
$$- (\lambda_0-\lambda_\infty)\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}\big(e^{-(\beta-\alpha)(t-\tau)}I - e^{S^{\mathrm{T}}(t-\tau)}\big)\theta\Big)\Big)^{\mathrm{T}} + \lambda_\infty\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}$$
$$\cdot\Big(e^{-(\beta-\alpha)\tau}I - e^{S^{\mathrm{T}}\tau}\Big)\theta\Big(\lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}\big(I - e^{S^{\mathrm{T}}(t-\tau)}\big)\theta - (\lambda_0-\lambda_\infty)\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}$$
$$\cdot\big(e^{-(\beta-\alpha)(t-\tau)}I - e^{S^{\mathrm{T}}(t-\tau)}\big)\theta\Big)^{\mathrm{T}} + \frac{\alpha(2\beta-\alpha)\lambda_\infty}{2(\beta-\alpha)}\left((\beta-\alpha)I - S^{\mathrm{T}}\right)^{-1}e^{S^{\mathrm{T}}\tau}\Bigg(2(\beta-\alpha)e^{S^{\mathrm{T}}(t-\tau)}$$
$$\cdot M_{0,\theta,S}(t-\tau)e^{S(t-\tau)} + \theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}(t-\tau)}\theta\theta^{\mathrm{T}}e^{S(t-\tau)} + e^{S^{\mathrm{T}}(t-\tau)}\theta\theta^{\mathrm{T}}\left(e^{-(\beta-\alpha)(t-\tau)}I - e^{S(t-\tau)}\right)$$
$$\cdot((\beta-\alpha)I+S)^{-1}((\beta-\alpha)I-S) + \left((\beta-\alpha)I - S^{\mathrm{T}}\right)\left((\beta-\alpha)I+S^{\mathrm{T}}\right)^{-1}\Big(e^{-(\beta-\alpha)(t-\tau)}I$$
$$- e^{S^{\mathrm{T}}(t-\tau)}\Big)\theta\theta^{\mathrm{T}}e^{S(t-\tau)}\Bigg)\left((\beta-\alpha)I-S\right)^{-1} + \frac{\alpha\beta(\lambda_0-\lambda_\infty)}{\beta-\alpha}\left(S^{\mathrm{T}}\right)^{-1}e^{S^{\mathrm{T}}\tau}\Bigg((\beta-\alpha)e^{S^{\mathrm{T}}(t-\tau)}$$
$$\cdot M_{-(\beta-\alpha),\theta,S}(t-\tau)e^{S(t-\tau)} + e^{-(\beta-\alpha)(t-\tau)}\theta\theta^{\mathrm{T}} - e^{S^{\mathrm{T}}(t-\tau)}\theta\theta^{\mathrm{T}}e^{S(t-\tau)} - e^{S^{\mathrm{T}}(t-\tau)}\theta\theta^{\mathrm{T}}$$
$$\cdot\left(e^{-(\beta-\alpha)(t-\tau)}I - e^{S(t-\tau)}\right)((\beta-\alpha)I+S)^{-1}S - S^{\mathrm{T}}\left((\beta-\alpha)I+S^{\mathrm{T}}\right)^{-1}\Big(e^{-(\beta-\alpha)(t-\tau)}I$$
$$- e^{S^{\mathrm{T}}(t-\tau)}\Big)\theta\theta^{\mathrm{T}}e^{S(t-\tau)}\Bigg)S^{-1} - \frac{\alpha^2(2\lambda_0-\lambda_\infty)}{2(\beta-\alpha)}\left((\beta-\alpha)I+S^{\mathrm{T}}\right)^{-1}e^{S^{\mathrm{T}}\tau}\Bigg(e^{-2(\beta-\alpha)(t-\tau)}\theta\theta^{\mathrm{T}}$$
$$- e^{S^{\mathrm{T}}(t-\tau)}\theta\theta^{\mathrm{T}}e^{S(t-\tau)} - e^{S^{\mathrm{T}}(t-\tau)}\theta\theta^{\mathrm{T}}\big(e^{-(\beta-\alpha)(t-\tau)}I - e^{S(t-\tau)}\big) - \Big(e^{-(\beta-\alpha)(t-\tau)}I - e^{S^{\mathrm{T}}(t-\tau)}\Big)$$
$$\cdot\theta\theta^{\mathrm{T}}e^{S(t-\tau)}\Bigg)\left((\beta-\alpha)I+S\right)^{-1} - \lambda_\infty e^{S^{\mathrm{T}}\tau}\mathrm{diag}\Big(\left(S^{\mathrm{T}}\right)^{-1}\Big(I - e^{S^{\mathrm{T}}(t-\tau)}\Big)\theta\Big) - (\lambda_0-\lambda_\infty)$$
$$\cdot e^{S^{\mathrm{T}}\tau}\mathrm{diag}\Big(\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}\Big(e^{-(\beta-\alpha)(t-\tau)}I - e^{S^{\mathrm{T}}(t-\tau)}\Big)\theta\Big) + \Big(\lambda_\infty\left(-S^{\mathrm{T}}\right)^{-1}\big(e^{S^{\mathrm{T}}\tau}-I\big)\theta$$
$$- (\lambda_0-\lambda_\infty)\big(S^{\mathrm{T}}+(\beta-\alpha)I\big)^{-1}\big(e^{-(\beta-\alpha)It+(S^{\mathrm{T}}+(\beta-\alpha)I)\tau}I - e^{-(\beta-\alpha)t}I + e^{S^{\mathrm{T}}t} - e^{S^{\mathrm{T}}t}\big)\theta\Big)\Big(\lambda_\infty\big(-S^{\mathrm{T}}\big)^{-1}$$
$$\cdot\big(I - e^{S^{\mathrm{T}}t}\big)\theta - (\lambda_0-\lambda_\infty)\left(S^{\mathrm{T}}+(\beta-\alpha)I\right)^{-1}\big(e^{-(\beta-\alpha)t}I - e^{S^{\mathrm{T}}t}\big)\theta\Big)^{\mathrm{T}}.$$

Proof. The stated result follows directly from substitution of the expressions in Theorem 3 into Equation 61. □