Queues Driven by Hawkes Processes

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

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

July 15, 2017

Abstract
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. In this 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 mean, variance, and covariances. Furthermore, we compare our analytical results with simulation, and compare the simulated limiting distributions to queues that are driven by Poisson processes instead of Hawkes processes. As motivation for our Hawkes queueing model, we demonstrate its usefulness through three applications. These applications are calls to airline help centers, trending internet data traffic, and even arrivals to night clubs. Lastly, in the night club or Club Queue setting, we design an optimal control problem that determines the optimal rate at which a bouncer should allow club-goers to enter the club.

Keywords: Infinite-Server Queues, Hawkes Processes, Phase-type Distributions, Moments. Mathematics 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 [26] 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 [2]. 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 [17]. Even patterns of violent crime have been known to occur in clusters, as victims may decide to retaliate [13]. 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 tend to cluster together. 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.

One stochastic arrival process that captures clustering of arrivals was introduced in 1971 by Hawkes [7], and is referred to as the Hawkes process. This stochastic system counts the number of arrivals but, 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 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 work together as a pair representing the system as a whole.

Historically, the Hawkes process has been well studied in financial settings. However, it has only recently received significant amounts of attention in broader and more general contexts. For a general overview, a fantastic review of the Hawkes process was given by Laub et al. [8] in 2015. In this paper 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, they 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 viewing the act it is more enticing to join the group and see what is happening. This is the basic idea 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 [14]. These examples include cellular companies paying employees to join the lines outside stores during product launches, pastry enthusiasts waiting hours in queue to buy baked goods from the famed New York bakery Dominique Ansel, and even 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.
In this paper we apply our results to three main applications: the mass cancellations of flights, the viral nature of modern web traffic, and the appeal associated with the lengths of queues for nightclubs. For airlines, wide-reaching events like weather or technological infrastructure failure can force the cancellations of many flights at once [18]. However, even a single plane experiencing mechanical issues can cause several flight cancellations downstream. A natural service system arises out of these cancellations as help desks, ticket counters, and call centers respond to the customers who have had their travels interrupted. 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 data. Finally, 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. 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 [6]. Conceptually however, our motivation is most similar to Debo et al. [4]. While the model in [4] 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 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 functions for the mean, variance, and covariance of the Hawkes counting process, Hawkes intensity, and Hawkes process driven queue for all time, in both transient and steady state. These moments are derived for general phase-distributed service; we also provide examples for exponential, hyper-exponential, and Erlang service. These results are built from linear ordinary differential equations we find for all moments both of the arrival process and of the queue. We verify these functions via comparisons to simulations. We also derive a partial differential equation for the moment generating
function for the queueing system. Throughout this work we show the relevance of the Hawkes process by comparison to the Poisson process. In our applications, we demonstrate Markov and Chebyshev bounds on the queue, investigate the long run effect of the self-excitement structure, and design an optimal control problem and describe how to solve it numerically.

1.2 Organization of Paper

In Section 2 we analyze the Hawkes process independently. After the introductory definition, we compare and contrast the process to the Poisson process in Subsection 2.1. We then find the Hawkes process mean dynamics in Subsection 2.2. We use these findings in Section 3 to analyze the Hawkes process driven queue with phase-type service and infinite servers. We start by defining phase-type distributions in Subsection 3.1, followed by the Markovian construction of this queue and its mean dynamics in Subsection 3.2. We complete the section with a brief comparison to Poisson based queues in Subsection 3.3. We next discuss three applications in Section 4, with Subsection 4.1 devoted to airline call centers, Subsection 4.2 to viral internet traffic, and Subsection 4.3 to nightlife and clubbing. We conclude in Section 5 followed by references and appendix. 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 reading.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_t$</td>
<td>Hawkes counting process, the self-exciting point process</td>
</tr>
<tr>
<td>$\lambda_t$</td>
<td>Hawkes process intensity, represents the excitement of the process at time $t$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Hawkes process jump parameter, represents the jump in intensity upon an arrival</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Hawkes process decay parameter, governs the exponential decrease of $\lambda_t$</td>
</tr>
<tr>
<td>$\lambda^*$</td>
<td>Hawkes process baseline intensity, which is the minimum value of the intensity</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>Initial value of $\lambda_t$</td>
</tr>
<tr>
<td>$\lambda_\infty$</td>
<td>Equal to $\frac{\beta \lambda^*}{\beta - \alpha}$, represents the limit of the mean intensity as $t \to \infty$</td>
</tr>
<tr>
<td>$\mathcal{F}_t$</td>
<td>The filtration up to time $t$ generated by $(N_t)_{t \geq 0}$</td>
</tr>
<tr>
<td>$Q_t$</td>
<td>Queueing system, where $Q_{t,i}$ is the number in phase $i$ of service at time $t$</td>
</tr>
<tr>
<td>$S$</td>
<td>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 $\mu_{ij}$ and diagonal elements are $-\mu_i$</td>
</tr>
<tr>
<td>$\mu_{ij}$</td>
<td>Transition rate from phase $i$ to phase $j$ where $i \neq j$</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>Overall transition rate out of phase $i$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Queueing system initial distribution of arrivals over the $n$ phases of service</td>
</tr>
<tr>
<td>$v$</td>
<td>The $n$-dimensional vector of all ones$^1$</td>
</tr>
<tr>
<td>$v_i$</td>
<td>The $n$-dimensional vector of all zeros other than a one at the $i^{th}$ element</td>
</tr>
<tr>
<td>$V_i$</td>
<td>The $n \times n$ matrix with one at $(i,i)$ and zero otherwise</td>
</tr>
</tbody>
</table>

$^1$While this is commonly denoted as $e$, we avoid that notation as these vector are frequently used near matrix exponentials thus making $v$ more distinct from $e$. |
2 Hawkes Arrival Process

Introduced in [7], the Hawkes process 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 $\lambda_t$:

\[ \mathbb{P}(N_{t+h} - N_t = 1 | \mathcal{F}_t) = \lambda_t \cdot h + o(h) \quad (2.1) \]
\[ \mathbb{P}(N_{t+h} - N_t > 1 | \mathcal{F}_t) = o(h) \quad (2.2) \]
\[ \mathbb{P}(N_{t+h} - N_t = 0 | \mathcal{F}_t) = 1 - \lambda_t \cdot h + o(h) \quad (2.3) \]

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:

\[ d\lambda_t = \beta \cdot (\lambda^* - \lambda_t) dt + \alpha \cdot dN_t. \quad (2.4) \]

$\lambda^*$ represents the minimum value of the intensity, referred to as the baseline intensity. $\alpha > 0$ is the height of the jump in the intensity upon an arrival. $\beta > 0$ 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 by amount $\alpha$, and this increases the probability of another jump occurring. This is why the Hawkes process is called self-exciting, as its prior activity increases the likelihood of its future activity. However, as soon as an arrival occurs the intensity begins to decay exponentially with rate $\beta$ back to the baseline intensity $\lambda^*$. 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

\[ \lambda_t = \lambda^* + e^{-\beta t} \cdot (\lambda_0 - \lambda^*) + \alpha \int_0^t e^{-\beta(t-s)} dN_s. \quad (2.5) \]

This process is known to be stable for $\alpha < \beta$ [8]. Additionally, it is Markovian when conditioned on the present value of both the counting process and the intensity, as stated formally in the following proposition.

**Proposition 2.1.** The process $X_t = (\lambda_t, N_t)$ is a Markov process in the state space $\mathbb{D} = \mathbb{R}^+ \times \mathbb{N}$ when the excitation kernel function is chosen to be exponential.

**Proof.** The proof is given in the appendix.

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 analysis of the Hawkes process, we motivate it by showing both similarities and differences with the Poisson process.

2.1 Comparison to the Poisson Process

In Equation 2.5, 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 $\lambda^*$. 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. Further, a Hawkes process with $\lambda_0 = \lambda^*$ is in essence a stationary Poisson process until the first arrival occurs. However, once an arrival does occur the intensity jumps by amount $\alpha$ from the initial value and then begins to decay according to the exponential rate $\beta$. This is demonstrated in the example in Figure 2.1 below.

![Figure 2.1: Simulated $\lambda_t$, where $\alpha = \frac{3}{4}$, $\beta = 1$, and $\lambda^* = 1$](image)

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 the 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 2.2: Limit Distributions for $\lambda^* = \beta = 1$ and $\alpha = 0$ (left) and 0.6 (right).](image)

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 $\alpha$ increases the similarity between the two fades. Through these examples, we observe that the Hawkes process behaves quite differently from the Poisson process since it has heavy tails and is more variable. Thus, this provides theoretical motivation for our following investigation.

---

2 The simulations throughout this work are performed using the algorithm provided in [16].
2.2 Hawkes Process Mean Dynamics

The Markov property from Proposition 2.1 allows us to use powerful tools to analyze various properties and the distributional behavior of the Hawkes process. Among these tools is the infinitesimal generator. Consider a differentiable function $f : \mathbb{D} \to \mathbb{R}$. The infinitesimal generator of the process, denoted $L$, is the operator acting on $f$ such that

$$L f(x) = \lim_{h \to 0} \frac{E^x_t[f(X_{t+h})] - f(x)}{h}$$

with $E^x_t[\cdot] = E^x[\cdot|\mathcal{F}_t]$ [11]. In the case of Hawkes process, the infinitesimal generator can be written as the following operator

$$L f(x) = \beta (\lambda^* - \lambda_t) \cdot \frac{\partial f}{\partial \lambda}(x) + \lambda_t \cdot [f(\lambda_t + \alpha, N_t + 1) - f(x)].$$

(2.7)

It can be shown that for every function in the domain of the infinitesimal generator, the process

$$M_t = f(X_t) - f(X_0) - \int_0^t L f(X_u) du$$

(2.8)

is a martingale relative to the natural filtration. Thus, for $s \geq t$, we have by the martingale property that

$$E_t \left[ f(X_s) - \int_0^s L f(X_u) du \right] = f(X_t) - \int_0^t L f(X_u) du.$$  

(2.9)

Thus, we can also obtain Dynkin’s formula for the point process as

$$E_t [f(X_s)] = f(X_t) + E_t \left[ \int_t^s L f(X_u) du \right].$$

(2.10)

This leads us to a technical lemma that will support the rest of our analysis.

**Lemma 2.2.** Let $f$ be a function such that Equation 2.10 holds. Then,

$$\frac{d}{dt} E[f(X_t)] = E[L f(X_t)]$$

for all $t \geq 0$.

**Proof.** This follows through use of Fubini’s theorem and the fundamental theorem of calculus. Using Equation 2.10 we see

$$\frac{d}{dt} E[f(X_t)] = \frac{d}{dt} \left( f(X_0) + E \left[ \int_0^t L f(X_u) du \right] \right)$$

$$= \frac{d}{dt} E \left[ \int_0^t L f(X_u) du \right] = \frac{d}{dt} \int_0^t E[L f(X_u)] du = E[L f(X_t)].$$

and this is the stated result. \qed

Using the martingale representation, we can now find differential equations that describe the moments of the Hawkes process.
Theorem 2.3. Given a Hawkes process \( X_t = (\lambda_t, N_t) \) with dynamics given by Equation 2.4, then we have the following differential equations for the moments of \( N_t \) and \( \lambda_t \),

\[
\frac{d}{dt} E[N_t^m] = \sum_{j=0}^{m-1} \binom{m}{j} \cdot E[N_j^i \cdot \lambda_t] \\
\frac{d}{dt} E[\lambda_t^m] = m \cdot \beta \cdot \lambda^* \cdot E[\lambda_t^{m-1}] - m \cdot \beta \cdot E[\lambda_t^m] \\
\quad + \sum_{j=0}^{m-1} \binom{m}{j} \cdot \alpha^{m-j} \cdot E[\lambda_t^{j+1}] \\
\frac{d}{dt} E[\lambda_t^m \cdot N_t^l] = m \cdot \beta \cdot \lambda^* \cdot E[\lambda_t^{m-1} \cdot N_t^l] - m \cdot \beta \cdot E[\lambda_t^m \cdot N_t^l] \\
\quad + \sum_{(j,k) \in S} \binom{m}{j} \cdot \binom{l}{k} \cdot \alpha^{m-j} \cdot E[\lambda_t^{j+1} \cdot N_t^k]
\]

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

Proof. We first note that the moments of the intensity and the moments of the counting process can both be found from the product moment, and so we only derive that moment’s differential equation. By Lemma 2.2, we have that

\[
\frac{d}{dt} E[\lambda_t^m N_t^n] = E[m\beta(\lambda^* - \lambda_t)\lambda_t^{m-1}N_t^n + \lambda_t((\lambda_t + \alpha)^m(N_t + 1)^n - \lambda_t^m N_t^n)]
\]

and this distributes to the following.

\[
= m\beta \lambda^* E[\lambda_t^{m-1}N_t^n] - m\beta E[\lambda_t^m N_t^n] + E[\lambda_t(\lambda_t + \alpha)^m(N_t + 1)^n] - E[\lambda_t^{m+1} N_t^n]
\]

From the binomial theorem, this simplifies to

\[
= m\beta \lambda^* E[\lambda_t^{m-1}N_t^n] - m\beta E[\lambda_t^m N_t^n] + \sum_{(j,k) \in S} \binom{m}{j} \binom{n}{k} \alpha^{m-j} E[\lambda_t^{j+1} N_t^k]
\]

where \( S = \{(0, \ldots, m) \times (0, \ldots, n) \} \setminus \{(m, n)\} \), and this is the stated result.

Theorem 2.3 provides convenient forms for the differential equations of the means, variances, and covariance of the Hawkes process.

Corollary 2.4. Given a Hawkes process \( X_t = (\lambda_t, N_t) \) with dynamics given by Equation 2.4, then the following are differential equations for the means, variances, and covariance of \( N_t \).
and $\lambda_t$.

\[
\frac{d}{dt} E[\lambda_t] = \beta \cdot \lambda^* - (\beta - \alpha) \cdot E[\lambda_t] \tag{2.14}
\]

\[
\frac{d}{dt} E[N_t] = E[\lambda_t] \tag{2.15}
\]

\[
\frac{d}{dt} \text{Var}(\lambda_t) = \alpha^2 \cdot E[\lambda_t] - 2 \cdot (\beta - \alpha) \cdot \text{Var}(\lambda_t) \tag{2.16}
\]

\[
\frac{d}{dt} \text{Var}(N_t) = 2 \cdot \text{Cov}[\lambda_t, N_t] + E[\lambda_t] \tag{2.17}
\]

\[
\frac{d}{dt} \text{Cov}[\lambda_t, N_t] = \text{Var}(\lambda_t) + \alpha \cdot E[\lambda_t] - (\beta - \alpha) \cdot \text{Cov}[\lambda_t, N_t] \tag{2.18}
\]

Here we recognize that these equations form a closed system of linear ordinary differential equations, and thus we can find each function explicitly. The proposition below provides these closed form expressions. Before proceeding, we define $\lambda_\infty := \frac{\beta \lambda^*}{\beta - \alpha}$. We will show in Corollary 2.6 that this value is the limit of the mean Hawkes intensity, and it occurs throughout our analysis.

**Proposition 2.5.** Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 2.4, then the mean, variance, and covariance of $N_t$ and $\lambda_t$ are provided by the following equations for all $t \geq 0$.

\[
E[\lambda_t] = \lambda_\infty + (\lambda_0 - \lambda_\infty) e^{-(\beta - \alpha)t} \tag{2.19}
\]

\[
E[N_t] = \lambda_\infty t + \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha} \left(1 - e^{-(\beta - \alpha)t}\right) \tag{2.20}
\]

\[
\text{Var}(\lambda_t) = \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)} + \frac{\alpha^2(\lambda_0 - \lambda_\infty)}{\beta - \alpha} e^{-(\beta - \alpha)t} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} e^{-2(\beta - \alpha)t} \tag{2.21}
\]

\[
\text{Var}(N_t) = \left(\lambda_\infty + \frac{2\alpha \lambda_\infty}{\beta - \alpha} + \frac{\alpha^2 \lambda_\infty}{(\beta - \alpha)^2} \right) t + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} \left(1 - e^{-2(\beta - \alpha)t}\right) - \frac{2\alpha \beta (\lambda_0 - \lambda_\infty)}{(\beta - \alpha)^2} \tag{2.22}
\]

\[
\cdot te^{-(\beta - \alpha)t} + \left(\frac{\beta + \alpha}{(\beta - \alpha)^2} (\lambda_0 - \lambda_\infty) - \frac{2\alpha \beta}{(\beta - \alpha)^3} \lambda_\infty \right) (1 - e^{-(\beta - \alpha)t})
\]

\[
\text{Cov}[\lambda_t, N_t] = \left(\frac{\alpha \lambda_\infty}{\beta - \alpha} + \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)^2} \right) \left(1 - e^{-(\beta - \alpha)t}\right) + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)^2} \left(e^{-2(\beta - \alpha)t} - e^{-(\beta - \alpha)t}\right)
\]

\[
+ \frac{\alpha \beta (\lambda_0 - \lambda_\infty)}{(\beta - \alpha)} te^{-(\beta - \alpha)t} \tag{2.23}
\]

**Proof.** First consider the mean intensity. From Corollary 2.4, we know that

\[
\frac{d}{dt} E[\lambda_t] = \beta \lambda^* - (\beta - \alpha) E[\lambda_t]
\]

and so we can see that this is equivalently expressed as follows.

\[
\frac{d}{dt} \left(e^{(\beta - \alpha)t} E[\lambda_t]\right) = e^{(\beta - \alpha)t} \beta \lambda^*
\]
Integrating each side from time 0 to time \( t \) gives us that

\[
e^{(\beta-\alpha)t}E[\lambda_t] - \lambda_0 = e^{(\beta-\alpha)t} \frac{\beta \lambda^*}{\beta - \alpha} - \frac{\beta \lambda^*}{\beta - \alpha}
\]

where here we have used that \( \lambda_0 \) is the initial value of the intensity. After substituting \( \lambda_\infty \) for each \( \frac{\beta \lambda^*}{\beta - \alpha} \) as defined, this now rearranges to the stated result.

\[
E[\lambda_t] = \frac{\beta \lambda^*}{\beta - \alpha} + \left( \lambda_0 - \frac{\beta \lambda^*}{\beta - \alpha} \right) e^{-(\beta-\alpha)t} = \lambda_\infty + (\lambda_0 - \lambda_\infty) e^{-(\beta-\alpha)t}
\]

We can now simply integrate this from time 0 to time \( t \) to find the counting process mean.

\[
E[N_t] = \int_0^t E[\lambda_s] \, ds = \lambda_\infty t + \frac{\lambda_0 - \lambda_\infty}{\beta - \alpha} \left( 1 - e^{-(\beta-\alpha)t} \right)
\]

Note that here we have used that the counting process starts at 0, but other cases can be captured by simply shifting this equation by the desired amount thanks to linearity of expectation. We now move on to solving for the variance of the intensity. Using the same techniques, we see that

\[
\frac{d}{dt} (e^{2(\beta-\alpha)t} \text{Var}(\lambda_t)) = \alpha^2 e^{2(\beta-\alpha)t} E[\lambda_t] = \alpha^2 \lambda_\infty e^{2(\beta-\alpha)t} + \alpha^2 (\lambda_0 - \lambda_\infty) e^{(\beta-\alpha)t}
\]

and we can now again solve for the stated result by integrating from time 0 to \( t \) while using that the initial variance is 0.

\[
\text{Var}(\lambda_t) = \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)} \left( 1 - e^{-2(\beta-\alpha)t} \right) + \frac{\alpha^2 (\lambda_0 - \lambda_\infty)}{\beta - \alpha} \left( e^{-(\beta-\alpha)t} - e^{-2(\beta-\alpha)t} \right)
\]

\[
= \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)} + \frac{\alpha^2 (\lambda_0 - \lambda_\infty)}{\beta - \alpha} e^{-(\beta-\alpha)t} - \frac{\alpha^2 (2 \lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} e^{-2(\beta-\alpha)t}
\]

Because the variance of the counting process depends on the covariance between the counting process and the intensity, we now derive the equation for the covariance.

\[
\frac{d}{dt} \left( e^{(\beta-\alpha)t} \text{Cov}(\lambda_t, N_t) \right) = e^{(\beta-\alpha)t} \text{Var}(\lambda_t) + \alpha e^{(\beta-\alpha)t} E[\lambda_t]
\]

\[
= \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)} e^{(\beta-\alpha)t} + \frac{\alpha^2 (\lambda_0 - \lambda_\infty)}{\beta - \alpha} - \frac{\alpha^2 (2 \lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} e^{-(\beta-\alpha)t}
\]

\[
+ \alpha \lambda_\infty e^{(\beta-\alpha)t} + \alpha (\lambda_0 - \lambda_\infty)
\]

\[
= \left( \frac{\alpha \lambda_\infty}{\beta - \alpha} + \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)} \right) e^{(\beta-\alpha)t} - \frac{\alpha^2 (2 \lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} e^{-(\beta-\alpha)t}
\]

\[
+ \frac{\alpha \beta(\lambda_0 - \lambda_\infty)}{\beta - \alpha} t e^{-(\beta-\alpha)t}
\]

After performing the same definite integration and simplifying, we get the stated result.

\[
\text{Cov}(\lambda_t, N_t) = \left( \frac{\alpha \lambda_\infty}{\beta - \alpha} + \frac{\alpha^2 \lambda_\infty}{2(\beta - \alpha)} \right) \left( 1 - e^{-(\beta-\alpha)t} \right) + \frac{\alpha^2 (2 \lambda_0 - \lambda_\infty)}{2(\beta - \alpha)^2} \left( e^{-(2(\beta-\alpha)t} - e^{-(\beta-\alpha)t} \right)
\]

\[
+ \frac{\alpha \beta(\lambda_0 - \lambda_\infty)}{\beta - \alpha} t e^{-(\beta-\alpha)t}
\]
Finally, we solve for the variance of the counting process. Because we know that

$$\frac{d}{dt} \text{Var}(N_t) = 2 \text{Cov} [\lambda_t, N_t] + E[\lambda_t]$$

$$= \left( \frac{2\alpha\lambda_\infty}{\beta - \alpha} + \frac{\alpha^2\lambda_\infty}{(\beta - \alpha)^2} \right) (1 - e^{-(\beta - \alpha)t}) + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{(\beta - \alpha)^2} (e^{-2(\beta - \alpha)t} - e^{-(\beta - \alpha)t})$$

$$+ \frac{2\alpha\beta(\lambda_0 - \lambda_\infty)}{\beta - \alpha} t e^{-(\beta - \alpha)t} + \lambda_\infty + (\lambda_0 - \lambda_\infty) e^{-(\beta - \alpha)t}$$

we can solve for the variance by taking the same definite integral over the above equation.

$$\text{Var}(N_t) = \left( \frac{\lambda_\infty}{\beta - \alpha} + \frac{2\alpha\lambda_\infty}{(\beta - \alpha)^2} + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} \right) t + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)^3} (1 - e^{-2(\beta - \alpha)t}) - \frac{2\alpha\beta(\lambda_0 - \lambda_\infty)}{(\beta - \alpha)^2} t e^{-(\beta - \alpha)t}$$

$$+ \left( \frac{\beta + \alpha}{(\beta - \alpha)^2} (\lambda_0 - \lambda_\infty) - \frac{2\alpha\beta}{(\beta - \alpha)^3} \lambda_\infty \right) (1 - e^{-(\beta - \alpha)t})$$

This yields the stated result, thus completing the proof.

By further observation of either Corollary 2.4 or Proposition 2.5, we see that we can find the steady-state behavior of various Hawkes process statistics. We show this in the following corollary.

**Corollary 2.6.** Given a Hawkes process $X_t = (\lambda_t, N_t)$ with dynamics given by Equation 2.4, then the are following equilibrium solutions for differential equations presented in Corollary 2.4. Additionally, the differential equations imply that the functions converge to the equilibria, and hence they are limits of the functions.

$$\lim_{t \to \infty} E[\lambda_t] = \frac{\beta \lambda^*}{\beta - \alpha} = \lambda_\infty$$  \hspace{1cm} (2.24)

$$\lim_{t \to \infty} \text{Var}(\lambda_t) = \frac{\alpha^2\lambda_\infty}{2(\beta - \alpha)}$$  \hspace{1cm} (2.25)

$$\lim_{t \to \infty} \text{Cov} [\lambda_t, N_t] = \frac{\alpha \lambda_\infty}{\beta - \alpha} + \frac{\alpha^2\lambda_\infty}{2(\beta - \alpha)^2}$$  \hspace{1cm} (2.26)

**Proof.** The proof follows directly from taking the limit as $t \to \infty$ in the equations in Proposition 2.5 or by finding the equilibrium solutions of the differential equations in Corollary 2.4.

To investigate the performance of these findings numerically, we provide several plots that compare the simulated mean and variance of the Hawkes process with the values given by the differential equations.
The left-hand plot in Figure 2.3 compares the simulated mean of the intensity and the mean intensity found from the differential equations. Likewise, the right-hand pair of curves compares the simulated variance of the intensity with its differential equation counterpart. Below, we show the same comparisons for the mean and variance of the counting process. Each plot is simulated from 100,000 sample paths.

Now that we have analyzed the Hawkes process, we will use it as an arrival process to a queueing system to model clustering and self-exciting behavior in service systems.

3 Hawkes/PH/∞ Queue

In this section we will explore queueing systems in which arrivals occur according to a Hawkes process. To begin, we define key components of the general queueing system discussed throughout this section, including 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 reasons. The first is that this is more general than just exponential service, and it can be shown that phase-type distributions can be used to approximate any other non-negative continuous distribution
[3]. 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 analysis.

3.1 Phase-Type Distributions

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,

\[
\Gamma = \begin{pmatrix} 0 & 0 \\ s & S \end{pmatrix}.
\]

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

\[
S = \begin{pmatrix} -\mu_1 & \cdots & \mu_{1,n} \\ \vdots & \ddots & \vdots \\ \mu_{n,1} & \cdots & -\mu_n \end{pmatrix}, \quad s = (\mu_{1,0}, \ldots, \mu_{n,0})^T, \tag{3.1}
\]

where the \( \mu_{ij} \)'s agree with the definition of the infinitesimal generator matrix \( \Gamma \). For notational consistency, we use a term phase to indicate the state of CTMC of the phase-type distributions throughout this paper.

3.2 Mean Dynamics of the Hawkes/PH/\( \infty \) Queue

With the phase-type distributions described in Subsection 3.1, 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 \( \mu_{ij} \). 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 \( \theta_i \), with \( \sum_{i=1}^n \theta_i = 1 \). An entity departs the system at rate \( \mu_{i0} \), where \( i \) is the entity’s phase of service before leaving. For brevity of notation, define \( \mu_i := \mu_{i0} + \mu_{i1} + \cdots + \mu_{i,i-1} + \mu_{i,i+1} + \mu_{i,n} \). Let \( Q_t \in \mathbb{N}^n \) represent the number of entities in the queueing system, with \( Q_{t,i} \) representing the number in phase \( i \) of service i.e.

\[
Q_t = \sum_{i=1}^n Q_{t,i} v_i \tag{3.2}
\]

where \( v_i \) is the unit column vector in the \( i \)th coordinate. We let \( (\lambda_t, N_t) \) represent a Hawkes process as described in Equation 2.4. We will now use this to 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 a 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.

\[
\mathcal{L} f(x) = \beta (\lambda^* - \lambda_t) \cdot \frac{\partial f(x)}{\partial \lambda_t} + \sum_{i=1}^{n} \lambda_t \theta_i \left( f(\lambda_t + \alpha, N_t + 1, Q_{t,i} + 1) - f(x) \right) + \sum_{i=1}^{n} \mu_{ij} Q_{t,i} \left( f(\lambda_t, N_t, Q_{t,i} - 1, Q_{t,j} + 1) - f(x) \right) + \sum_{i=1}^{n} \mu_{i0} Q_{t,i} \left( f(\lambda_t, N_t, Q_{t,i} - 1) - f(x) \right)
\]

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/∞ queueing system. We have that

\[
E_t [f(X_s)] = f(X_t) + E_t \left[ \int_t^s \mathcal{L} f(X_u) du \right],
\]

where now \( X_t = (\lambda_t, N_t, Q_t) \). This gives rise to a lemma equivalent to Lemma 2.2 for the queueing system state space.

**Lemma 3.1.** Let \( f \) be a function such that Equation 3.4 holds. Then,

\[
\frac{d}{dt} E[f(X_t)] = E[\mathcal{L} f(X_t)]
\]

for all \( t \geq 0 \).

**Proof.** Like in Lemma 2.2, this is achieved through use of Fubini’s theorem and the fundamental theorem of calculus. Using Equation 3.4 we have that

\[
\frac{d}{dt} E[f(X_t)] = \frac{d}{dt} \left( f(X_0) + E \left[ \int_0^t \mathcal{L} f(X_u) du \right] \right)
\]

\[
= \frac{d}{dt} E \left[ \int_0^t \mathcal{L} f(X_u) du \right] = \frac{d}{dt} \int_0^t E[\mathcal{L} f(X_u)] du = E[\mathcal{L} f(X_t)]
\]

and this completes this proof.

We now use this to find the moments for the full queueing system.

**Theorem 3.2.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process \((\lambda_t, N_t)\) with dynamics given in Equation 2.4 and phase-type distributed service. Then we have differential equations for the moments of \( Q_{t,i} \) given by

\[
\frac{d}{dt} E \left[ Q_{t,i}^m \right] = \theta_i \sum_{g=0}^{m-1} \binom{m}{g} E \left[ \lambda_t Q_{t,i}^g \right] + \sum_{g=0}^{m-1} \sum_{j=1}^{n} \binom{m}{g} \mu_{ji} E \left[ Q_{t,j} Q_{t,i}^g \right] + \sum_{g=1}^{m} \binom{m}{g-1} \mu_i (-1)^{m-g+1} E \left[ Q_{t,i}^g \right],
\]

14
for the products of \( Q_{t,i} \) and \( Q_{t,j} \) where \( i \neq j \) given by

\[
\frac{d}{dt} \mathbb{E} [Q_{t,i}^m Q_{t,j}^k] = \theta_i \sum_{g=0}^{m-1} \binom{m}{g} \mathbb{E} \left[ \lambda_t Q_{t,j}^l Q_{t,i}^{g} \right] + \theta_j \sum_{h=0}^{l-1} \binom{l}{h} \mathbb{E} \left[ \lambda_t Q_{t,i}^m Q_{t,j}^h \right] + \sum_{k=1}^{n} \sum_{g=0}^{m-1} \binom{m}{g} \mu_{k} \mathbb{E} \left[ Q_{t,k} Q_{t,i}^g Q_{t,j}^l \right] + \sum_{k=1}^{n} \sum_{h=0}^{l-1} \binom{l}{h} \mu_{kJ} \mathbb{E} \left[ Q_{t,k} Q_{t,i}^m Q_{t,j}^k \right] + \mu_{i} \sum_{g=0}^{m-1} \binom{m}{g} (-1)^{m-g} \mathbb{E} \left[ Q_{t,j}^l Q_{t,i}^{g-1} \right] + \mu_{ij} \sum_{g=0}^{m-1} \binom{m}{g} \frac{(-1)^{m-g} E [Q_{t,i}^{g+1} Q_{t,j}^h]}{g} + \mu_{j} \sum_{h=0}^{l-1} \binom{l}{h} (-1)^{l-h} \mathbb{E} \left[ Q_{t,i}^m Q_{t,j}^{k-1} \right] + \mu_{ji} \sum_{h=0}^{l-1} \binom{l}{h} \frac{(-1)^{l-h} E [Q_{t,i}^{h+1} Q_{t,j}^g]}{l} (3.6)
\]

and for the products of \( \lambda_t \) and \( Q_{t,i} \) given by

\[
\frac{d}{dt} \mathbb{E} [\lambda_t^m Q_{t,i}^l] = \beta \lambda^* m \mathbb{E} [\lambda_t^{m-1} Q_{t,i}^l] - \beta m \mathbb{E} [\lambda_t^m Q_{t,i}^l] + \theta_i \sum_{g=0}^{m-1} \binom{m}{g} \frac{(-1)^{m-g} E [\lambda_t^{l+1} Q_{t,i}^g]}{g} + \mu_{i} \sum_{h=0}^{l-1} \binom{l}{h} \frac{(-1)^{l-h} E [\lambda_t^m Q_{t,i}^{h+1}]}{l} (3.7)
\]

where \( t \geq 0 \).

**Proof.** We can first observe that each of these moments can be generalized to \( \mathbb{E} [\lambda_t^m Q_{t,i}^l Q_{t,j}^k] \). From Lemma 3.1 we see that

\[
\frac{d}{dt} \mathbb{E} [\lambda_t^m Q_{t,i}^l Q_{t,j}^k] = \mathbb{E} \left[ \beta (\lambda^* - \lambda_t) m \lambda_t^{m-1} Q_{t,i}^l Q_{t,j}^k + \lambda_t \theta_i ((\lambda_t + \alpha)^m (Q_{t,i} + 1)^l Q_{t,j}^k - \lambda_t^m Q_{t,i}^l Q_{t,j}^k) + \lambda_t \theta_j ((\lambda_t + \alpha)^m Q_{t,j} + 1)^k - \lambda_t^m Q_{t,i}^l Q_{t,j}^k) + \sum_{x=1}^{n} \lambda_t \theta_x Q_{t,i}^l Q_{t,j}^k ((\lambda_t + \alpha)^m - \lambda_t^m)
\]

\[
+ \sum_{x=1}^{n} \mu_{x} Q_{t,x}^l \lambda_t^m Q_{t,j}^k ((Q_{t,i} + 1)^l - Q_{t,i}^l) + \sum_{x=1}^{n} \mu_{x} Q_{t,x}^l \lambda_t^m Q_{t,i}^l ((Q_{t,j} + 1)^k - Q_{t,j}^k)
\]

\[
+ \sum_{x=0}^{n} \mu_{x} Q_{t,x}^l \lambda_t^m Q_{t,j}^k ((Q_{t,i} + 1)^l - Q_{t,i}^l) + \sum_{x=0}^{n} \mu_{x} Q_{t,x}^l \lambda_t^m Q_{t,i}^l ((Q_{t,j} + 1)^k - Q_{t,j}^k)
\]

\[
+ \mu_{ij} Q_{t,i}^l \lambda_t^m ((Q_{t,j} + 1)^l - Q_{t,i}^l) + \mu_{ji} Q_{t,j}^k \lambda_t^m ((Q_{t,i} + 1)^k - Q_{t,j}^k)
\]

\[
\]
where here 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.

\[ \beta \lambda^* m E \left[ \lambda_t^{n-1} Q_{t,i}^l Q_{t,j}^k \right] - \beta m E \left[ \lambda_t^m Q_{t,i}^l Q_{t,j}^k \right] + \sum_{n=1}^{m-1} \sum_{y=0}^{m-1} \binom{m}{y} \theta_x \alpha^{m-y} E \left[ \lambda_t^{y+1} Q_{t,i}^l Q_{t,j}^k \right] \]

\[ + \theta_i \left( \sum_{x=0}^{m} \sum_{y=0}^{l} \binom{m}{x} \binom{l}{y} \alpha^{m-x} E \left[ \lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^k \right] - E \left[ \lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k \right] \right) + \theta_j \left( \sum_{x=0}^{m} \sum_{y=0}^{k} \binom{m}{x} \binom{k}{y} \alpha^{m-x} E \left[ \lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^k \right] - E \left[ \lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k \right] \right) \]

\[ \cdot \alpha^{m-x} E \left[ \lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^k \right] - E \left[ \lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k \right] \]

\[ + \sum_{x=1}^{l} \sum_{y=0}^{l} \binom{k}{y} \mu_{xj} E \left[ \lambda_t^{m} Q_{t,x}^l Q_{t,j}^{l+1} \right] + \sum_{x=0}^{l-1} \sum_{y=0}^{l-1} \binom{l}{y} \mu_{ix} E \left[ \lambda_t^{m} Q_{t,i}^l Q_{t,j}^{l+1} \right] \]

\[ + \sum_{x=0}^{l-1} \sum_{y=0}^{l-1} \sum_{z=0}^{l-1} \binom{l}{x} \binom{l}{y} \binom{l}{z} \mu_{xj} \mu_{ix} E \left[ \lambda_t^{m} Q_{t,x}^l Q_{t,j}^{l+1} \right] - E \left[ \lambda_t^{m} Q_{t,i}^l Q_{t,j}^{l+1} \right] \]

Now here 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

\[ \beta \lambda^* m E \left[ \lambda_t^{n-1} Q_{t,i}^l Q_{t,j}^k \right] - \beta m E \left[ \lambda_t^m Q_{t,i}^l Q_{t,j}^k \right] + \sum_{y=0}^{m-1} \binom{m}{y} \alpha^{m-y} E \left[ \lambda_t^{y+1} Q_{t,i}^l Q_{t,j}^k \right] \]

\[ + \theta_i \left( \sum_{x=0}^{m} \sum_{y=0}^{l} \binom{m}{x} \binom{l}{y} \alpha^{m-x} E \left[ \lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^k \right] - E \left[ \lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k \right] \right) + \theta_j \left( \sum_{x=0}^{m} \sum_{y=0}^{k} \binom{m}{x} \binom{k}{y} \alpha^{m-x} E \left[ \lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^k \right] - E \left[ \lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k \right] \right) \]

\[ \cdot \alpha^{m-x} E \left[ \lambda_t^{x+1} Q_{t,i}^l Q_{t,j}^k \right] - E \left[ \lambda_t^{m+1} Q_{t,i}^l Q_{t,j}^k \right] \]

\[ + \sum_{x=1}^{l} \sum_{y=0}^{l} \sum_{z=0}^{l} \binom{k}{y} \mu_{xj} E \left[ \lambda_t^{m} Q_{t,x}^l Q_{t,j}^{l+1} \right] + \sum_{x=0}^{l-1} \sum_{y=0}^{l-1} \sum_{z=0}^{l-1} \binom{l}{x} \binom{l}{y} \binom{l}{z} \mu_{xj} \mu_{ix} E \left[ \lambda_t^{m} Q_{t,x}^l Q_{t,j}^{l+1} \right] - E \left[ \lambda_t^{m} Q_{t,i}^l Q_{t,j}^{l+1} \right] \]

which is equivalent to each stated result when \( m = k = 0 \), \( k = 0 \), and \( m = 0 \), respectively.

This now gives rise to Corollary 3.3, which finds the differential equations for the mean,
variance, and covariances of queues driven by Hawkes processes.
Corollary 3.3. Consider a queueing system with arrivals occurring in accordance to a Hawkes process \((\lambda_t, N_t)\) with dynamics given in Equation 2.4 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.

\[
\frac{d}{dt} E[Q_{t,i}] = \theta_i E[\lambda_t] + \sum_{j=1, j \neq i}^n \mu_{ji} E[Q_{t,j}] - \mu_i E[Q_{t,i}] \quad (3.8)
\]

\[
\frac{d}{dt} \text{Var}(Q_{t,i}) = \theta_i E[\lambda_t] + 2\theta_i \text{Cov}[\lambda_t, Q_{t,i}] + 2 \sum_{j=1, j \neq i}^n \mu_{ji} \text{Cov}[Q_{t,j}, Q_{t,i}] + \mu_i E[Q_{t,i}] \quad (3.9)
\]

\[
\frac{d}{dt} \text{Cov}[\lambda_t, Q_{t,i}] = (\alpha - \beta - \mu_i) \text{Cov}[\lambda_t, Q_{t,i}] + \alpha \theta_i E[\lambda_t] + \sum_{j=1, j \neq i}^n \mu_{ji} \text{Cov}[\lambda_t, Q_{t,j}] \quad (3.10)
\]

\[
\frac{d}{dt} \text{Cov}[Q_{t,i}, Q_{t,j}] = \theta_i \text{Cov}[\lambda_t, Q_{t,j}] + \theta_j \text{Cov}[\lambda_t, Q_{t,i}] - (\mu_i + \mu_j) \text{Cov}[Q_{t,i}, Q_{t,j}] \quad (3.11)
\]

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. First, we find a form for the indefinite integral of the exponential of a non-singular matrix.

Lemma 3.4. Let \(L \in \mathbb{R}^{n \times n}\) be invertible. Then, if the integral of \(e^{Lt}\) exists it can be expressed

\[
\int e^{Lt} \, dt = L^{-1}e^{Lt} + c
\]

where \(c\) is some constant of integration.

Proof. By the definition of the matrix exponential, we have that

\[
\int e^{Lt} \, dt = \sum_{k=0}^{\infty} \frac{t^k}{k!} e^{Lt} \, dt = \sum_{k=0}^{\infty} \frac{t^k}{k!} \int e^{Lt} \, dt
\]

where we have exchanged the order of integration and summation using Fubini’s theorem. Then,

\[
\sum_{k=0}^{\infty} L^k \int \frac{t^k}{k!} \, dt = \sum_{k=0}^{\infty} L^k \frac{t^{k+1}}{(k+1)!} + c_{k+1}
\]
where \( c_{k+1} \) is the corresponding constant of integration for each term. Now, we have assumed that the integral of \( e^{Lt} \) exists and so we introduce a general constant of integration, \( c \) such that
\[
\sum_{k=0}^{\infty} L^k \frac{t^{k+1}}{(k+1)!} + c = L^{-1} \sum_{k=0}^{\infty} L^{k+1} \frac{t^{k+1}}{(k+1)!} + c = L^{-1} e^{Lt} + c
\]
where here \( c \) has both absorbed the \( c_k \) series and the shift from adjusting the summand.

We now use this lemma in proving a second technical lemma, which 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 3.5.** 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,
\[
\int_0^t e^{Ls} \nu s^\eta e^{\gamma s} \, ds = \sum_{k=0}^{\eta} \frac{\eta!}{(\eta-k)!} (-1)^k (L + \gamma I)^{-(k+1)} (e^{Lt} \nu t^{\eta-k} e^{\gamma t}) - \eta! (-1)^\eta (L + \gamma I)^{-(\eta+1)} \nu
\]
for \( t > 0 \).

**Proof.** We show this via induction on \( \eta \), starting with a base case of \( \eta = 0 \). In that case we have that
\[
\int_0^t e^{Ls} \nu e^{\gamma s} \, ds = L^{-1} (e^{Lt} e^{\gamma t} - I) \nu - \int_0^t L^{-1} e^{Ls} \nu e^{\gamma s} \, ds
\]
via integration by parts and Lemma 3.4. This rearranges to
\[
(I + \gamma L^{-1}) \int_0^t e^{Ls} \nu e^{\gamma s} \, ds = L^{-1} (e^{Lt} e^{\gamma t} - I) \nu
\]
and because the non-singularity of \((I + \gamma L^{-1})\) is implied by the assumed non-singularity of \( L \) and \( L + \gamma I \), we have that
\[
\int_0^t e^{Ls} \nu e^{\gamma s} \, ds = (I + \gamma L^{-1})^{-1} (e^{Lt} e^{\gamma t} - I) \nu = (L + \gamma I)^{-1} (e^{Lt} e^{\gamma t} - I) \nu
\]
and this satisfies the inductive hypothesis. We now suppose that the hypothesis holds for \( \eta \) and will investigate for \( \eta + 1 \). Again through integration by parts and Lemma 3.4, we have
\[
\int_0^t e^{Ls} \nu s^{\eta+1} e^{\gamma s} \, ds = L^{-1} e^{Lt} \nu t^{\eta+1} e^{\gamma t} - \int_0^t L^{-1} e^{Ls} \nu (\gamma s^{\eta+1} e^{\gamma s} + (\eta+1)s^\eta e^{\gamma s}) \, ds
\]
This rearranges to
\[
(I + \gamma L^{-1}) \int_0^t e^{Ls} \nu s^{\eta+1} e^{\gamma s} \, ds = L^{-1} e^{Lt} \nu t^{\eta+1} e^{\gamma t} - (\eta + 1) L^{-1} \int_0^t e^{Ls} \nu s^\eta e^{\gamma s} \, ds
\]
which can be rewritten again through the invertibility of \((I + \gamma L^{-1})\) to the following.

\[
\int_0^t e^{Lt}e^{\gamma s}ds = (L + \gamma I)^{-1}\left(e^{Lt}e^{\gamma}t - (\eta + 1)\int_0^t e^{Lt}e^{\gamma}ds\right)
\]

We can now substitute the inductive hypothesis for \(\eta\) into the right hand side integral.

\[
\int_0^t e^{Lt}e^{\gamma}ds = (L + \gamma I)^{-1}\left(e^{Lt}e^{\gamma}t - (\eta + 1)\left(\sum_{k=0}^{\eta} \frac{\eta!}{(\eta - k)!}(-1)^k \right.ight.

\[
\cdot (L + \gamma I)^{-(k+1)}(e^{Lt}e^{\gamma}t) - \eta!(-1)^{\eta}(L + \gamma I)^{-(\eta+1)}\nu\right)
\]

After distributing the \((L + \gamma I)^{-1}, \eta + 1,\) and \(-1\) and simplifying, we have the following.

\[
\int_0^t e^{Lt}e^{\gamma}ds = (L + \gamma I)^{-1}e^{Lt}e^{\gamma}t + \sum_{k=0}^{\eta} \frac{(\eta + 1)!}{(\eta - k)!}(-1)^{k+1}(L + \gamma I)^{-(k+2)}(e^{Lt}e^{\gamma}t)

\[
- (\eta + 1)!(-1)^{\eta+1}(L + \gamma I)^{-(\eta+2)}\nu
\]

We now shift the summation index by 1 so that \(k\) over 0 to \(\eta\) becomes \(k\) from 1 to \(\eta + 1\). This replaces each \(k\) with a \(k - 1\) and allows us to absorb \((L + \gamma I)^{-1}e^{Lt}e^{\gamma}t\) into the summation so that it again starts at 0.

\[
\int_0^t e^{Lt}e^{\gamma}ds = \sum_{k=0}^{\eta+1} \frac{(\eta + 1)!}{(\eta + 1 - k)!}(-1)^k(L + \gamma I)^{-(k+1)}(e^{Lt}e^{\gamma}t)

\[
- (\eta + 1)!(-1)^{\eta+1}(L + \gamma I)^{-(\eta+2)}\nu
\]

Finally, we observe that this is now the hypothesis for \(\eta + 1\) and thus by induction we conclude the proof.

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

**Lemma 3.6.** Let \(A \in \mathbb{R}^{n \times n}\) be invertible and let \(b \in \mathbb{R}\) be such that \(A + bI\) is also invertible. Then,

\[
e^{-A}(cA + bI)^{-1} = (cA + bI)^{-1}e^{-A}.
\]

**Proof.** By the definition of the matrix exponential and the fact that for invertible \(X\) and \(Y\) \((XY)^{-1} = Y^{-1}X^{-1}\), we have that

\[
e^{-A}(cA + bI)^{-1} = \left(\sum_{k=0}^{\infty} \frac{A^k}{k!}\right)^{-1}(A + bI)^{-1} = \left(cA \sum_{k=0}^{\infty} \frac{A^k}{k!} + bI \sum_{k=0}^{\infty} \frac{A^k}{k!}\right)^{-1}

\[
= \left(\sum_{k=0}^{\infty} \frac{A^k}{k!}(cA) + \sum_{k=0}^{\infty} \frac{A^k}{k!}(bI)\right)^{-1} = (cA + bI)^{-1}\left(\sum_{k=0}^{\infty} \frac{A^k}{k!}\right)^{-1}

\[
= (cA + bI)^{-1}e^{-A}
\]

where here we have used the commutativity of \(A\) and itself and of \(A\) and the identity.
We now use these lemmas to prove a general solution to differential equations of a certain form.

**Lemma 3.7.** Let \( g(t) \in \mathbb{R}^n \) be a function described by the dynamics

\[
\dot{g}(t) = -Lg(t) + \sum_{i \in S} \nu_i t^{n_i} e^{\gamma_i t}
\]

with an initial condition of \( g(0) = g_0 \), where \( L \in \mathbb{R}^{n \times n} \) is invertible and \( S \) 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 \( i \in S \). Then, if \( L + \gamma_i I \) is invertible for all \( i \in S \) the explicit function for \( g(t) \) is given by

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

for all \( t \geq 0 \).

**Proof.** From the assumed dynamics, we see that

\[
\dot{g}(t) + Lg(t) = \sum_{i \in S} \nu_i t^{n_i} e^{\gamma_i t}
\]

and this can be evaluated by multiplying each side of the equation by \( e^{Lt} \) and recognizing an induced product rule [23].

\[
\frac{d}{dt} (e^{Lt} g(t)) = e^{Lt} \dot{g}(t) + e^{Lt} Lg(t) = \sum_{i \in S} e^{Lt} \nu_i t^{n_i} e^{\gamma_i t}
\]

After taking the definite integral from 0 to \( t \) of each side, we have

\[
e^{Lt} g(t) - g_0 = \int_0^t \sum_{i \in S} e^{Lt} \nu_i s^{n_i} e^{\gamma_i s} ds
\]

which gives us the following equation of \( g(t) \), which we rearrange via the fact that the summation is over a finite set.

\[
g(t) = e^{-Lt} \int_0^t \sum_{i \in S} e^{Ls} \nu_i s^{n_i} e^{\gamma_i s} ds + e^{-Lt} g_0 = e^{-Lt} \sum_{i \in S} \int_0^t e^{Ls} \nu_i s^{n_i} e^{\gamma_i s} ds + e^{-Lt} g_0
\]

Now, via Lemma 3.5 we can replace the integrals as follows.

\[
e^{-Lt} \sum_{i \in S} \int_0^t e^{Ls} \nu_i s^{n_i} e^{\gamma_i s} ds + e^{-Lt} g_0 = e^{-Lt} \sum_{i \in S} \left( \sum_{k=0}^{\eta_i} \frac{\eta_i!}{(\eta_i - k)!} (L + \gamma_i I)^{-(k+1)} \left( \nu_i t^{n_i-k} e^{\gamma_i t} \right) - \eta_i! (L + \gamma_i I)^{-(\eta_i+1)} \nu_i \right) + e^{-Lt} g_0
\]

(3.12)
We can distribute both the summation over \( S \) and \( e^{-Lt} \). Ignoring \( e^{-Lt} g_0 \) for the moment to conserve space, this is as follows.

\[
e^{-Lt} \sum_{i \in S} \left( \sum_{k=0}^{\eta_i} \frac{\eta_i!}{(\eta_i - k)!} (-1)^k (L + \gamma_i I)^{-(k+1)} (e^{Lt} \nu_t^{\eta_i - k} e^{\gamma_t}) - \eta_i! (-1)^{\eta_i} (L + \gamma_i I)^{-(\eta_i+1)} \nu_i \right)
\]

\[
= \sum_{i \in S} \sum_{k=0}^{\eta_i} \frac{\eta_i! (-1)^k}{(\eta_i - k)!} (L + \gamma_i I)^{-(k+1)} (\nu_t^{\eta_i - k} e^{\gamma_t}) - \eta_i! (-1)^{\eta_i} (L + \gamma_i I)^{-(\eta_i+1)} e^{-Lt} \nu_i
\]

where we have made repeated use of Lemma 3.6 to commute \( e^{-Lt} \) with \( (L + \gamma_i I)^{-(k+1)} \) and \( (L + \gamma_i I)^{-(\eta_i+1)} \). Substituting this back into 3.12, we see that we have the desired result. \( \square \)

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

\[
M_{\gamma, \nu, L}(t) = \int_0^t e^{(\gamma I - L^T)s} \nu \nu^T e^{-Lt} \, ds
\]

(3.13)

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

\[
(M_{\gamma, \nu, L}(t))_{i,j} = \left\{ \begin{array}{ll}
\sum_{k=1}^n \sum_{l=1}^n \nu_k \nu_l \sum_{r=0}^\infty \sum_{w=0}^\infty \frac{(L')_{k,i}(L')_{l,j}}{r^w(r!w!)^2} \left( e^{\gamma t} \sum_{z=0}^\infty \frac{(-\gamma)^z}{z!} - 1 \right) & \text{if } \gamma \neq 0, \\
\sum_{k=1}^n \sum_{l=1}^n \nu_k \nu_l \sum_{r=0}^\infty \sum_{w=0}^\infty \frac{(L')_{k,i}(L')_{l,j}}{r^w(r!w!)^2} \left( e^{\gamma t} \sum_{z=0}^\infty \frac{(-\gamma)^z}{z!} - 1 \right) & \text{if } \gamma = 0.
\end{array} \right.
\]

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 \( \nu \nu^T \) need not commute. With defining \( M_{\gamma, \nu, L}(t) \) we circumvent this issue by integrating on the element-level, but if \( L \) and \( \nu \nu^T \) 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 3.8.** Let \( \eta, \gamma, c \in \mathbb{R} \), \( \nu \in \mathbb{R}^n \), \( L \in \mathbb{R}^{n \times n} \) be such that \( L, \eta \gamma I \pm L \), and \( (\eta + 1) \gamma I - L \) are each invertible. Then,

\[
\int_0^t \left( ((\eta + 1) \gamma I - L^T)^{-1} \left( e^{(\eta \gamma I - L^T)s} - e^{-\gamma I s} \right) \nu \nu^T e^{-Lt} + e^{-Lt} \nu \nu^T c \left( e^{(\eta \gamma I - L)s} - e^{-\gamma I s} \right) \right) \, ds
\]

\[
= e \left( ((\eta + 1) \gamma I - L^T)^{-1} \left( (\eta + 2) \gamma M_{\eta, \nu, L}(t) + e^{(\gamma I - L^T)t} \nu \nu^T e^{-Lt} - \nu \nu^T + \nu \nu^T \left( e^{-(\gamma I + L)t} - I \right) (\gamma I + L)^{-1} \right) (\gamma I + L)^{-1} \right)
\]

where

\[
(\eta + 1) \gamma I - L^T
\]

for all \( t \geq 0 \).
Proof. To begin, we bring the inverted terms outside the integral as follows.

\[
\int_0^t \left( (\eta + 1) \gamma I - L^T \right)^{-1} \left( e^{(\eta \gamma I - L^T)s} - e^{-\gamma Is} \right) \nu_T \nu^T e^{-Ls} \\
+ e^{-Ls} \nu_T \nu^T c \left( e^{(\eta \gamma I - L)s} - e^{-\gamma Is} \right) ((\eta + 1) \gamma I - L)^{-1} \right) \right) ds \\
= ((\eta + 1) \gamma I - L^T)^{-1} \int_0^t \left( (\eta + 2) \gamma e^{(\eta \gamma I - L^T)s} \nu_T \nu^T e^{-Ls} + (\eta \gamma I - L^T) e^{(\eta \gamma I - L^T)s} \nu_T \nu^T e^{-Ls} \\
- e^{(\eta \gamma I - L^T)s} \nu_T \nu^T e^{-Ls} - \nu_T \nu^T e^{-(\gamma I + L)s} ((\eta + 1) \gamma I - L) - ((\eta + 1) \gamma I - L^T)^{-1} \right) \\
. e^{-\gamma I + L^T s} \nu_T c \right) ds ((\eta + 1) \gamma I - L) \\
\]

where here we have used that for a scalar \( x e^x = e^x \), which implies that this exponential commutes with any matrix. We can now distribute the integral and evaluate each of the five terms inside. For the first, we can recognize that this is \((\eta + 2) \gamma c M_{\eta \gamma , \nu , L}(t)\) after integration. Now for the second and third terms we recognize that it is of the form \( X e^{Xt} e^{-Yt} + X e^{Xt} e^{Yt} \), which by the product rule is simply \( \frac{d}{dt} e^{Xt} c e^{Yt} \). Finally, we see that the fourth and fifth terms each only have one exponential function, and so we integrate them as given in Lemma 3.4. All together, we get the stated result.

\[
((\eta + 1) \gamma I - L^T)^{-1} \int_0^t \left( (\eta + 2) \gamma e^{(\eta \gamma I - L^T)s} \nu_T \nu^T e^{-Ls} + (\eta \gamma I - L^T) e^{(\eta \gamma I - L^T)s} \nu_T \nu^T e^{-Ls} \\
- e^{(\eta \gamma I - L^T)s} \nu_T \nu^T e^{-Ls} - \nu_T \nu^T e^{-(\gamma I + L)s} ((\eta + 1) \gamma I - L) - ((\eta + 1) \gamma I - L^T)^{-1} \right) \\
. e^{-\gamma I + L^T s} \nu_T c \right) ds ((\eta + 1) \gamma I - L) \\
= c((\eta + 1) \gamma I - L^T)^{-1} \left( (\eta + 2) \gamma M_{\eta \gamma , \nu , L}(t) + e^{(\eta \gamma I - L^T)s} \nu_T e^{-Ls} - \nu_T + \nu_T e^{-(\gamma I + L)t} - I) (\gamma I + L)^{-1} \\
. (\eta + 1) \gamma I - L + ((\eta + 1) \gamma I - L^T) (\gamma I + L^T)^{-1} \left( e^{-(\gamma I + L)t} - I \right) \nu_T \right) ((\eta + 1) \gamma I - L)^{-1}
\]
We can now use these lemmas to find explicit linear algebraic solutions to the closed system of differential equations in Corollary 3.3. In this case we use the covariance matrix of $Q_t$ and itself to capture both the covariance between phases (in off-diagonal elements) and the variance of each phase alone (in the elements on the diagonal). Now define $\text{diag}(x) \in \mathbb{R}^{n \times n}$ for $x \in \mathbb{R}^n$ as $\text{diag}(x) = \sum_{i=1}^n v_i x v_i^T$, where $v_i \in \mathbb{R}^n$ is the unit column vector in the direction of the $i$th coordinate and $V_i = v_i v_i^T$, meaning that the $i$th diagonal element is 1 and the rest are 0.

**Theorem 3.9.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process $(\lambda_t, N_t)$ with dynamics given in Equation 2.4 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. If $S$ and $S + (\beta - \alpha)I$ are invertible, then

$$E[Q_t] = \lambda_\infty (I - e^{St}) \theta - (\lambda_0 - \lambda_\infty) (S^T + (\beta - \alpha)I)^{-1} (e^{-(\beta - \alpha)t} - e^{St}) \theta \quad (3.14)$$

where $\lambda_\infty = \frac{\beta \lambda^*}{\alpha}$. Further, if $S - (\beta - \alpha)I$ is invertible as well, then

$$\text{Cov}[\lambda_t, Q_t] = \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)} ((\beta - \alpha)I - S^T)^{-1} \left( (I - e^{S^T((\beta - \alpha)I)T}) \theta - \frac{\alpha \beta (\lambda_0 - \lambda_\infty)}{\beta - \alpha} \right)$$

$$+ \left( e^{-(\beta - \alpha)t} I - e^{S^T((\beta - \alpha)I)T} \right) \theta + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} (S^T + (\beta - \alpha)I)^{-1}$$

$$\cdot \left( e^{2(\beta - \alpha)t} I - e^{S^T((\beta - \alpha)I)T} \right) \theta. \quad (3.15)$$

Finally, if $S + 2(\beta - \alpha)I$ is also invertible, then

$$\text{Cov}[Q_t, Q_t] = \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)} ((\beta - \alpha)I - S^T)^{-1} \left( 2(\beta - \alpha) e^{S^T M_{0, \theta, S}(t) e^S} + \theta \theta^T e^S + e^{S^T \theta \theta^T} e^S \right)$$

$$+ e^{S^T \theta \theta^T} e^S ((\beta - \alpha)I + S)^{-1} ((\beta - \alpha)I - S) + ((\beta - \alpha)I - S) \left( 2(\beta - \alpha) e^{S^T M_{-((\beta - \alpha), \theta, S}(t) e^S} ight)$$

$$\cdot \left( e^{-(\beta - \alpha)t} I - e^{S^T(T)} \theta \theta^T e^S \right) ((\beta - \alpha)I + S)^{-1} + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} (S^T)^{-1} \left( (\beta - \alpha) e^{S^T M_{-((\beta - \alpha), \theta, S}(t) e^S} ight)$$

$$+ e^{S^T \theta \theta^T} e^S - e^{S^T \theta \theta^T} e^S \left( e^{-(\beta - \alpha)t} I - e^{S^T(T)} \theta \theta^T e^S \right) ((\beta - \alpha)I + S)^{-1} S - S^T ((\beta - \alpha)I + S)^{-1}$$

$$\cdot \left( e^{-(\beta - \alpha)t} I - e^{S^T(T)} \theta \theta^T e^S \right) S^{-1} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} ((\beta - \alpha)I + S)^{-1} \left( e^{2(\beta - \alpha)t} \theta \theta^T e^S + e^{S^T \theta \theta^T} e^S ight)$$

$$- e^{S^T \theta \theta^T} e^S - \left( e^{-(\beta - \alpha)t} I - e^{S^T(T)} \theta \theta^T e^S \right) ((\beta - \alpha)I + S)^{-1} - \lambda_\infty \text{diag}(S^T)^{-1}$$

$$\cdot \left( I - e^{S^T(T)} \theta \right) - (\lambda_0 - \lambda_\infty) \text{diag}(I + S)^{-1} \left( e^{-(\beta - \alpha)t} I - e^{S^T(T)} \theta \right) \quad (3.16)$$

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 3.3 that

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

23
and so we apply Lemma 3.7. 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

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

which now simplifies to the stated result. We find the stated result for \( \text{Cov}[\lambda_t, Q_t] \) through repeating the same technique to the corresponding differential equation systems, where in this case we note that

\[
\frac{d}{dt} \text{Cov}[\lambda_t, Q_t] = (S^T - (\beta - \alpha) I) \text{Cov}[\lambda_t, Q_t] + \alpha \theta E[\lambda_t] + \theta \text{Var}(\lambda_t).
\]

We now solve for the covariance matrix. Note that from Corollary 3.3, the variance of each phase and the covariance between phases can form one linear algebraic form as the covariance matrix, as shown below.

\[
\frac{d}{dt} \text{Cov}[Q_t, Q_t] = S^T \text{Cov}[Q_t, Q_t] + \text{Cov}[Q_t, Q_t] S + \theta \text{Cov}[\lambda_t, Q_t] + \text{Cov}[\lambda_t, Q_t] \theta^T + \text{diag}(\theta E[\lambda_t] + S^T E[Q_t]) - S^T \text{diag}(E[Q_t]) - \text{diag}(E[Q_t]) S
\]

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

\[
\frac{d}{dt} \left( e^{-S^T t} \text{Cov}[Q_t, Q_t] e^{-St} \right) = e^{-S^T t} \theta \text{Cov}[\lambda_t, Q_t] e^{-St} + e^{-S^T t} \text{Cov}[\lambda_t, Q_t] \theta^T e^{-St} + e^{-S^T t} \text{diag}(\theta E[\lambda_t] + S^T E[Q_t]) e^{-St} - e^{-S^T t} \text{diag}(E[Q_t]) S e^{-St}
\]

For the pair of \( \text{Cov}[\lambda_t, Q_t] \) terms, we use Lemma 3.8 in conjunction with the explicit function for \( \text{Cov}[\lambda_t, Q_t] \) to find

\[
\int_0^t \left( e^{-S^T s} \theta \text{Cov}[\lambda_s, Q_s]^T e^{-Ss} + e^{-S^T s} \text{Cov}[\lambda_s, Q_s] \theta^T e^{-Ss} \right) ds
\]

\[
= \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)} ((\beta - \alpha) I - S)^{-1} \left( 2(\beta - \alpha) M_{\theta, \theta, S}(t) + e^{-S^T t} \theta \theta^T e^{-St} - \theta \theta^T \left( e^{-(\beta - \alpha) S^T t} I \right) \theta \theta^T \right)
\]

\[
\cdot ((\beta - \alpha) I + S)^{-1} ((\beta - \alpha) I - S) + ((\beta - \alpha) I - S^T) ((\beta - \alpha) I + S^T)^{-1} \left( e^{-(\beta - \alpha) S^T t} I \right) \theta \theta^T
\]

\[
\cdot ((\beta - \alpha) I - S)^{-1} + \frac{\alpha(\beta_0 - \lambda_\infty)}{\beta - \alpha} (S^T)^{-1} \left( (\beta - \alpha) M_{\theta, \theta, S}(t) + e^{-(\beta - \alpha) S^T t} \theta \theta^T e^{-St} - \theta \theta^T \left( e^{-(\beta - \alpha) S^T t} I \right) \theta \theta^T
\]

\[
- \theta \theta^T \left( e^{-(\beta - \alpha) S^T t} I \right) ((\beta - \alpha) I + S)^{-1} S - S^T ((\beta - \alpha) I + S^T)^{-1} \left( e^{-(\beta - \alpha) S^T t} I \right) \theta \theta^T
\]

\[
S^{-1} - \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} ((\beta - \alpha) I + S^T)^{-1} \left( e^{-(2\beta - \alpha) S^T t} \theta \theta^T e^{-St} - \theta \theta^T - \theta \theta^T \left( e^{-(\beta - \alpha) S^T t} I \right) \theta \theta^T
\]

\[
- \left( e^{-(\beta - \alpha) S^T t} I \right) \theta \theta^T \right) ((\beta - \alpha) I + S)^{-1}
\]

24
and so we now integrate the remaining terms in the covariance matrix differential equations. Note that the product rule for three terms is \((fg)’ = f’g + fg’ + fgh’\). 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 \(\frac{d}{dt} E[Q_t] = S^T E[Q_t] + \theta E[\lambda_t]\). Using this realization, the integral of the remaining three terms is

\[
\int_0^t \left( e^{-St} \text{diag} \left( \theta E[\lambda_s] \right) + S^T E[Q_s] e^{-Ss} - e^{-St} \text{diag} \left( E[Q_s] \right) e^{-Ss} - e^{-St} \text{diag} \left( E[Q_s] \right) S e^{-Ss} \right) ds = e^{-St} \text{diag} \left( E[Q_t] \right) e^{-St} \theta e^{-St} - e^{-St} \text{diag} \left( \left( S^T \right)^{-1} \left( I - e^{St} \right) \right) e^{-St} \text{diag} \left( \left( S^T + (\beta - \alpha)I \right)^{-1} \left( e^{-(\beta - \alpha)t} I - e^{St} \right) \right)
\]

where here 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.

We now apply Theorem 3.9 to common phase-type distributions, starting with the 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, \(S^T = n\mu(N - I)\) 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 \(\theta = v_1\) as all arrivals occur in the first phase. With this in hand, we see that

\[
(M_{\gamma,v_1,n\mu(N^T-I)}(t)_{i,j} = \left(M_{\gamma-n\mu,v_1,n\mu N^T}(t)_{i,j}\right) = \left\{ \begin{array}{ll}
\binom{i+j-2}{i-1} \binom{n\mu}{i+j-2} e^{(\gamma-n\mu)t} \sum_{k=0}^{(i+j-2)(i-1)/2} \binom{(i+j-2)(i-1)/2}{k} e^{-(\gamma-n\mu)(i-j-1)} & \text{if } \gamma \neq n\mu \\
\binom{i+j-2}{i-1} \binom{n\mu}{i+j-1} e^{-(\gamma-n\mu)(i-j-1)} & \text{if } \gamma = n\mu
\end{array} \right.
\]

and we make use of this in the following corollary.

**Corollary 3.10.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process \((\lambda_t, N_t)\) with dynamics given in Equation 2.4 and Erlang distributed service with \(n\) phases and mean \(\frac{1}{\mu}\). Then, if \(n\mu - \beta + \alpha \neq 0\),

\[
E[Q_t] = \frac{\lambda_\infty}{n\mu} (I - N)^{-1} (I - e^{-n\mu(I-N)t}) v_1 - (\lambda_\infty - \lambda_\infty) (n\mu N - (n\mu - \beta + \alpha)I)^{-1} e^{-(\beta - \alpha)t} I - e^{-n\mu(I-N)t}) v_1
\]

where \(\lambda_\infty = \frac{\beta \lambda^*}{\beta - \alpha}\). Further,

\[
\text{Cov}[\lambda_t, Q_t] = \lambda_\infty \left( \alpha + \frac{\alpha^2}{2(\beta - \alpha)} \right) ((n\mu + \beta - \alpha) I - n\mu N)^{-1} \left( I - e^{(n\mu N - (n\mu + \beta - \alpha))t} \right) v_1 + \frac{\alpha^2(2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} \left( (n\mu N - (n\mu - \beta + \alpha)I)^{-1} (I - e^{n\mu N - (n\mu + \beta - \alpha)I}) v_1 \right.
\]

\[
\cdot (n\mu N - (n\mu - \beta + \alpha) I)^{-1} (e^{-(\beta - \alpha)t} I - e^{(n\mu N - (n\mu + \beta - \alpha))t}) v_1.
\]
Finally, if \( n\mu - 2(\beta - \alpha) \neq 0 \) as well, then

\[
\text{Cov} [Q_t, Q_t] = \frac{\alpha (2\beta - \alpha) \lambda_\infty}{2(\beta - \alpha)} \left( ((n\mu + \beta - \alpha)I - n\mu N)^{-1} \left( 2(\beta - \alpha) e^{n\mu(N-I)t} M_{-2n\mu, \nu, 1, n\mu N} (t) e^{n\mu(N-I)t} \\
+ \nu_1 \nu_1^T - e^{n\mu(N-I)t} \nu_1 \nu_1^T e^{n\mu(N-I)t} + e^{n\mu(N-I)t} \nu_1 \nu_1^T \right) (n\mu N^T - (n\mu - \beta + \alpha)I)^{-1} \\
\cdot ((n\mu + \beta - \alpha)I - n\mu N^T) + ((n\mu + \beta - \alpha)I - n\mu N)(n\mu N - (n\mu - \beta + \alpha)I)^{-1} \left( e^{-(\beta - \alpha)t} I - e^{n\mu(N-I)t} \right) \\
\cdot \nu_1 \nu_1^T e^{n\mu(N-I)t} \right) \right) \left( (n\mu + \beta - \alpha)I - n\mu N^T \right)^{-1} + \frac{\alpha \beta (\lambda_0 - \lambda_\infty)}{(n\mu)^2 (\beta - \alpha)} (N - I)^{-1} \left( (\beta - \alpha) e^{n\mu(N-I)t} \\
\cdot M_{-(n\mu + \beta - \alpha), \nu_1, n\mu N} (t) e^{n\mu(N-I)t} + e^{-(\beta - \alpha)t} \nu_1 \nu_1^T e^{n\mu(N-I)t} \nu_1 \nu_1^T e^{n\mu(N-I)t} - n\mu e^{n\mu(N-I)t} \nu_1 \nu_1^T \\
\cdot \left( e^{-(\beta - \alpha)t} I - e^{n\mu(N-I)t} \right) \nu_1 \nu_1^T e^{n\mu(N-I)t} \right) (N^T - I)^{-1} - \frac{\alpha^2 (2\lambda_0 - \lambda_\infty)}{2(\beta - \alpha)} (n\mu N - (n\mu - \beta + \alpha)I)^{-1} \\
\cdot \left( e^{-2(\beta - \alpha)t} \nu_1 \nu_1^T e^{n\mu(N-I)t} - e^{-(\beta - \alpha)t} \nu_1 \nu_1^T e^{n\mu(N-I)t} \nu_1 \nu_1^T \right) e^{-(\beta - \alpha)t} I - e^{n\mu(N-I)t} \\
- \left( e^{-(\beta - \alpha)t} I - e^{n\mu(N-I)t} \right) \nu_1 \nu_1^T e^{n\mu(N-I)t} \right) (n\mu N^T - (n\mu - \beta + \alpha)I)^{-1} - \frac{\lambda_\infty}{n\mu} \text{diag} \left( (N - I)^{-1} \\
\cdot \left( I - e^{n\mu(N-I)t} \right) \nu_1 \right) \left( (n\mu N - (n\mu - \beta + \alpha)I)^{-1} \left( e^{-(\beta - \alpha)t} I - e^{n\mu(N-I)t} \right) \nu_1 \right)
\right)
\]

where all \( t \geq 0 \).

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 \( \theta \theta^T \), giving us

\[
M_{\gamma, \beta, -D} (t) = \int_0^t e^{(\gamma I + D)s} \theta \theta^T e^{Ds} ds = \int_0^t e^{(\gamma I + 2D)s} ds \theta \theta^T = (\gamma I + 2D)^{-1} (e^{(\gamma I + 2D)t} - I) \theta \theta^T
\]

as long as \( \gamma I + 2D \) is invertible. Like with the Erlang, we now make use of this in the following corollary.

**Corollary 3.11.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process \( (\lambda_i, N_i) \) with dynamics given in Equation 2.4 and hyper-exponential distributed service with \( n \) phases each with service rate \( \mu_i \) for \( 1 \leq i \leq n \). Then, if \( D - (\beta - \alpha)I \) is invertible for \( D_{i,i} = \mu_i \) and \( \theta \) otherwise,

\[
E [Q_t] = \lambda_\infty D^{-1} (I - e^{-Dt}) \theta - (\lambda_0 - \lambda_\infty) (\beta - \alpha)I - D)^{-1} (e^{-(\beta - \alpha)t} I - e^{-Dt}) \theta
\]

(3.20)
where \( \lambda_\infty = \frac{\beta \lambda}{\beta - \alpha} \). Further, if \( D + (\beta - \alpha)I \) is invertible as well, then

\[
\text{Cov} [\lambda_t, Q_t] = \lambda_\infty \left( \alpha + \frac{\alpha^2}{2(\beta - \alpha)} \right) (D + (\beta - \alpha)I)^{-1} (I - e^{-(D + (\beta - \alpha)I)t}) \begin{pmatrix} \alpha \beta \lambda_0 - \lambda_\infty \\ \beta - \alpha \end{pmatrix} \]

where all \( t \geq 0 \).

Finally, if \( \beta \lambda_0 - \lambda_\infty \) is invertible as well, then

\[
\text{Cov} [\lambda_t, Q_t] = \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)} ((\beta - \alpha)I + D)^{-1} \left( \begin{array}{c} \alpha \beta \lambda_0 - \lambda_\infty \\ \beta - \alpha \end{array} \right) D^{-2} (\beta - \alpha) (2D - (\beta - \alpha)I)^{-1} (e^{-(\beta - \alpha)t} - e^{-2Dt}) \begin{pmatrix} \alpha \beta \lambda_0 - \lambda_\infty \\ \beta - \alpha \end{pmatrix} \]

where all \( t \geq 0 \).

In either the Erlang setting or the hyper-exponential setting, if \( n = 1 \) then we recover a simple exponential service setting. In that case, we have the following.

**Corollary 3.12.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process \((\lambda_t, N_t)\) with dynamics given in Equation 2.4 and exponentially distributed service with rate \( \mu \). Then, if \( \mu \neq \beta - \alpha \),

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

where \( \lambda_\infty = \frac{\beta \lambda}{\beta - \alpha} \). Further,

\[
\text{Cov} [\lambda_t, Q_t] = \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)(\mu + \beta - \alpha)} (1 - e^{-(\mu + \beta - \alpha)t}) + \frac{\alpha \beta (\lambda_0 - \lambda_\infty)}{\mu (\beta - \alpha)} (e^{-(\beta - \alpha)t} - e^{-(\mu + \beta - \alpha)t})
\]

Finally, if \( 2\mu \neq \beta - \alpha \),

\[
\text{Var} (Q_t) = \frac{\alpha(2\beta - \alpha)\lambda_\infty}{2(\beta - \alpha)(\mu + \beta - \alpha)} \left( \frac{1 - e^{-2\mu t}}{\mu} - 2 \frac{e^{-(\mu + \beta - \alpha)t} - e^{-2\mu t}}{\mu - \beta + \alpha} \right) + \frac{\lambda_\infty}{\mu} (1 - e^{-\mu t})
\]

where all \( t \geq 0 \).
We now use these findings to describe the equilibria of the queueing system.

**Corollary 3.13.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process \((\lambda_t, N_t)\) with dynamics given in Equation 2.4 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. If \(S\) is invertible, then

\[
\lim_{t \to \infty} E[Q_t] = \lambda_\infty (-S^T)^{-1} \theta
\]

(3.26)

where \(\lambda_\infty = \frac{\beta \lambda^*}{\beta - \alpha}\). Further, if \(S - (\beta - \alpha)I\) is invertible as well, then

\[
\lim_{t \to \infty} \text{Cov}[\lambda_t, Q_t] = \lambda_\infty \frac{\alpha(2\beta - \alpha)}{2(\beta - \alpha)} ((\beta - \alpha)I - S^T)^{-1} \theta.
\]

(3.27)

Finally,

\[
\lim_{t \to \infty} \text{Cov}[Q_t, Q_t] = \lambda_\infty \frac{\alpha(2\beta - \alpha)}{2(\beta - \alpha)} ((\beta - \alpha)I - S^T)^{-1} \theta \theta^T - \text{diag}\left( (S^T)^{-1} \theta \right) \lambda_\infty
\]

(3.28)

where all \(t \geq 0\).

**Proof.** The proof follows by either taking the limit of the equations in Proposition 3.9 or setting the corresponding differential equations to 0 and finding the equilibrium solution. In the equations method, each matrix exponential term is diagonally dominant with strictly negative values on the diagonal and thus has non-positive eigenvalues.

As before, these results compare quite nicely in numerical demonstrations. Below are plots of the simulated and deterministic mean and variance of the queueing system for the single phase case.

![Figure 3.1](image-url)

Figure 3.1: Mean (left) & variance (right) of \(Q_t\) in Hawkes/M/\(\infty\), \(\alpha = \frac{1}{2}\), \(\beta = \frac{3}{4}\), \(\lambda^* = \mu = 1\)

To complement these findings, we also derived a form for the moment generating function for a general queueing system driven by a Hawkes process.
Theorem 3.14. Consider a queueing system with arrivals occurring in accordance to a Hawkes process \((\lambda_t, N_t)\) with dynamics given in Equation 2.4 and phase-type distributed service. Let \(\delta \in \mathbb{R}^{n+1}_+\) and let \(M(\delta, t) = M(\delta_0, \ldots, \delta_n, t) = \mathbb{E}[e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}]\). Then, the moment generating function for the queueing system \(M(\delta, t)\) is given by the solution to the following partial differential equation,

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

(3.29)

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

\[
\frac{\partial M(\delta, t)}{\partial t} = \mathbb{E}\left[ \delta_0 \beta x M(\lambda_t, e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}}) + \sum_{j=1}^n \lambda_t \theta_j \left( e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} - 1 \right) + \sum_{k=1}^n \sum_{j \neq k} \mu_{jk} Q_{t,j} e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} - \sum_{j=1}^n \mu_{j0} Q_{t,j} e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} \right]
\]

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

\[
= \mathbb{E}\left[ \delta_0 \beta x e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} - \delta_0 \beta x e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} + \sum_{j=1}^n \lambda_t \theta_j \left( e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} - 1 \right) + \sum_{k=1}^n \sum_{j \neq k} \mu_{jk} Q_{t,j} e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} - \sum_{j=1}^n \mu_{j0} Q_{t,j} e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} \right]
\]

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

\[
= \delta_0 \beta x \mathbb{E}\left[ e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} \right] + \left( \delta_0 \beta + \sum_{j=1}^n \theta_j (e^{\delta_0 \alpha + \delta_j} - 1) \right) \mathbb{E}\left[ e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} \right] + \sum_{j=1}^n \left( \mu_{j0} (e^{-\delta_j} - 1) + \sum_{k \neq j} \mu_{jk} (e^{\delta_k - \delta_j} - 1) \right) \mathbb{E}\left[ Q_{t,j} e^{\delta_0 \lambda_t + \sum_{i=1}^n \delta_i Q_{t,i}} \right]
\]

29
Finally, here we recognize the form of partial derivatives of $M(\delta, t)$ in each expectation, and so we re-express accordingly.

$$
= \delta_0 \beta \lambda^* M(\delta, t) + \left( \delta_0 + \sum_{i=1}^{n} \theta_i (e^{\delta_0 \alpha + \delta_i} - 1) \right) \frac{\partial M(\delta, t)}{\partial \delta_0} \\
+ \sum_{j=1}^{n} \left( \mu_j^0 (e^{-\delta_j} - 1) + \sum_{k \neq j} \mu_{jk} (e^{\delta_k - \delta_j} - 1) \right) \frac{\partial M(\delta, t)}{\partial \delta_j}
$$

This is 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 3.15.** Consider a queueing system with arrivals occurring in accordance to a Hawkes process $(\lambda_t, N_t)$ with dynamics given in Equation 2.4 and phase-type distributed service. Let $\delta \in \mathbb{R}^{n+1}_+$ and let $G(\delta, t) = G(\delta_0, \ldots, \delta_n, t) = \log \left( \mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right] \right)$. Then, the cumulant moment generating function for the queueing system $G(\delta, t)$ is given by the solution to the following partial differential equation,

$$
\frac{\partial G(\delta, t)}{\partial t} = \delta_0 \beta \lambda^* + \left( \delta_0 + \sum_{i=1}^{n} \theta_i (e^{\delta_0 \alpha + \delta_i} - 1) \right) \frac{\partial G(\delta, t)}{\partial \delta_0} \\
+ \sum_{i=1}^{n} \left( \mu_i^0 (e^{-\delta_i} - 1) + \sum_{k \neq i} \mu_{ik} (e^{\delta_k - \delta_i} - 1) \right) \frac{\partial G(\delta, t)}{\partial \delta_i}.
$$

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

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

and here we can recognize that these expectations are the moment generating function. Using Theorem 3.14, we have

$$
= \delta_0 \beta \lambda^* + \left( \delta_0 + \sum_{i=1}^{n} \theta_i (e^{\delta_0 \alpha + \delta_i} - 1) \right) \frac{\partial}{\partial \delta_0} \frac{\mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right]}{\mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right]}
\\
+ \sum_{i=1}^{n} \left( \mu_i^0 (e^{-\delta_i} - 1) + \sum_{k \neq i} \mu_{ik} (e^{\delta_k - \delta_i} - 1) \right) \frac{\partial}{\partial \delta_i} \frac{\mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right]}{\mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right]}.
$$

Now we recognize that $\frac{\partial}{\partial \delta_i} \frac{\mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right]}{\mathbb{E} \left[ e^{\delta_0 \lambda_t + \sum_{i=1}^{n} \delta_i Q_{t,i}} \right]} = \frac{\partial G(\delta, t)}{\partial \delta_i}$, 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.
3.3 Comparison to Poisson Arrival Queues

While we have already observed that the Hawkes process differs from the Poisson process alone, it is not clear that this will remain true in a queue system. To investigate this, we compare the limit distribution of a simulated Hawkes driven queue and the known theoretical limit distribution of a queue with Poisson arrivals. Note that for the latter the limit distribution is equivalent to a Poisson distribution with rate equal to the arrival rate divided by the service rate. Using this, we create the following plots via 1,000 simulation replications.

![Figure 3.2: Limit Distributions for $\lambda^* = 10, \mu = 1, \beta = 2, \text{ and } \alpha = 0$ (left) and $1.2$ (right).](image)

Observe that like in process comparison, the Hawkes arrival queue and Poisson arrival queue are equivalent when $\alpha = 0$. Additionally, as $\alpha \to \beta$ the two queues become dissimilar. In particular, when $\alpha$ is very near $\beta$ we see that the variance of the Hawkes driven queue is larger than that of the Poisson queue.

4 Applications

To motivate this study and demonstrate its findings, we now list three different applications. The first model is motivated by flight cancellations. These often occur in clusters because some reasons impact many flights at once and one cancelled flight often causes several more to be cancelled in a chain reaction. This then causes passengers to contact the airline help centers in clusters. For this case we investigate Markov and Chebyshev bounds via simulation. The second model is for internet users arriving to websites in clusters according to trending social links. For this setting we describe the phenomena through observed data and investigate the value and influence a website gets from a click. The third model considers the managerial aspect of operating and promoting a night club, based on the observation that in this setting a customer infers value from seeing other customers being willing to wait. This setting motivates an optimal control problem for deciding how quickly to admit new customers, as delaying earning service revenue can increase the potential for attracting additional business.
4.1 Flight Cancellations with Impacts on Call Centers

As noted in a recent Business Insider article, flights are typically cancelled in mass [29]. This occurs for two primary reasons. First, large scale events like inclement weather and technological failure will affect a number of flights occurring within the same time range, naturally causing a cluster of cancellations. Secondly, even when a flight is cancelled for a seemingly confined event like mechanical failure, the downstream flights will also be cancelled if a replacement plane cannot be found. Thus, a service system arises out of handling these cancellations and the passengers they have left stranded. These passengers will seek assistance at airport help counters and through airline customer service phone lines. This motivates us to model this airline call center system as a queue driven by Hawkes process arrivals, as clusters of cancelled flights will produce clusters of stranded passengers. Using a phase-type distribution of size $n$, we can bound the total number of calls in system as

$$P \left( \sum_{i=1}^{n} Q_{t,i} \geq x \right) \leq \frac{E \left[ \sum_{i=1}^{n} Q_{t,i} \right]}{x} = \sum_{i=1}^{n} \frac{E [Q_{t,i}]}{x}$$

via the Markov inequality where here we are simply using generic phase-type service [28]. However, because of the clustered arrivals, variability should be addressed. By Chebyshev’s inequality and the definitions of variance and covariance, we have that

$$P \left( \left| \sum_{i=1}^{n} Q_{t,i} - \sum_{i=1}^{n} E [Q_{t,i}] \right| \geq x \right) \leq \frac{\text{Var} \left( \sum_{i=1}^{n} Q_{t,i} \right)}{x^2} = \sum_{i=1}^{n} \frac{\text{Var} (Q_{t,i})}{x^2} + 2 \sum_{j \neq i} \text{Cov} [Q_{t,i}, Q_{t,j}]$$

where $E [Q_t]$, Var $(Q_t)$, and Cov $[Q_t, Q_t]$ can be supplied via Theorem 3.9. We now examine this bound for single phase service via simulation. Below are plots of the gaps between the theoretical bounds and observed probability from the simulations.

Figure 4.1: Markov bound probability gap, where $\lambda^* = \beta = \mu = 1$ and $\alpha = .5$. 

32
Each of these figures is formed by calculating the theoretical inequality and then subtracting the empirical probability based on 10,000 simulation replications. So that the Markov bound was not trivial, the \( x \) values start at \( \lambda_{\infty} = 2 \), which in this case is equal to the limit of \( E[Q_t] \). While both bounds become more reliable as \( x \) grows large, the Chebyshev bound approaches tightness faster than the Markov bound does.

Figure 4.2: Chebyshev bound probability gap, where \( \lambda^* = \beta = \mu = 1 \) and \( \alpha = .5 \).

### 4.2 Trending Web Traffic

In the May 2017 Alexa website rankings for the Unites 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 [1]. 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 mean radical jumps in exposure, interest, and revenue.

As a basic example of a link that became trending, we analyzed publicly available Twitter data [12]. 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. However, even outside of this window
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 4.3: Tweets of Young Jeezy - My President music video from November 5 - 7, 2012.](image)

While performing significant inference on one sample path can risk overfitting, if one
assumes that the process starts in steady state (i.e. $\lambda_0 = \lambda_\infty$) then we can see from Theo-
rem 3.9 that $E[N_t] = \lambda_\infty t$. Thus, by the Law of Large Numbers we could take
\[
\frac{\text{Number of Arrivals}}{\text{Time}}
\]
as an estimate of $\lambda_\infty$. We can also make observations about the conditions in which this type
of intense viral activity may occur in a Hawkes process. 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,
suggesting that small baseline intensities combined with very large intensity jump sizes may
create the conditions needed for a majority-containing cluster like we see in the Twitter
example.

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 $\lambda^*$, initial intensity $\lambda_0$, jump size $\alpha$, and decay parameter $\beta$. Now, let $N^*_t$
represent an independent Hawkes process that is identical to $N_t$ 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 $N^*_t$ as they were for $N_t$, but the
initial intensity is $\lambda_0 + \alpha$ and $N_0^* = 1$. Then, by Theorem 3.9,

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

which shows that the gap between the two expectations is positive and grows throughout time. In the web clicks context, this shows how much influence a single user can have on a site’s traffic. However, this is simply tracking the number of visitors but does not account for the time the users spend on the site consuming content, viewing advertisements, and otherwise interacting with the page. To capture this, we can extend the web user arrival model to a queueing model in which the service represents the time the user spends on the page.

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 $E[Q_T]$. Then, from time 0 to time $T$ the expected total time spent on the page across all users $\sigma(T)$ is

$$
\sigma(T) = \int_0^T E[Q_t] \, dt = \int_0^T \left(\frac{\lambda_\infty}{\mu} \left(1 - e^{-\mu t}\right) + \frac{\lambda_0 - \lambda_\infty}{\mu - \beta + \alpha} \left(e^{-(\beta - \alpha)t} - e^{-\mu t}\right)\right) \, dt
$$

where we have applied the results of Theorem 3.9 for exponential service with $\mu \neq \beta - \alpha$. 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 $\mu$, infinite servers, and Hawkes process arrivals with parameters $\lambda^*$, $\alpha$, and $\beta$ and assume the queue starts empty. Then, let $Q_t^*$ be an identically distributed queueing system with the exception of initial intensity $\lambda^* + \alpha$ and one user in the system at the start. Let $A(T)$ and $A^*(T)$ be the corresponding expected ad 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,

$$
A^*(T) - A(T) = m\frac{1 - e^{-\mu T}}{\mu} + m\frac{\alpha}{\mu - \beta + \alpha} \left(1 - e^{-(\beta - \alpha)T}\right) - \frac{1 - e^{-\mu T}}{\mu}
$$

which can be shown to also always grow with $T$ via its first derivative.
based services and allow customers to upload pictures from their smart phones as soon as they are taken, the Hawkes/M/∞ queue can be used to describe the number of pictures being uploaded at once. In this case Hawkes process arrivals capture that smart phone users often take several pictures in one short setting and then not again for a longer interval. Again, for reliable data center support this is appropriately modeled as an infinite server queue.

### 4.3 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. However, this scenario actually claims partial inspiration for this work as it 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. For this application, consider the phase-type to be a two-step service. The first step can be considered “admittance” to the service with the second step being the service itself. This process is visualized below, where $\mu_O$ and $\mu_I$ are the rates of each step of service.

![Figure 4.4: Club Queue Process Diagram](image)

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. This is non-trivial as a short line outside the club might signal to others that the club is not interesting and no one wants to go inside, however, if the line is too long, there are many customers not actively generating revenue for the club. 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 to party.

$$\zeta(\mu_O(t)) = r_O \mu_O E[Q_O(t)] + r_I E[Q_I(t)] - c(\mu_O E[Q_O(t)] - k)^2 - w \mu_O^2$$  \hspace{1cm} (4.1)

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 4.1 (Unconstrained Club Profit Model).

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

subject to
\[
\begin{align*}
\bullet & \quad E[\lambda(t)] = \beta \cdot (\lambda^* - E[\lambda(t)]) + \alpha \cdot E[\lambda(t)] \\
\bullet & \quad E[Q_O(t)] = E[\lambda(t)] - \mu_O(t) \cdot E[Q_O(t)] \\
\bullet & \quad \dot{E}[Q_I(t)] = \mu_O(t) \cdot E[Q_O(t)] - \mu_I \cdot E[Q_I(t)]
\end{align*}
\]

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 4.1 (Unconstrained Club Profit Optimal Solution). The optimal solution to 4.1 is given by \( \mu_O^*(t) \), where

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

for all \( t \geq 0 \).

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

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

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 \( \frac{d\mathcal{H}}{d\mu_O(t)} = 0 \) for all \( t \geq 0 \). We see that the derivative of the Hamiltonian with respect to \( \mu_O(t) \) is

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

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

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

for \( \mu_O^*(t) \), which gives

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

for \( t \geq 0 \). Because the objective function is concave in \( \mu_O(t) \) at every \( t \), we have that this solution corresponds to a maximum. \( \square \)
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 [15], Qin and Pender [27], Lenhart and Workman [9]. We now give two example outputs of this method below.

Figure 4.5: 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, \( w \), is 150. On the right, \( w \) is instead 100 and is \( k = 12 \). 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 \( \mu_O \) 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 distributions. We are able to derive the exact moments and moment generating function for the Hawkes driven queueing process 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 [11], Pender [19], Engblom and Pender [5], Pender [23, 21, 22, 24]. 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 somewhat analytically 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 [10], Pender [20, 21, 23], Engblom and Pender [5] 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 [25]. 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. We plan to explore these extensions in subsequent work.

References


Appendix

Proof of Proposition 2.1

Proof. This can be demonstrated by first re-expressing $\lambda_{t+\delta}$ in terms of time before $t$ and after $t$ for any $\delta > 0$.

\[
\lambda_{t+\delta} = \lambda^* + e^{-\beta(t+\delta)}(\lambda_0 - \lambda^*) + \alpha \int_0^{t+\delta} e^{-\beta(t+\delta-u)}dN_u
\]

\[
= \lambda^* + e^{-\beta(t+\delta)}(\lambda_0 - \lambda^*) + \alpha \int_0^{t} e^{-\beta(t+\delta-u)}dN_u + \alpha \int_t^{t+\delta} e^{-\beta(t+\delta-u)}dN_u
\]

\[
= (1 - e^{-\beta\delta} + e^{-\beta\delta})(\lambda^* + e^{-\beta(t+\delta)}(\lambda_0 - \lambda^*) + \alpha e^{-\beta\delta} \int_0^{t} e^{-\beta(t-u)}dN_u + \alpha \int_t^{t+\delta} e^{-\beta(t+\delta-u)}dN_u
\]

\[
= (1 - e^{-\beta\delta})\lambda^* + e^{-\beta\delta}\alpha + \alpha \int_t^{t+\delta} e^{-\beta(t+\delta-u)}dN_u
\]

Note that this form shows that the future values of the intensity function depend on the future arrivals and the present value of the intensity function, but none of the past. Now, we continue by conditioning the density of $(\lambda_t, N_t)$ on the future value of $N_{t+\delta}$, allowing for two separate probability functions.

\[
f((\lambda_{t+\delta}, N_{t+\delta}) = (l, j) \mid (\lambda_t, N_t) = (l_0, j_0), F_t)
\]

\[
= \sum_{n=0}^\infty f(\lambda_{t+\delta} = l \mid N_{t+\delta} = j, (\lambda_t, N_t) = (l_0, j_0), F_t) \cdot P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0), F_t)
\]

Now that the density is only evaluated on $\lambda_{t+\delta}$, we can use the form just shown above.

\[
= f\left((1 - e^{-\beta\delta})\lambda^* + e^{-\beta\delta}\lambda_t + \alpha \int_t^{t+\delta} e^{-\beta(t+\delta-u)}dN_u = l \mid N_{t+\delta} = j, (\lambda_t, N_t) = (l_0, j_0), F_t\right)
\]

\[
\cdot P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0), F_t)
\]

To follow conventional notation, we rearrange the terms inside the density to separate the uncertain stochastic process from the deterministic constant and conditioned terms.

\[
= f\left(\int_t^{t+\delta} e^{-\beta(t+\delta-u)}dN_u = \frac{l - (1 - e^{-\beta\delta})\lambda^* - e^{-\beta\delta}\lambda_t}{\alpha} \mid N_{t+\delta} = j, (\lambda_t, N_t) = (l_0, j_0), F_t\right)
\]

\[
\cdot P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0), F_t)
\]

Now, for this particular choice of excitation kernel function, the Lebesgue integral is equivalent to summing over the arrival times of the process’s events. Since this density is conditioned on $j - j_0$ arrivals occurring in the time period $[t, t + \delta)$, this can be substituted into the equation.

\[
= f\left(\sum_{k=j_0}^j e^{-\beta(t+\delta-S_k)} = \frac{l - (1 - e^{-\beta\delta})\lambda^* - e^{-\beta\delta}\lambda_t}{\alpha} \mid N_{t+\delta} = j, (\lambda_t, N_t) = (l_0, j_0), F_t\right)
\]

\[
\cdot P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0), F_t)
\]
Now we can investigate each of these two probabilities separately. For $N_{t+\delta}$, we know that the probability of $j-j_0$ arrivals occurring in this interval depends only on the intensity at that point in time. Since the intensity only depends on the most recent value, the full filtration is not needed and so $P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0), \mathcal{F}_t) = P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0))$. By the same argument, the density of the arrival times is also conditionally independent of the filtration given the present value of the intensity, as it again only relies on the intensity and not the full history. Thus, each of our probability functions is independent of the filtration.

\[= f \left( \sum_{k=j_0}^{j} e^{-\beta(t+\delta-S_k)} = \frac{l - (1 - e^{-\beta\delta})\lambda^* - e^{-\beta\delta}l_0}{\alpha} \right \mid N_{t+\delta} = j, (\lambda_t, N_t) = (l_0, j_0)) \cdot P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0)) \]

Now, by rearranging we have the following.

\[= f (\lambda_{t+\delta} = l \mid N_{t+\delta} = j, (\lambda_t, N_t) = (l_0, j_0)) \cdot P(N_{t+\delta} = j \mid (\lambda_t, N_t) = (l_0, j_0)) \]

We now recover an un-enforced value $j$ for $N_{t+\delta}$ by reversing the conditioning, leaving us with

\[= f ((\lambda_{t+\delta}, N_{t+\delta}) = (l, j) \mid (\lambda_t, N_t) = (l_0, j_0)) \]

which finally shows that the Markov property holds for $(\lambda_t, N_t)$. \qed