On the time-dependent behavior of finite-state birth-death processes: new insights on known results

Xiaoyuan Liu and Brian Fralix
School of Mathematical and Statistical Sciences, and CORI
Clemson University
Clemson, SC, USA

May 5, 2019

Abstract

We illustrate how the knockout queue construction given recently in [8] can be used to derive, probabilistically, expressions for the transition functions of any finite-state birth-death process. The expressions we derive can also be found in the work of A. Zeifman on deriving useful/simple upper and lower bounds for the convergence rate of a birth-death process: see e.g. [6]. Our approach makes use of a combination of point process theory, phase-type distributions, and the 'knockout queue' construction recently featured in [8]. We also illustrate how various observations made throughout our derivations can be used to establish other interesting (known and unknown) structural properties of birth-death processes.

Keywords: birth-death process, convergence rate, time-dependent behavior
2010 MSC: 60J27, 60G55, 60K25

1 Introduction

Suppose $\{Q(t); t \geq 0\}$ is a finite-state birth death process, whose state space is given by $S := \{0, 1, 2, \ldots, N-1, N\}$, for some finite positive integer N. Associated with $\{Q(t); t \geq 0\}$ is its generator matrix $\mathbf{Q} := [q(i,j)]_{i,j \in S}$ as well as a collection of transition functions $\{p_{i,j}\}_{i,j \in S}$, where, for each $i, j \in S$ (where possibly $i = j$), $p_{i,j} : [0, \infty) \to [0, 1]$ is defined as

$$p_{i,j}(t) := \mathbb{P}(Q(t) = j \mid Q(0) = i), \qquad t \geq 0.$$

The generator matrix $\mathbf{Q}$ is completely determined by the birth rates $\{\lambda_i\}_{i=0}^{N-1}$ and the death rates $\{\mu_i\}_{i=1}^{N}$ of $\{Q(t); t \geq 0\}$, since

$$q(0,1) = \lambda_0 = -q(0,0), \quad q(N, N-1) = \mu_N = -q(N,N)$$

and for each state $i \in \{1, 2, \ldots, N-1\}$,

$$q(i, i-1) = \mu_i, \quad q(i, i+1) = \lambda_i, \quad q(i,i) = -(\lambda_i + \mu_i).$$

All other elements of $\mathbf{Q}$ are equal to zero.

We assume throughout that $\{Q(t); t \geq 0\}$ is irreducible, which clearly holds if and only if all birth rates $\{\lambda_i\}_{0 \leq i \leq N-1}$ and all death rates $\{\mu_i\}_{1 \leq i \leq N}$ are positive. This assumption, combined with the assumption that $N < \infty$, further implies $\{Q(t); t \geq 0\}$ has a unique stationary distribution $\mathbf{p} := (p_0, p_1, \ldots, p_N)$ which satisfies, for each $i, j \in S$,

$$p_j = \lim_{t \to \infty} p_{i,j}(t).$$

It is well-known that $p_{i,j}(t)$ converges exponentially fast to p_j as $t \to \infty$, and the rate governing this speed of convergence is the largest nonzero eigenvalue of $\mathbf{Q}$ (which is a negative number: recall that due to reversibility, all eigenvalues of a birth-death process are real and nonpositive). In this finite-state setting, exponential convergence quickly follows from the Jordan canonical form of $\mathbf{Q}$, see e.g. Horn and Johnson [11]. Readers should recall also that one eigenvalue of $\mathbf{Q}$ is the zero eigenvalue, associated with which is its row eigenvector $\mathbf{p}$, the stationary distribution of $\{Q(t); t \geq 0\}$.

Our objective is to derive a computable expression for each transition function of $\{Q(t) : t \geq 0\}$ that yields additional insight into structural properties of the underlying birth-death process, but in order to present our main result we first need to set up some additional notation. Let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N \in \mathbb{R}^{N \times 1}$ be unit column vectors, where for each $j \in \{1, 2, \ldots, N\}$, the jth element of $\mathbf{e}_j$ is equal to one, and all other elements of $\mathbf{e}_j$ are equal to zero. Both $\mathbf{e}_0$ and $\mathbf{e}_{N+1}$ will represent the zero vector $\mathbf{0} \in \mathbb{R}^{N \times 1}$; it will be helpful to make use of both symbols in our formulas.

Next, given a row vector $\boldsymbol{\gamma} := [\gamma_0, \gamma_1, \ldots, \gamma_N] \in \mathbb{R}^{1 \times (N+1)}$ whose elements are nonnegative and sum to one, we define the row vector $\boldsymbol{\gamma}_e := [\gamma_{e,1}, \ldots, \gamma_{e,N}] \in \mathbb{R}^{1 \times N}$, whose jth element satisfies, when $\gamma_0 < 1$,

$$\gamma_{e,j} := \frac{\sum_{k=j}^{N} \gamma_k}{\sum_{k=1}^{N} k\gamma_k}, \quad 1 \leq j \leq N.$$

$\boldsymbol{\gamma}_e$ is sometimes referred to as the *equilibrium* distribution associated wtih $\boldsymbol{\gamma}$. If $\boldsymbol{\gamma}$ has mean zero, meaning if $\gamma_0 = 1$, we simply set $\boldsymbol{\gamma}_e$ to be the zero vector in $\mathbb{R}^{1 \times N}$.

We further associate with $\{Q(t); t \geq 0\}$ the matrix $\mathbf{C} := [c(i,j)]_{1 \leq i,j \leq N}$, whose elements are defined as follows:

$$c(i,j) := \begin{cases} \lambda_i, & 1 \leq i \leq N-1, j = i+1; \\ \mu_{i-1}, & 2 \leq i \leq N, j = i-1; \\ -(\lambda_{i-1} + \mu_i), & 1 \leq i \leq N, j = i; \\ 0, & \text{otherwise}. \end{cases} \tag{1}$$

It will be shown in Section 2 that this matrix $\mathbf{C}$ has a special probabilistic interpretation for an important class of birth-death processes, but in general, this matrix will still appear in the transition functions even when it does not have a clear probabilistic interpretation.

The main objective of this study is to establish the following result.

Theorem 1.1 *The transition functions $\{p_{i,j}\}_{i,j \in S}$ of $\{Q(t); t \geq 0\}$ are as follows: for each $i, j \in S$,*

$$p_{i,j}(t) = p_j - \mathbb{E}[Q(\infty)]\mathbf{p}_e^T e^{\mathbf{C}t}(\mathbf{e}_j - \mathbf{e}_{j+1}) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}(\mathbf{e}_j - \mathbf{e}_{j+1}) \tag{2}$$

for each real $t \geq 0$, where $Q(\infty)$ is a random variable whose probability mass function is $\mathbf{p}$. More generally, when the probability mass function of $Q(0)$ is $\boldsymbol{\gamma}$, we have

$$\mathbb{P}(Q(t) = j) = p_j - \mathbb{E}[Q(\infty)]\mathbf{p}_e^T e^{\mathbf{C}t}(\mathbf{e}_j - \mathbf{e}_{j+1}) + \mathbb{E}[Q(0)]\boldsymbol{\gamma}_e^T e^{\mathbf{C}t}(\mathbf{e}_j - \mathbf{e}_{j+1}). \tag{3}$$

It is important at this point to note that Theorem 1.1 has been discovered before by A. Zeifman in [19]. In [19], Zeifman uses the Kolmogorov Forward Equations associated with $\{Q(t); t \geq 0\}$ to derive an alternative linear system of ordinary differential equations, from which the matrix $\mathbf{C}$ within Theorem 1.1 can be found through a similarity transformation. This analytic approach is clearly explained in the paper of Van Doorn, Zeifman and Panfilova [6], and we refer readers interested in this approach to [6], as well as to many of the papers cited therein. Our contribution to this stream of research is a new derivation of this result, that explains how the matrix $\mathbf{C}$ can be interpreted probabilistically, through the 'knockout queue' construction found in [8]. Moreover, our derivations also yield additional observations that can be used to derive, very quickly, other interesting structural properties of birth-death processes as well, both known and unknown. For example, one can use our derivation to show that when $Q(0) = 0$, $Q(t)$ is stochastically increasing in t (this is a known fact, see e.g. Theorem 6.1 of Van Doorn [5]). Not only that, our derivation also gives what appear to be new sufficient conditions for $\mathbb{E}[Q(t)]$ to be a concave function in t when $Q(0) = 0$.

Theorem 1.1 clearly indicates that we can express each transition function in terms of a constant, plus a function that is bounded, uniformly continuous, and integrable on $[0, \infty)$. It will be shown later that $(-\mathbf{C})^{-1}$ can be expressed in closed-form (meaning each element can be expressed explicitly in terms of the elements of $\mathbf{p}$) which means the integral of $p_{i,j}(t) - p_j$ (i.e. the (i, j)th element of the deviation matrix, see e.g. Coolen-Schrijner and Van Doorn [4] for more information on deviation matrices) and related integrals can be calculated exactly. Theorem 1.1 also allows us to make use of the uniformization-like technique outlined in Chapter 2 of Latouche and Ramaswami [15], typically used to calculate phase-type distributions and densities, to numerically calculate the transition functions as well: what is interesting about this procedure is that the number of terms needed in the approximation is independent of t, and the procedure yields an approximation that is uniformly close to the true function over the entire nonnegative real line.

This paper is organized as follows. In Section 2, we introduce the knockout queue from [8], which is relevant to this study because a birth-death process can alternatively be interpreted as a queue-length process of a knockout queue if its birth and death rates satisfy certain ordering properties. Once we state and prove a few simple results on the behavior of terminating phase-type renewal processes, we then show how the knockout queue interpretation can be used to derive, probabilistically, the transition functions of birth-death processes that can be associated with a knockout queue. Next, in Section 3 we show that the transition function values derived in Section 2 still hold for arbitrary finite-state birth-death processes. Finally, in Section 4, we illustrate how these formulas can be used to study structural properties of birth-death processes.

2 The Knockout Queue

Throughout this section, we assume that the birth and death rates of $\{Q(t); t \geq 0\}$ are nondecreasing and nonincreasing, respectively, with respect to the state variable. In other words, the birth rates satisfy

$$\lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{N-1} \tag{4}$$

and the death rates satisfy

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_N. \tag{5}$$

There are many examples of queueing systems whose rates satisfy (4) and (5): examples include the M/M/1/N queue, the M/M/s/N queue, as well as simple variations of these queueing systems where an arriving customer may choose not to enter the system based on its state upon his/her arrival, and/or an arriving customer may choose to abandon the system after waiting in the queue for a certain (exponentially distributed) amount of time.

In [8], it was shown that when $\{Q(t); t \geq 0\}$ satisfies (4) and (5), $\{Q(t); t \geq 0\}$ can be interpreted as the queue-length process of what is referred to in [8] as a 'knockout queue', in that for each $t \geq 0$, $Q(t)$ represents the number of customers present in this type of queueing system at time t. This is an interesting observation, primarily because the knockout queue is a queueing system where the sojourn time of an arbitrary customer only depends on both the work brought by that customer, as well as both the work brought, and behavior of, future customers arriving after that customer. An example of a queueing system that satisfies this property is the M/G/1 queue that processes work in accordance to the Last-Come-First-Served Preemptive-Resume discipline, and it is precisely this property that makes many performance measures of this queue tractable. In [8], the knockout queue interpretation of $\{Q(t); t \geq 0\}$ is used to study properties of the mean $\mathbb{E}[Q(t)]$ and the probabilities $\mathbb{P}(Q(t) = N)$ as t varies: for example, one can use the knockout queue interpretation to quickly show that $\mathbb{E}[Q(t)]$ is a concave function of t. This concavity result was shown in [8] by applying the transient Little's law (see [7]) to the knockout queue, but later we will see that this concavity structure is still sometimes preserved when $\{Q(t); t \geq 0\}$ can no longer be interpreted as the queue-length process of a knockout queue.

Throughout this study, any birth-death process whose birth and death rates satisfy (4) and (5) will be said to satisfy the *knockout queue criterion*.

2.1 Constructing the Knockout Queue

This construction was first presented in [8], but we review it here for convenience.

We consider a single-server queueing system consisting of N slots, where each slot can hold at most one customer. Customers arrive to this queueing system in accordance to a homogeneous Poisson process $\{N_0(t); t \geq 0\}$ having arrival rate λ_0, and each arrival brings with it a unit exponentially distributed amount of work for processing, independently of everything else. Customers always occupy the lowest-numbered slots in the system, meaning that if there are k customers present in the system, those customers can be found in slots $1, 2, \ldots, k$. While a customer is present in slot k, it receives processing from the server at a rate of $\mu_k - \mu_{k-1}$, $1 \leq k \leq N$, where $\mu_0 := 0$. As soon as a customer in slot k is served, it leaves the system, and all customers occupying slots $k + 1$ and higher shift down one slot. For instance, if there are currently 5 customers in the system, and the customer in slot 3 has just completed service, then the customers previously in slots 4 and 5 shift to slots 3 and 4, respectively.

This queueing system is referred to in [8] as a 'knockout queue', due to how an arriving customer behaves at each arrival instant. Namely, when a customer arrives to the system, it chooses to 'knockout', or eliminate, the customer currently occupying slot k with probability $(\lambda_{k-1} - \lambda_k)/\lambda_0$, for $1 \leq k \leq N$. If there is no customer present in slot k at this arrival time, then no customers are eliminated from the system, each customer currently in the system moves up one slot, and the new customer enters slot 1: otherwise, if there is a customer present in slot k at this arrival time, that customer is eliminated, and for each $\ell \leq k-1$, the customer formerly in slot ℓ moves to slot $\ell+1$, and the new arrival moves to slot 1. We note that when a customer in slot N is eliminated from the system by a new arrival, we always think of that elimination as occurring due to overcrowding: this will be a simple, yet important trick that we will make use of later on in our analysis. Previously in [8], this knockout procedure was described in terms of conditional distributions, but that description is probabilistically equivalent to the description given here, and this description is arguably simpler to comprehend.

A bit of thought—see [8]—shows that $\{Q(t); t \geq 0\}$ is probabilistically equivalent to the queue-length process of the knockout queue, where $Q(t)$ is interpreted as the number of customers present in the system at time t. This system has the following interesting property: the sojourn time of each customer in the system depends only on the amount of work brought by that customer, as well as the amounts of work, and 'knockout behaviors' of, all future arrivals to the system. In fact, we can measure the sojourn time of a given customer by keeping track of how that customer moves among slots in the queueing system, and these movements form another CTMC $\{B(t); t \geq 0\}$—referred to throughout as the customer CTMC—whose state space is $\{0, 1, 2, \ldots, N\}$, where state 0 is an absorbing state. When the chain reaches state 0, the customer is said to have left the system.

The generator of this CTMC is given by $\mathbf{B} := [b(i,j)]_{i,j}$, whose elements are as follows:

$$b(i,j) := \begin{cases} \lambda_{i-1} - \lambda_i + \mu_i - \mu_{i-1}, & 1 \leq i \leq N, j = 0; \\ \lambda_i, & 1 \leq i \leq N-1, j = i+1; \\ \mu_{i-1}, & 2 \leq i \leq N, j = i-1; \\ -(\lambda_{i-1} + \mu_i), & 1 \leq i \leq N, j = i; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

We can express $\mathbf{B}$ in block-partitioned form as

$$\mathbf{B} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{c}_0 & \mathbf{C} \end{pmatrix} \tag{7}$$

where $\mathbf{C} \in \mathbb{R}^{N \times N}$ is a subintensity matrix as defined in (1) and $\mathbf{c}_0 = -\mathbf{C}\mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^{N \times 1}$ is a column vector of ones.

2.2 An Aside: Terminating Phase-Type Renewal Processes

Our approach towards deriving transition functions requires us to gain a slightly deeper understanding of terminating, phase-type renewal processes. This will be made clearer when we begin our derivations. Readers seeking an introduction to both phase-type distributions, and phase-type renewal processes are referred to e.g. Chapters 2 and 3 of Latouche and Ramaswami [15], and Chapter 3 of Bladt and Nielsen [2].

Consider a continuous-time Markov chain (CTMC) $\{Z(t); t \geq 0\}$, whose state space is given by $E = \{0, 1, 2, \ldots, N, \Delta\}$ and whose generator is given by $\mathbf{Z}$, where $\mathbf{Z}$ is of the form

$$\mathbf{Z} = \begin{pmatrix} 0 & \mathbf{0}^T & 0 \\ \mathbf{t}_0 & \mathbf{T} & \mathbf{t}_\Delta \\ 0 & \mathbf{0}^T & 0 \end{pmatrix}. \tag{8}$$

Here $\mathbf{T} := [t(i,j)]_{i,j} \in \mathbb{R}^{N\times N}$ is a subintensity matrix, $\mathbf{t}_0$ and $\mathbf{t}_\Delta := [t(k,\Delta)]_k$ are column vectors in $\mathbb{R}^{N\times 1}$, and $\mathbf{0} \in \mathbb{R}^{N\times 1}$ is a column vector whose elements are all equal to zero. In light of $\mathbf{Z}$ being a generator matrix, the elements of both $\mathbf{t}_0$ and $\mathbf{t}_\Delta$ must be nonnegative, and together $\mathbf{t}_0$ and $\mathbf{t}_\Delta$ must satisfy

$$\mathbf{Te} + \mathbf{t}_0 + \mathbf{t}_\Delta = \mathbf{0} \tag{9}$$

where $\mathbf{e} \in \mathbb{R}^{N\times 1}$ is a column vector whose components are all equal to one. We assume in our construction of $\{Z(t); t \geq 0\}$ that states $\{1, 2, \ldots, N\}$ are all transient states, and both states 0 and Δ are accessible from the set $\{1, 2, \ldots, N\}$.

Let $(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_N, \alpha_\Delta)$ represent the initial distribution of $\{Z(t); t \geq 0\}$, and define $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathbb{R}^{1\times N}$, so that the initial distribution of $\{Z(t); t \geq 0\}$ can alternatively be expressed in partitioned vector form as $(\alpha_0, \boldsymbol{\alpha}, \alpha_\Delta)$. From both $\{Z(t); t \geq 0\}$ and its initial distribution, we define the random variable τ_Δ as

$$\tau_\Delta := \inf\{t \geq 0 : Z(t) = \Delta\}. \tag{10}$$

Readers should note that τ_Δ has a distribution that looks very similar to the distribution of a phase-type random variable, but here it is possible that $\tau_\Delta = \infty$ with positive probability: this will happen if and only if $\{Z(t); t \geq 0\}$ reaches state 0 before state Δ.

Our next result shows how to calculate the probability that $\{Z(t); t \geq 0\}$ is at state Δ at time t, when $Z(0)$ satisfies $\mathbb{P}(Z(0) = j) = \alpha_j$ for each $j \in E$.

Proposition 2.1 *For each $t \geq 0$, we have*

$$\mathbb{P}(Z(t) = \Delta) = \alpha_\Delta + \boldsymbol{\alpha}\mathbf{T}^{-1}(e^{\mathbf{T}t} - \mathbf{I})\mathbf{t}_\Delta \tag{11}$$

Proof This result can be established by thinking of a CTMC as a stochastic process governed by a collection of independent, homogeneous Poisson processes, à la Chapter 9 of Brémaud [3]. Indeed, letting $\{M_{i,j}\}_{i,j\in E: i\neq j}$ be a collection of independent, homogeneous Poisson processes that govern the transition times of $\{Z(t); t \geq 0\}$, where $M_{i,j}$ has rate $t(i,j)$, we may say that for each $t \geq 0$,

$$\mathbf{1}(Z(t) = \Delta) = \mathbf{1}(Z(0) = \Delta) + \sum_{k=1}^{N} \int_0^t \mathbf{1}(Z(s-) = k) M_{k,\Delta}(ds) \tag{12}$$

where $Z(s-) := \lim_{u\uparrow s} Z(u)$ for each $s > 0$. Taking expectations of both sides of (12), then applying the Campbell-Mecke formula and the time-dependent PASTA property to the right-hand-side (see [9]) yields

$$\begin{aligned}
\mathbb{P}(Z(t) = \Delta) &= \alpha_\Delta + \sum_{k=1}^{N} t(k,\Delta) \int_0^t \mathbb{P}(Z(s) = k) ds \\
&= \alpha_\Delta + \int_0^t \sum_{k=1}^{N} \mathbb{P}(Z(s) = k) t(k,\Delta) ds \\
&= \alpha_\Delta + \int_0^t \boldsymbol{\alpha} e^{\mathbf{T}s} \mathbf{t}_\Delta ds \\
&= \alpha_\Delta + \boldsymbol{\alpha}\mathbf{T}^{-1}(e^{\mathbf{T}t} - \mathbf{I})\mathbf{t}_\Delta
\end{aligned}$$

which establishes Proposition 2.1. Another way to prove this result involves applying Lévy's Formula: see Section 4.2 of Serfozo [17], Example 53 on page 271 of Serfozo [18], or Corollary 7.5.3 on page 232 of Last and Brandt [14]. $\diamondsuit$

We now turn our attention to terminating phase-type renewal processes. Consider a sequence of independent random variables (referred to throughout as interrenewals) $\{X_k\}_{k\geq 1}$, where X_1 is a random variable that is equal in distribution to τ_Δ when $Z(0)$ has initial distribution $(\beta_0, \boldsymbol{\beta}, \beta_\Delta)$, while $\{X_k\}_{k\geq 2}$ is an independent, identically distributed sequence of random variables that are equal in distribution to τ_Δ when $Z(0)$ has initial distribution $(\alpha_0, \boldsymbol{\alpha}, \alpha_\Delta)$. We assume throughout the rest of this section that $\alpha_0 = \beta_0 = \beta_\Delta = \alpha_\Delta = 0$, which ensures that each interrenewal of $\{A(t); t \geq 0\}$ is strictly positive with probability one. This assumption will hold each time we apply these results towards the study of finite-state birth-death processes.

From these interrenewals, we construct the terminating, delayed renewal process $\{A(t); t \geq 0\}$, where for each $t \geq 0$,

$$A(t) := \sum_{n=1}^{\infty} \mathbf{1}\left(\sum_{k=1}^{n} X_k \leq t\right).$$

Here $\{A(t); t \geq 0\}$ is a counting process which counts the number of times we reach state Δ—which we leave instantaneously, and transition to another state in $\{1, 2, \ldots, N\}$ with distribution $\boldsymbol{\alpha}$—before eventually visiting state 0.

Further associated with $\{A(t); t \geq 0\}$ is the CTMC $\{J(t); t \geq 0\}$, whose state space is $\{0, 1, 2, \ldots, N\}$ and whose generator $\mathbf{J}$ can be expressed in block-partitioned form as

$$\mathbf{J} := \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{t}_0 & \mathbf{D} \end{pmatrix} \tag{13}$$

where $\mathbf{D} := \mathbf{T} + \mathbf{t}_\Delta \boldsymbol{\alpha}$. This CTMC governs the behavior of $\{A(t); t \geq 0\}$, and we will use it to derive both the renewal function, as well as the renewal density of $\{A(t); t \geq 0\}$. The renewal function of the terminating phase-type renewal process is actually much easier to calculate than the renewal function associated with an ordinary phase-type renewal process whose interrenewals are all finite with probability one, because in this setting the matrix $\mathbf{D}$ is invertible.

Proposition 2.2 *For each $t \geq 0$, let $R(t) := \mathbb{E}[A(t)]$ represent the expected number of renewals that occur in $(0, t]$, and let $r(t) := R^{'}(t)$ denote the renewal density of $\{A(t); t \geq 0\}$. Then for each $t \geq 0$,*

$$r(t) = \boldsymbol{\beta} e^{\mathbf{D}t} \mathbf{t}_\Delta \tag{14}$$

and

$$R(t) = \boldsymbol{\beta} \mathbf{D}^{-1} (e^{\mathbf{D}t} - \mathbf{I}) \mathbf{t}_\Delta. \tag{15}$$

Proof In order to prove this result it suffices, in light of $\mathbf{D}$ being invertible, to calculate $r(t)$ for each $t > 0$. This is the case because (15) follows immediately from (14) by integrating r over $(0, t]$.

Our derivation of $r(t)$ is essentially the same as the derivation given in Chapter 3 of [15] of the renewal density of a nonterminating phase-type renewal process. Fix a real number $t > 0$, and first observe that for each real number $h > 0$,

$$R(t+h) - R(t) = \mathbb{E}[A(t+h) - A(t)] \quad = \quad \sum_{\ell=1}^{N} \mathbb{E}[A(t+h) - A(t) \mid J(t) = \ell] \mathbb{P}(J(t) = \ell).$$

Furthermore, for each $\ell \in \{1, 2, \ldots, N\}$,

$$\mathbb{E}[A(t+h) - A(t) \mid J(t) = \ell] = \sum_{m=0}^{N} (t(\ell, \Delta)\alpha_m h + o(h))$$

as $h \downarrow 0$, where we recall that a function $g : [0, \infty) \to \mathbb{R}$ is $o(h)$ as $h \downarrow 0$ if $\lim_{h\downarrow 0} g(h)/h = 0$. Hence,

$$\mathbb{E}[A(t+h) - A(t)] = \sum_{\ell=1}^{N} \sum_{m=0}^{N} (t(\ell, \Delta)\alpha_m h + o(h))\mathbb{P}(J(t) = \ell)$$

as $h \downarrow 0$. Dividing both sides by h and letting $h \downarrow 0$ then gives

$$\lim_{h\downarrow 0} \frac{R(t+h) - R(t)}{h} = \sum_{\ell=1}^{N} \sum_{m=0}^{N} t(\ell, \Delta)\alpha_m \mathbb{P}(J(t) = \ell) = \sum_{\ell=1}^{N} \mathbb{P}(J(t) = \ell)t(\ell, \Delta)$$

which proves that $r(t) := R^{'}(t)$ exists, and satisfies $r(t) = \boldsymbol{\beta} e^{\mathbf{D}t}\mathbf{t}_{\Delta}$. $\diamondsuit$

A faster, yet less elementary way to establish this result involves again making use of the framework found in [3], combined with the Campbell-Mecke formula and the time-dependent PASTA property: for each $t > 0$,

$$A(t) = \sum_{\ell=1}^{N} \int_0^t \mathbf{1}(J(s-) = \ell) M_{\ell,\Delta}(ds) \tag{16}$$

where $M_{\ell,\Delta}$ is a point process that counts the number of transitions made by $\{J(t); t \geq 0\}$ from ℓ to the absorbing state Δ, then instantaneously moving to another state in $\{1, 2, \ldots, N\}$ via the probability law $\boldsymbol{\alpha}$. Taking expectations of both sides of (16) and applying both the Campbell-Mecke formula and the time-dependent PASTA property to the right-hand-side further yields

$$R(t) = \sum_{\ell=1}^{N} \int_0^t \mathbb{P}(J(s) = \ell)t(\ell, \Delta)ds = \int_0^t \boldsymbol{\beta} e^{\mathbf{D}s}\mathbf{t}_{\Delta} ds$$

which also proves Proposition 2.2.

2.3 Relating the law of $Q(t)$ to the customer CTMC

The next step involves showing that we can express the probability mass function of $Q(t)$, conditional on $Q(0) = i$, in terms of random elements associated with the customer CTMC. In order to do this, we first need to introduce some additional notation.

Assume $Q(0) = i$ with probability one, where i is some fixed integer in the set $\{0, 1, 2, \ldots, N\}$. We associate to the customer found in slot $\ell \in \{1, 2, \ldots, i\}$ at time zero the customer CTMC $\{B_\ell(t); t \geq 0\}$, which governs how long that customer stays in the system. Further associated with that customer CTMC is another CTMC $\{Z_\ell^{\Delta}(t); t \geq 0\}$, as well as the renewal processes $\{N_{k,k+1,\ell}(t); t \geq 0\}$, for $1 \leq k \leq N-1$. Here $N_{k,k+1,\ell}(t)$ represents the number of times that particular customer moves from slot k to slot $k+1$ in the interval $(0,t]$ and $\{Z_\ell^\Delta(t); t \geq 0\}$ is a CTMC having state space $\{0, 1, 2, \ldots, N, \Delta\}$, initial distribution $(0, \mathbf{e}_\ell^T, 0)$, and generator $\mathbf{Z}_\Delta$, where

$$\mathbf{Z}_\Delta = \begin{pmatrix} 0 & \mathbf{0}^T & 0 \\ \mathbf{t}_0 & \mathbf{T} & \mathbf{t}_\Delta \\ 0 & \mathbf{0}^T & 0 \end{pmatrix} \tag{17}$$

where $\mathbf{T} = \mathbf{C}$, and $\mathbf{t}_\Delta = \lambda_{N-1}\mathbf{e}_N$: obviously $\mathbf{T}$ and $\mathbf{t}_\Delta$ together determine $\mathbf{t}_0$. Essentially, the CTMC $\{Z_\ell^\Delta(t); t \geq 0\}$ is the process you get when you take the original customer CTMC $\{B_\ell(t); t \geq 0\}$ and split the single absorbing state 0 into two absorbing states, 0 and Δ, where a transition from state N to state Δ occurs with rate λ_{N-1}: in $\{B_\ell(t); t \geq 0\}$, this type of transition would move you from state N to state 0, and this transition would physically correspond to the customer leaving the system due to overcrowding, not due to having its service completed.

Similarly, if a customer arrives to the system at time s, associate with that customer the customer CTMC $\{B^{(s)}(t); t \geq s\}$, and associate with this CTMC the CTMC $\{Z^{(s),\Delta}(t); t \geq s\}$ as well as, for each $k \in \{1, 2, \ldots, N-1\}$, the renewal processes $\{N_{k,k+1}^{(s)}(t); t \geq s\}$, where $N_{k,k+1}^{(s)}(t)$ represents the number of times that customer moves from slot k to slot $k+1$ in the interval $(s,t]$. Clearly $\{N_{k,k+1}^{(s)}(t); t \geq s\}$ behaves almost identically to the renewal process $\{N_{k,k+1,1}(t); t \geq 0\}$, so we will focus primarily on the latter processes when we derive various quantities associated with them.

Let $\{J_{k,\ell}(t); t \geq 0\}$ denote the underlying CTMC with state space $\{0, 1, 2, \ldots, N\}$ that governs the terminating, phase-type renewal process $\{N_{k,k+1,\ell}(t); t \geq 0\}$. Our next proposition finds the generator $\mathbf{J}_{k,\ell}$ of this underlying CTMC.

Proposition 2.3 *The generator* $\mathbf{J}_{k,\ell}$ *of* $\{J_{k,\ell}(t); t \geq 0\}$ *is as follows:*

$$\mathbf{J}_{k,\ell} = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{c}_0 & \mathbf{C} \end{pmatrix}.$$

Proof We first need to think about precisely what is being counted by the renewal process $\{N_{k,k+1,\ell}(t); t \geq 0\}$. This process keeps track of the number of times the customer occupying slot ℓ at time zero moves from slot k to slot $k+1$ before leaving the system, so we construct an extra state Δ, and we construct the modified customer CTMC $\{Z_\ell^{(k)}(t); t \geq 0\}$, whose state space is $\{0, 1, 2, \ldots, N-1, N, \Delta\}$ and whose generator $\mathbf{Z}_{k,\ell}$ is of the form

$$\mathbf{Z}_{k,\ell} := \begin{pmatrix} 0 & \mathbf{0}^T & \mathbf{0} \\ \mathbf{t}_{k,\ell,0} & \mathbf{T}_{k,\ell} & \mathbf{t}_{k,\ell,\Delta} \\ 0 & \mathbf{0}^T & 0 \end{pmatrix}$$

where $\mathbf{t}_{k,\ell,\Delta} := \lambda_k \mathbf{e}_k$, and the elements of $\mathbf{T}_{k,\ell} := [t_{k,\ell}(i,j)]_{i,j}$ satisfy

$$t_{k,\ell}(i,j) = c(i,j) - \lambda_k \mathbf{1}(i = k, j = k+1).$$

Next, notice that in our construction of $\{N_{k,k+1,\ell}(t); t \geq 0\}$, each time the underlying CTMC $\{J_{k,\ell}(t); t \geq 0\}$ reaches state Δ, it is instantaneously sent to state $k+1$ with probability one. Letting $\boldsymbol{\alpha}_{k,\ell} := \mathbf{e}_{k+1}^T$ denote the probability mass function that corresponds to sending the CTMC from state Δ to state $k+1$ with probability one, we see that since $\mathbf{J}_{k,\ell} = \mathbf{T}_{k,\ell} + \mathbf{t}_{k,\ell,\Delta}\boldsymbol{\alpha}_{k,\ell}$, we conclude that for each

$i, j \in \{1, 2, \ldots, N\}$,

$$\begin{aligned} j_{k,\ell}(i,j) &= t_{k,\ell}(i,j) + \lambda_k \mathbf{1}(i = k)\mathbf{1}(j = k+1) \\ &= c(i,j) - \lambda_k \mathbf{1}(i = k, j = k+1) + \lambda_i \mathbf{1}(i = k, j = k+1) \\ &= c(i,j) \end{aligned}$$

which proves $\mathbf{J}_{k,\ell} = \mathbf{C}$ for each $k \in \{1, 2, \ldots, N-1\}$, and each $\ell \in \{1, 2, \ldots, N\}$. $\diamondsuit$

Our next result shows how to express the probability mass function of $Q(t)$ in terms of random elements associated with the customer CTMC. Readers should note that throughout this paper, each usage of the symbol $\mathbb{P}_i$ refers to conditioning on a CTMC being in state i at time zero: it will be clear from the context which CTMC is being referred to whenever we use this symbol in our analysis.

Proposition 2.4 *For each real number $t \geq 0$, each integer $i \in \{0, 1, 2, \ldots, N\}$ and each integer $k \in \{1, 2, \ldots, N-1\}$, we have that when $Q(0) = i$ with probability one,*

$$\mathbb{P}_i(Q(t) \geq k) = \frac{1}{\lambda_k} \sum_{\ell=1}^{i} \frac{d}{dt} \mathbb{E}[N_{k,k+1,\ell}(t)] + \frac{\lambda_0}{\lambda_k} \mathbb{E}[N_{k,k+1,1}(t)] \tag{18}$$

Furthermore,

$$\mathbb{P}_i(Q(t) = N) = \frac{1}{\lambda_{N-1}} \sum_{\ell=1}^{i} \frac{d}{dt} \mathbb{P}(Z_\ell^\Delta(t) = \Delta) + \frac{\lambda_0}{\lambda_{N-1}} \mathbb{P}(Z_1^\Delta(t) = \Delta). \tag{19}$$

Proof Formula (19) is technically contained in Theorem 2.2 of [8], but the statement given in [8] (as well as its proof) contains a small error, so we fix it here (although we note that the idea behind the approach given in [8] is exactly the same as it is here). Formula (18) can be derived using similar methods that involve a more subtle application of the Campbell-Mecke formula.

We begin by deriving (19). For each $t > 0$, let $D(t)$ denote the number of customers that reach state Δ in the interval $(0, t]$, and if a customer arrives to the system at time s, let $E(s)$ denote the slot that arrival chooses for elimination. On the one hand, we see that for each $t \geq 0$,

$$D(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t, Q(T_n-) = N, E(T_n) = N) \tag{20}$$

where $\{T_n\}_{n \geq 1}$ denotes the collection of arrival times used to construct the counting process $\{N_0(t); t \geq 0\}$. The right-hand-side of (20) simply counts the number of arrivals that choose to eliminate a customer present at slot N upon arrival, when a customer is in fact present at slot N at the arrival time.

After taking expectations of both side of (20) and simplifying, we find that

$$\begin{aligned}
\mathbb{E}[D(t)] &= \sum_{n=1}^{\infty}\mathbb{P}(T_n \le t, Q(T_n-) = N, E(T_n) = N) \\
&= \sum_{n=1}^{\infty}\mathbb{P}(E(T_n) = N \mid T_n \le t, Q(T_n-) = N)\mathbb{P}(T_n \le t, Q(T_n-) = N) \\
&= \frac{\lambda_{N-1}}{\lambda_0}\sum_{n=1}^{\infty}\mathbb{P}(T_n \le t, Q(T_n-) = N) \\
&= \frac{\lambda_{N-1}}{\lambda_0}\mathbb{E}\left[\int_0^t \mathbf{1}(Q(s-) = N)N_0(ds)\right] \\
&= \lambda_{N-1}\int_0^t \mathbb{P}(Q(s) = N)ds
\end{aligned}$$

where the last line follows from an application of the Campbell-Mecke formula, along with the time-dependent PASTA property.

On the other hand, $D(t)$ can alternatively be expressed as

$$D(t) = \sum_{\ell=1}^{i}\mathbf{1}(Z_\ell(t) = \Delta) + \int_0^t \mathbf{1}(Z^{(s)}(t) = \Delta)N_0(ds) \tag{21}$$

and taking expectations of both sides of (21) while applying the Campbell-Mecke formula to the right-hand-side gives

$$\mathbb{E}[D(t)] = \sum_{\ell=1}^{i}\mathbb{P}_\ell(Z(t) = \Delta) + \lambda_0\int_0^t \mathbb{P}_1(Z(t-s) = \Delta)ds.$$

Hence, for each $t > 0$,

$$\lambda_{N-1}\int_0^t \mathbb{P}(Q(s) = N)ds = \sum_{\ell=1}^{i}\mathbb{P}_\ell(Z(t) = \Delta) + \lambda_0\int_0^t \mathbb{P}_1(Z(t-s) = \Delta)ds$$

and taking derivatives and solving for $\mathbb{P}(Q(t) = N)$ establishes (19).

It remains to derive (18). For each integer $k \in \{1, 2, \ldots, N-1\}$ and each real number $t \ge 0$, define $N_k(t)$ as the total number of times customers move from slot k to slot $k+1$ in the interval $(0, t]$. From the definition of $N_k(t)$, we can see that for each integer $k \in \{1, 2, \ldots, N-1\}$, and each $t \ge 0$,

$$N_k(t) = \sum_{\ell=1}^{i} N_{k,k+1,\ell}(t) + \int_{(0,t]} N^s_{k,k+1,1}(t)N_0(ds). \tag{22}$$

Starting with (22), taking expectations of both sides while applying the Campbell-Mecke formula on the right-hand-side gives

$$\begin{aligned}\mathbb{E}[N_k(t)] &= \sum_{\ell=1}^{i}\mathbb{E}[N_{k,k+1,\ell}(t)] + \lambda_0\int_0^t \mathbb{E}[N_{k,k+1,1}(t-s)]ds \\ &= \sum_{\ell=1}^{i}\mathbb{E}[N_{k,k+1,\ell}(t)] + \lambda_0\int_0^t \mathbb{E}[N_{k,k+1,1}(s)]ds. \qquad (23)\end{aligned}$$

Our next task is to relate $\mathbb{E}[N_k(t)]$ to the birth-death process $\{Q(t); t \geq 0\}$. Here

$$N_k(t) = \sum_{n=1}^{\infty}\mathbf{1}(T_n \leq t, Q(T_n-) \geq k, E(T_n) > k) \qquad (24)$$

because in order for a transition from k to $k+1$ to occur at time T_n, $Q(T_n-)$—the number of customers in the system immediately before time T_n—must be at least k, and none of the customers in slots $1, 2, \ldots, k$ can be eliminated at that time.

The next step is to take expectations of both sides of (24), then simplify the right-hand-side. Indeed,

$$\begin{aligned}\mathbb{E}[N_k(t)] &= \sum_{n=1}^{\infty}\mathbb{P}(T_n \leq t, Q(T_n-) \geq k, E(T_n) > k) \\ &= \sum_{n=1}^{\infty}\mathbb{P}(E(T_n) > k \mid T_n \leq t, Q(T_n-) \geq k)\mathbb{P}(T_n \leq t, Q(T_n-) \geq k) \\ &= \frac{\lambda_k}{\lambda_0}\sum_{n=1}^{\infty}\mathbb{P}(T_n \leq t, Q(T_n-) \geq k) \\ &= \frac{\lambda_k}{\lambda_0}\mathbb{E}\left[\int_{(0,t]}\mathbf{1}(Q(s-) \geq k)N_0(ds)\right] \\ &= \lambda_k\int_0^t \mathbb{P}(Q(s) \geq k)ds \qquad (25)\end{aligned}$$

where the third equality follows from $\mathbb{P}(E(T_n) > k \mid Q(T_n-) \geq k, T_n \leq t) = \lambda_k/\lambda_0$ since the distribution associated with how an arrival eliminates a customer is not affected by when that arrival occurs, and the fifth equality follows from another application of the Campbell-Mecke formula, combined with the time-dependent PASTA property.

Equating the right-hand-side of (25) with the right-hand-side of (23) shows that

$$\lambda_k\int_0^t \mathbb{P}(Q(s) \geq k)ds = \sum_{\ell=1}^{i}\mathbb{E}[N_{k,k+1,\ell}(t)] + \lambda_0\int_0^t \mathbb{E}[N_{k,k+1,1}(t-s)]ds \qquad (26)$$

and taking derivatives of (26) and simplifying establishes (18). ♢

Our next result expresses the probability mass function of $Q(t)$ in terms of the matrix $\mathbf{C}$.

Proposition 2.5 *For each real number $t \geq 0$, we have for each integer $k \in \{1, 2, \ldots, N\}$ that*

$$\mathbb{P}_i(Q(t) \geq k) = \lambda_0 \mathbf{e}_1^T(-\mathbf{C})^{-1}\mathbf{e}_k + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1} e^{\mathbf{C}t}\mathbf{e}_k + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\mathbf{e}_k \tag{27}$$

where $\mathbf{C}$ is the subintensity matrix found in the blocked partitioned representation of the generator of the customer CTMC.

Proof We first establish (27) for the case where $k = N$. From (19), we see that for each real $t \geq 0$,

$$\mathbb{P}_i(Q(t) = N) = \frac{1}{\lambda_{N-1}} \sum_{\ell=1}^{i} \frac{d}{dt}\mathbb{P}(Z_\ell^\Delta(t) = \Delta) + \frac{\lambda_0}{\lambda_{N-1}}\mathbb{P}(Z_1^\Delta(t) = \Delta). \tag{28}$$

Since, by Proposition 2.1,

$$\mathbb{P}(Z_\ell^\Delta(t) = \Delta) = \lambda_{N-1}\mathbf{e}_\ell^T \mathbf{C}^{-1}(e^{\mathbf{C}t} - \mathbf{I})\mathbf{e}_N$$

and thus

$$\frac{d}{dt}\mathbb{P}(Z_\ell^\Delta(t) = \Delta) = \lambda_{N-1}\mathbf{e}_\ell^T e^{\mathbf{C}t}\mathbf{e}_N$$

it quickly follows that

$$\mathbb{P}_i(Q(t) = N) \quad = \quad \sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\mathbf{e}_N + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1}(e^{\mathbf{C}t} - \mathbf{I})\mathbf{e}_N$$

which establishes (27) for the case where $k = N$.

It remains to derive (27) for the case where $k \in \{1, 2, \ldots, N-1\}$. From (18) within Proposition 2.4, we learned that

$$\mathbb{P}_i(Q(t) \geq k) = \frac{1}{\lambda_k} \sum_{\ell=1}^{i} \frac{d}{dt}\mathbb{E}[N_{k,k+1,\ell}(t)] + \frac{\lambda_0}{\lambda_k}\mathbb{E}[N_{k,k+1,1}(t)] \tag{29}$$

and the right-hand-side of (29) can be simplified further: combining Propositions 2.2 and 2.3 gives

$$\mathbb{E}[N_{k,k+1,\ell}(t)] = \lambda_k \mathbf{e}_\ell \mathbf{C}^{-1}(e^{\mathbf{C}t} - \mathbf{I})\mathbf{e}_\ell$$

and

$$\frac{d}{dt}\mathbb{E}[N_{k,k+1,\ell}(t)] = \lambda_k \mathbf{e}_\ell e^{\mathbf{C}t}\mathbf{e}_\ell$$

and after plugging these expressions into the right-hand-side of (29) and simplifying, we get (27), which proves Proposition 2.5. $\diamondsuit$

Corollary 2.1 *The stationary distribution* $\mathbf{p}$ *of* $\{Q(t); t \geq 0\}$ *is as follows:*

$$p_0 = 1 - \lambda_0 \mathbf{e}_1^T (-\mathbf{C})^{-1} \mathbf{e}_1$$

and for $j \in \{1, 2, \ldots, N\}$*,*

$$p_j = \lambda_0 \mathbf{e}_1^T (-\mathbf{C})^{-1} (\mathbf{e}_j - \mathbf{e}_{j+1}).$$

We can also use Proposition 2.5 to write down expressions for the transition functions of $\{Q(t); t \geq 0\}$.

Corollary 2.2 *For each* $i, j \in S$*, and each real number* $t \geq 0$*, we have*

$$p_{i,j}(t) = p_j + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1} e^{\mathbf{C}t} (\mathbf{e}_j - \mathbf{e}_{j+1}) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t} (\mathbf{e}_j - \mathbf{e}_{j+1}). \tag{30}$$

Proof This is a simple consequence of Proposition 2.5 and Corollary 2.1. $\diamondsuit$

3 The General Case

We now show that the transition functions of $\{Q(t); t \geq 0\}$ still satisfy (30), even when the knockout queue criterion is *not* satisfied. This fact is not difficult to check, as we only need to show that these functions are a solution to the Kolmogorov Forward Equations associated with $\{Q(t); t \geq 0\}$, but of course we did make use of what we learned in Section 2 to guess that (30) is the correct form of $p_{i,j}(t)$.

The next Lemma will be used to help verify this claim.

Lemma 3.1 *For each* $i \in \{1, 2, \ldots, N\}$*, we have*

$$\mathbf{C}\mathbf{e}_i = \lambda_{i-1}(\mathbf{e}_{i-1} - \mathbf{e}_i) - \mu_i(\mathbf{e}_i - \mathbf{e}_{i+1}). \tag{31}$$

Proof This immediately follows from the structure of each column of $\mathbf{C}$. $\diamondsuit$

The next lemma shows that the elements of $(-\mathbf{C})^{-1}$ can be expressed rather elegantly in terms of the stationary distribution $\mathbf{p}$ of $\{Q(t); t \geq 0\}$. For each $j \in \{0, 1, 2, \ldots, N\}$, we define

$$F_j := \sum_{k=0}^{j} p_k.$$

Lemma 3.2 *The values of* $\mathbf{e}_i^T (-\mathbf{C})^{-1} (\mathbf{e}_j - \mathbf{e}_{j-1})$ *are as follows:*
(1) for $1 \leq i \leq N$*, and* $1 \leq j \leq i$*,*

$$\mathbf{e}_i^T (-\mathbf{C})^{-1} (\mathbf{e}_j - \mathbf{e}_{j-1}) = \frac{p_{j-1}}{\lambda_{i-1} p_{i-1}} (1 - F_{i-1}). \tag{32}$$

and (2) for $1 \leq i \leq N$*, and* $i + 1 \leq j \leq N + 1$*,*

$$\mathbf{e}_i^T (-\mathbf{C})^{-1} (\mathbf{e}_j - \mathbf{e}_{j-1}) = \frac{-p_{j-1}}{\lambda_{i-1} p_{i-1}} F_{i-1}. \tag{33}$$

Moreover, for $1 \leq i \leq N$, and $1 \leq j \leq N$,

$$\mathbf{e}_i^T(-\mathbf{C})^{-1}\mathbf{e}_j = \frac{1}{\lambda_{i-1}p_{i-1}} F_{\min(i-1,j-1)}(1 - F_{\max(i-1,j-1)}). \tag{34}$$

Proof For each $i \in \{1, 2, \dots, N\}$, we define, for each $j \in \{0, 1, \dots, N\}$,

$$x_{i,j} := \mathbf{e}_i^T(-\mathbf{C})^{-1}(\mathbf{e}_{j+1} - \mathbf{e}_j).$$

Fix $i \in \{1, 2, \dots, N\}$: multiplying both sides of (31) by $\mathbf{C}^{-1}$, then by $\mathbf{e}_i^T$ yields, for each $j \in \{1, \dots, N\}$,

$$\mathbf{1}_{\{i=j\}} = \lambda_{j-1}x_{i,j-1} - \mu_j x_{i,j}. \tag{35}$$

Using the linear system given by (35), we can easily express each $x_{i,j}$ in terms of $x_{i,i-1}$. First, notice that when $j = i$ in (35), we get

$$1 = \lambda_{i-1}x_{i,i-1} - \mu_i x_{i,i} \tag{36}$$

which gives

$$x_{i,i} = \frac{\lambda_{i-1}}{\mu_i}x_{i,i-1} - \frac{1}{\mu_i} = \frac{\lambda_{i-1}}{\mu_i}\left[x_{i,i-1} - \frac{1}{\lambda_{i-1}}\right].$$

Furthermore, for each $j \in \{i+1, i+2, \dots, N\}$, (35) reduces to

$$0 = \lambda_{j-1}x_{i,j-1} - \mu_j x_{i,j}$$

or, equivalently,

$$x_{i,j} = \frac{\lambda_{j-1}}{\mu_j}x_{i,j-1}$$

for each $j \in \{i+1, i+2, \dots, N\}$. Further iterations of the same idea show that for $j \in \{i, i+1, \dots, N\}$,

$$x_{i,j} = \left[\prod_{\ell=i}^{j} \frac{\lambda_{\ell-1}}{\mu_\ell}\right]\left[x_{i,i-1} - \frac{1}{\lambda_{i-1}}\right].$$

Next, suppose $1 \leq j < i$: in this case, (35) becomes

$$0 = \lambda_{j-1}x_{i,j-1} - \mu_j x_{i,j}$$

or

$$x_{i,j-1} = \frac{\mu_j}{\lambda_{j-1}}x_{i,j}.$$

We can further iterate this observation to show that, for $0 \leq j < i$,

$$x_{i,j} = \left[\prod_{\ell=j+1}^{i-1} \frac{\mu_\ell}{\lambda_{\ell-1}}\right] x_{i,i-1}.$$

The remaining $x_{i,i-1}$ term can be found by noticing that

$$\sum_{j=0}^{N} x_{i,j} = 0$$

From this observation, we find that

$$\begin{aligned} 0 &= \sum_{j=0}^{i-1}\left[\prod_{\ell=j+1}^{i-1}\frac{\mu_\ell}{\lambda_{\ell-1}}\right]x_{i,i-1} + \sum_{j=i}^{N}\left[\prod_{\ell=i}^{j}\frac{\lambda_{\ell-1}}{\mu_\ell}\right]\left[x_{i,i-1} - \frac{1}{\lambda_{i-1}}\right] \\ &= \left[\sum_{j=0}^{i-1}\left[\prod_{\ell=j+1}^{i-1}\frac{\mu_\ell}{\lambda_{\ell-1}}\right] + \sum_{j=i}^{N}\left[\prod_{\ell=i}^{j}\frac{\lambda_{\ell-1}}{\mu_\ell}\right]\right]x_{i,i-1} - \frac{1}{\lambda_{i-1}}\sum_{j=i}^{N}\left[\prod_{\ell=i}^{j}\frac{\lambda_{\ell-1}}{\mu_\ell}\right] \end{aligned}$$

which gives

$$x_{i,i-1} = \frac{1}{\lambda_{i-1}}\frac{\sum_{j=i}^{N}\left[\prod_{\ell=i}^{j}\frac{\lambda_{\ell-1}}{\mu_\ell}\right]}{\left[\sum_{j=0}^{i-1}\left[\prod_{\ell=j+1}^{i-1}\frac{\mu_\ell}{\lambda_{\ell-1}}\right] + \sum_{j=i}^{N}\left[\prod_{\ell=i}^{j}\frac{\lambda_{\ell-1}}{\mu_\ell}\right]\right]} = \frac{1}{\lambda_{i-1}}(1 - F_{i-1}).$$

Thus, for $0 \le j \le i-1$,

$$x_{i,j} = \left[\prod_{\ell=j+1}^{i-1}\frac{\mu_\ell}{\lambda_{\ell-1}}\right]x_{i,i-1} = \frac{p_{j-1}}{\lambda_{i-1}p_{i-1}}(1 - F_{i-1})$$

and similarly, for $i \le j \le N$,

$$x_{i,j} = \left[\prod_{\ell=i}^{j}\frac{\lambda_{\ell-1}}{\mu_\ell}\right]\left[x_{i,i-1} - \frac{1}{\lambda_{i-1}}\right] = \frac{p_{j-1}}{p_{i-1}}\left[x_{i,i-1} - \frac{1}{\lambda_{i-1}}\right] = \frac{-p_{j-1}}{\lambda_{i-1}p_{i-1}}F_{i-1}$$

since

$$x_{i,i-1} - \frac{1}{\lambda_{i-1}} = \frac{(1 - F_{i-1})}{\lambda_{i-1}} - \frac{1}{\lambda_{i-1}} = \frac{-F_{i-1}}{\lambda_{i-1}}.$$

This establishes both (32) and (33). Finally, (34) can quickly be derived from (32) and (33). ◇

Proposition 3.1 *The equilibrium distribution* $\mathbf{p}_e$ *of* $\mathbf{p}$ *can be stated in terms of* $(-\mathbf{C})^{-1}$*, namely,*

$$\mathbf{p}_e = \frac{\lambda_0}{\mathbb{E}[Q(\infty)]}\mathbf{e}_1^T(-\mathbf{C})^{-1}.$$

Thus,

$$p_0 = 1 - \lambda_0\mathbf{e}_1^T(-\mathbf{C})^{-1}\mathbf{e}_1$$

and for each $j \in \{1, 2, \dots, N\}$*,*

$$p_j = \lambda_0\mathbf{e}_1^T(-\mathbf{C})^{-1}(\mathbf{e}_j - \mathbf{e}_{j+1})$$

Proof Using Lemma 3.2, we find that for each $j \in \{1, 2, \dots, N\}$,

$$\lambda_0 \mathbf{e}_1^T (-\mathbf{C})^{-1} \mathbf{e}_j = \lambda_0 \frac{1}{\lambda_0 p_0} F_{\min(0,j-1)} (1 - F_{\max(0,j-1)}) = 1 - F_{j-1}.$$

Thus,

$$p_0 = 1 - (1 - F_0) = 1 - \lambda_0 \mathbf{e}_1^T (-\mathbf{C})^{-1} \mathbf{e}_1$$

and for each $j \in \{1, 2, \dots, N\}$,

$$p_j = (1 - F_{j-1}) - (1 - F_j) = \lambda_0 \mathbf{e}_1^T (-\mathbf{C})^{-1} (\mathbf{e}_j - \mathbf{e}_{j+1}).$$

This completes the proof. $\diamondsuit$

We are now ready to state and prove the following theorem.

Theorem 3.1 *The transition functions of the birth-death process $\{Q(t); t \geq 0\}$ are as follows: for each $i, j \in \{0, 1, \dots, N\}$, we have*

$$p_{i,j}(t) = p_j + \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t} \mathbf{C}^{-1} (\mathbf{e}_j - \mathbf{e}_{j+1}) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t} (\mathbf{e}_j - \mathbf{e}_{j+1})$$

for each real $t \geq 0$.

Readers should note that once Theorem 3.1 has been proven, Theorem 1.1 simply follows from combining Theorem 3.1 with Proposition 3.1.

Proof For each $i, j \in \{0, 1, \dots, N\}$, we define the function $g_{i,j} : [0, \infty) \to \mathbb{R}$ as

$$g_{i,j}(t) := p_j + \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t} \mathbf{C}^{-1} (\mathbf{e}_j - \mathbf{e}_{j+1}) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t} (\mathbf{e}_j - \mathbf{e}_{j+1}).$$

It suffices, then, to show that these functions satisfy the Kolmogorov Forward Equations associated with $\{Q(t); t \geq 0\}$.

The first step is to show that the functions $g_{i,j}$, $i, j \in S$, satisfy $g_{i,j}(0) = \mathbf{1}(i = j)$. Observe first that for each $i \geq 0$,

$$\begin{aligned}
g_{i,0}(0) &= p_0 + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1} (\mathbf{e}_0 - \mathbf{e}_1) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T (\mathbf{e}_0 - \mathbf{e}_1) \\
&= 1 - \lambda_0 \mathbf{e}_1^T (-\mathbf{C})^{-1} \mathbf{e}_1 + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1} (\mathbf{e}_0 - \mathbf{e}_1) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T (\mathbf{e}_0 - \mathbf{e}_1) \\
&= 1 - \mathbf{1}(i \geq 1) \\
&= \mathbf{1}(i = 0).
\end{aligned}$$

Next, for each $i \geq 0$, and each $j \geq 1$,

$$\begin{aligned} g_{i,j}(0) &= p_j + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1}(\mathbf{e}_j - \mathbf{e}_{j+1}) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T(\mathbf{e}_j - \mathbf{e}_{j+1}) \\ &= \lambda_0 \mathbf{e}_1^T(-\mathbf{C})^{-1}(\mathbf{e}_j - \mathbf{e}_{j+1}) + \lambda_0 \mathbf{e}_1^T \mathbf{C}^{-1}(\mathbf{e}_j - \mathbf{e}_{j+1}) + \sum_{\ell=1}^{i} \mathbf{e}_\ell^T(\mathbf{e}_j - \mathbf{e}_{j+1}) \\ &= \sum_{\ell=1}^{i} \mathbf{e}_\ell^T(\mathbf{e}_j - \mathbf{e}_{j+1}) \\ &= \mathbf{1}(i = j) \end{aligned}$$

so the $g_{i,j}$ functions satisfy the correct initial conditions.

The next step of the argument involves showing that for each $i \in \{0, 1, \ldots, N\}$,

$$g_{i,0}^{'}(t) = \sum_{k=0}^{N} g_{i,k}(t) q(k, 0).$$

Taking derivatives of $g_{i,0}(t)$ yields

$$g_{i,0}^{'}(t) = \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t}(\mathbf{e}_0 - \mathbf{e}_1) + \left[\sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\right] \mathbf{C}(\mathbf{e}_0 - \mathbf{e}_1).$$

Next, observe that

$$\begin{aligned} \sum_{k=0}^{N} g_{i,k}(t) q(k, 0) &= g_{i,0}(t)(-\lambda_0) + g_{i,1}(t)\mu_1 \\ &= \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t}\left[(-\lambda_0)\mathbf{C}^{-1}(\mathbf{e}_0 - \mathbf{e}_1) + \mu_1 \mathbf{C}^{-1}(\mathbf{e}_1 - \mathbf{e}_2)\right] \\ &\quad + \left[\sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\right]\left[(-\lambda_0)(\mathbf{e}_0 - \mathbf{e}_1) + \mu_1(\mathbf{e}_1 - \mathbf{e}_2)\right]. \end{aligned}$$

It is clear, then, that in order to show $g_{i,0}^{'}(t) = \sum_{k=0}^{N} g_{i,k} q(k, 0)$, it suffices to prove

$$\mathbf{C}(\mathbf{e}_0 - \mathbf{e}_1) = (-\lambda_0)(\mathbf{e}_0 - \mathbf{e}_1) + \mu_1(\mathbf{e}_1 - \mathbf{e}_2)$$

but this is immediate from Lemma 3.1.

The next step is to show that, for each $i \in \{0, 1, \ldots, N\}$,

$$g_{i,N}^{'}(t) = \sum_{k=0}^{N} g_{i,k}(t) q(k, N).$$

Again, for each $t > 0$,

$$g_{i,N}^{'}(t) = \lambda_0 \mathbf{e}_1 e^{\mathbf{C}t} \mathbf{e}_N + \left[\sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\right] \mathbf{C}\mathbf{e}_N$$

and furthermore, one can show after some algebra that

$$\begin{aligned}\sum_{k=0}^{N} g_{i,k}(t)q(k,N) &= \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t}\left[\lambda_{N-1}\mathbf{C}^{-1}(\mathbf{e}_{N-1}-\mathbf{e}_N) - \mu_N \mathbf{C}^{-1}\mathbf{e}_N\right] \\ &\quad + \left[\sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\right]\left[\lambda_{N-1}(\mathbf{e}_{N-1}-\mathbf{e}_N) - \mu_N \mathbf{e}_N\right].\end{aligned}$$

Clearly, then, in order to show $g_{i,N}^{'}(t) = \sum_{k=0}^{N} g_{i,k}(t)q(k,N)$, it suffices to instead show that

$$\mathbf{C}\mathbf{e}_N = \lambda_{N-1}(\mathbf{e}_{N-1}-\mathbf{e}_N) - \mu_N \mathbf{e}_N$$

but this equality again follows from Lemma 3.1.

It remains to establish the result for each $j \in \{1, 2, \ldots, N-1\}$, and each $i \in \{0, 1, \ldots, N\}$. For each $t > 0$,

$$g_{i,j}^{'}(t) = \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t}(\mathbf{e}_j - \mathbf{e}_{j+1}) + \left[\sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\right]\mathbf{C}(\mathbf{e}_j - \mathbf{e}_{j+1})$$

and

$$\begin{aligned}\sum_{k=0}^{n} g_{i,k}(t)q(k,j) &= \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t}\mathbf{C}^{-1}\left[\lambda_{j-1}(\mathbf{e}_{j-1}-\mathbf{e}_j) - (\lambda_j+\mu_j)(\mathbf{e}_j-\mathbf{e}_{j+1}) + \mu_{j+1}(\mathbf{e}_{j+1}-\mathbf{e}_{j+2})\right] \\ &\quad + \left[\sum_{\ell=1}^{i} \mathbf{e}_\ell^T e^{\mathbf{C}t}\right]\left[\lambda_{j-1}(\mathbf{e}_{j-1}-\mathbf{e}_j) - (\lambda_j+\mu_j)(\mathbf{e}_j-\mathbf{e}_{j+1}) + \mu_{j+1}(\mathbf{e}_{j+1}-\mathbf{e}_{j+2})\right]\end{aligned}$$

so in order to prove the claim, it suffices to show

$$\mathbf{C}(\mathbf{e}_j - \mathbf{e}_{j+1}) = \lambda_{j-1}(\mathbf{e}_{j-1}-\mathbf{e}_j) - (\lambda_j+\mu_j)(\mathbf{e}_j-\mathbf{e}_{j+1}) + \mu_{j+1}(\mathbf{e}_{j+1}-\mathbf{e}_{j+2})$$

but again, this can be shown using Lemma 3.1. $\diamondsuit$

4 Applications

We close by illustrating how Theorem 1.1 can be used to derive many interesting properties of birth-death processes. Our first result shows that when $Q(0) = 0$ with probability 1, $Q(t)$ is stochastically increasing in t. This theorem is of course known, see e.g. Theorem 6.1 of Van Doorn [5] but our proof seems to be new.

Theorem 4.1 *Suppose $Q(0) = 0$ with probability one. Then for each $s, t \geq 0$ satisfying $s < t$,*

$$\mathbb{P}(Q(s) \geq k) \leq \mathbb{P}(Q(t) \geq k) \tag{37}$$

for each integer $k \geq 0$.

Proof First observe that when $Q(0) = 0$, we may use Theorem 1.1 to say that for each integer $k \geq 0$, and each $t \geq 0$,

$$\mathbb{P}(Q(t) \geq k) = \mathbb{P}(Q(\infty) \geq k) - \mathbb{E}[Q(\infty)]\mathbf{p}_e^T e^{\mathbf{C}t}\mathbf{e}_k. \tag{38}$$

Next, notice that since

$$\mathbb{E}[Q(\infty)]\mathbf{p}_e = \lambda_0 \mathbf{e}_1^T(-\mathbf{C})^{-1} \tag{39}$$

we can take derivatives on both sides of (38), while simultaneously making use of (39) to find that

$$\frac{d}{dt}\mathbb{P}(Q(t) \geq k) = \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}t}\mathbf{e}_k \tag{40}$$

and the right-hand-side of (40) is nonnegative, due to $e^{\mathbf{C}t}$ being a positive matrix (see Assertion 1 of Granovsky and Zeifman [10]). This establishes (37). $\diamondsuit$

Remark We can also establish that $e^{\mathbf{C}t}$ is a positive matrix by making use of a uniformization-like construction. Define

$$\lambda := \max_{1 \leq i \leq N} |c(i,i)|$$

and define $\mathbf{P} \in \mathbb{R}^{N \times N}$ as

$$\mathbf{P} = \mathbf{I} + \frac{1}{\lambda}\mathbf{C}.$$

While it is not necessarily true that $\mathbf{P}$ is a transition probability matrix, it is true that $\mathbf{P}$ contains only nonnegative elements. Furthermore, for each $t \geq 0$,

$$e^{\mathbf{C}t} = e^{\lambda \mathbf{P}t - \lambda \mathbf{I}t} = e^{\lambda \mathbf{P}t}e^{-\lambda t}$$

and clearly the matrix on the right-hand-side is a positive matrix.

Obviously, we can use Theorem 1.1 to derive analogous expressions for all moments of $Q(t)$ as well. Here is such a representation for the first moment of $Q(t)$.

Corollary 4.1 *Suppose $Q(0)$ has as its initial distribution the row vector $\boldsymbol{\alpha}$. Then*

$$\mathbb{E}[Q(t)] = \mathbb{E}[Q(\infty)] - \mathbb{E}[Q(\infty)]\mathbf{p}_e^T e^{\mathbf{C}t}\mathbf{e} + \mathbb{E}[Q(0)]\boldsymbol{\alpha}_e^T e^{\mathbf{C}t}\mathbf{e}.$$

Corollary 4.1 is interesting, particularly when $Q(0) = 0$ with probability one: in this case, we get

$$\mathbb{E}[Q(t)] = \mathbb{E}[Q(\infty)](1 - \mathbf{p}_e^T e^{\mathbf{C}t}\mathbf{e})$$

or, in other words, $\mathbb{E}[Q(t)]$ can be expressed as $\mathbb{E}[Q(\infty)]$ times the cumulative distribution function F_X of a nonnegative random variable X: more particularly,

$$F_X(t) := \mathbb{P}(X \leq t) = 1 - \mathbf{p}_e^T e^{\mathbf{C}t}\mathbf{e}, \qquad t \geq 0. \tag{41}$$

Observe too that when the row sums of $\mathbf{C}$ are all nonpositive, X can be interpreted as a phase-type random variable, which now represents the amount of time it takes the customer CTMC $\{B(t); t \geq 0\}$ to reach state 0, given its initial distribution is $(0, \mathbf{p}_e^T)$. More generally, one could also say that X is a matrix-exponential random variable, whose pdf f_X is of the form

$$f_X(x) = \lambda_0 \mathbf{e}_1^T e^{\mathbf{C}x}\mathbf{e}, \qquad x \geq 0.$$

Theorem 4.2 *Suppose that for each $i \in \{1, 2, \ldots, N\}$, the rates of $\{Q(t); t \geq 0\}$ satisfy*

$$(\lambda_{i-1} - \lambda_i) + (\mu_i - \mu_{i-1}) \geq 0.$$

Then $\mathbb{E}_0[Q(t)]$ is a concave function of t.

Proof This result simply follows from the fact that when $\mathbf{c}_0 \geq 0$, $\mathbf{e}_1^T e^{\mathbf{C}t}\mathbf{e} = \mathbb{P}_1(\tau_0 > t)$, where τ_0 is the amount of time it takes the customer CTMC $\{B(t); t \geq 0\}$ whose generator is defined in (7) to reach state zero. This probabilistic interpretation of $\mathbf{e}_1^T e^{\mathbf{C}t}\mathbf{e}$ appears to be lost when at least one element of $\mathbf{c}_0$ is strictly negative, yet readers should note that $\mathbf{e}_1^T e^{\mathbf{C}t}\mathbf{e}$ can still be interpreted probabilistically when the knockout queue conditions are no longer satisfied (we need $\mathbf{c}_0 \geq 0$ in order to understand $\mathbf{C}$ probabilistically). $\diamondsuit$

To the best of our knowledge, Theorem 4.2 was first established in Chapter 5 of [16] for the case where the birth and death rates are non-increasing and non-decreasing, respectively, with respect to the state variable (i.e. the knockout queue criterion, although the proof found in [16] does not make use of the knockout queue construction), while in [8], the result was derived through applying the transient Little's law to the knockout queue construction. Here we see that there is a slightly more general sufficient condition that preserves concavity of $\mathbb{E}_0[Q(t)]$. Readers interested in such structural results should also note that similar results have been found for various types of processes associated with Lévy processes: see Kella [12], Kella and Sverchkov [13], and Andersen and Mandjes [1].

References

[1] Andersen, L.N. and Mandjes, M. (2009). Structural properties of reflected Lévy processes. *Queueing Systems* **63**, 301-322.

[2] Bladt, M. and Nielsen, B.F. (2017). *Matrix-Exponential Distributions in Applied Probability*. Springer, New York.

[3] Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Springer, New York.

[4] Coolen-Schrijner, P., and Van Doorn, E.A. (2002). The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences* **16**, 351-366.

[5] Van Doorn, E.A. (1980). Stochastic monotonicity of birth-death processes. *Advances in Applied Probability* **12**, 59-80.

[6] Van Doorn, E.A., Zeifman, A.I., and Panfilova, T.L. (2010). Bounds and asymptotics for the rate of convergence of birth-death processes. *Theory of Probability and its Applications* **54**, 97-113.

[7] Fralix, B., and Riaño, G. (2010). A new look at transient versions of Little's law, and M/G/1 preemptive last-come-first-served queues. *Journal of Applied Probability* **47**, 459-473.

[8] Fralix, B. (2013). A time-dependent study of the knockout queue. *Probability in the Engineering and Informational Sciences* **27**, 309-317.

[9] Fralix, B., Van Leeuwaarden, J.S.H., and Boxma, O.J. (2013). Factorization identities for a general class of reflected processes. *Journal of Applied Probability* **50**, 632-653.

[10] Granovsky, B.L., and Zeifman, A.I. (2000). The N-limit of spectral gap of a class of birth-death Markov chains. *Applied Stochastic Models in Business and Industry* **16**, 235-248.

[11] Horn, R., and Johnson, C. (2013). *Matrix Analysis*, Second Edition. Cambridge University Press, New York.

[12] Kella, O. (1992). Concavity and reflected Lévy processes. *Journal of Applied Probability* **29**, 209-215.

[13] Kella, O., and Sverchkov, M. (1994). On concavity of the mean function and stochastic ordering for reflected processes with stationary increments. *Journal of Applied Probability* **31**, 1140-1142.

[14] Last, G., and Brandt, M. (1997). *Marked Point Processes on the Real Line: The Dynamic Approach*. Springer, New York.

[15] Latouche, G., and Ramaswami, V. (1999). *Introduction to Matrix-Analytic Methods in Stochastic Modeling*. ASA-SIAM publications, Philadelphia, PA, USA.

[16] Lindvall, T. (1992). *Lectures on the Coupling Method*. Dover Publications, New York.

[17] Serfozo, R.F. (1999). *Introduction to Stochastic Networks*. Springer, New York.

[18] Serfozo, R.F. (2009). *Basics of Applied Stochastic Processes*. Springer-Verlag, Berlin.

[19] Zeifman, A. (1991). Some estimates of the rate of convergence for birth and death processes. *Journal of Applied Probability* **28**, 268-277.