Qeios

Counting Processes with Multiple Randomness: Examples in Queuing Theory

Guang-Liang Li
University of Hong Kong
glli@eee.hku.hk

July 13, 2023

Abstract

This article introduces "counting processes with multiple randomness", which appear naturally in applications and differ essentially from known stochastic processes in the literature. Unlike times between consecutive "events" of a usual counting process, inter-event times of a counting process with multiple randomness are defined on proper subsets of the sample space, and not eligible to have marginal distributions. With examples in queuing theory, the existence of this new type of counting processes is demonstrated, and their properties are illustrated.

MSC2020: Primary 60G20; Secondary 90B15, 90B22, 60K25.

1 Introduction

A counting process $N(t)$ is a stochastic process, representing the total number of "events" occurred in the time interval $[0,t]$. If the jth event occurs at τ_j, where $j \geq 1$, the time $\tau_{j+1} - \tau_j$ between the jth and the $(j+1)$th events is a random variable. If $\tau_{j+1} - \tau_j$ for every j is defined on the whole sample space, $N(t)$ is a usual counting process. Poisson processes, and more generally, renewal processes, are popular examples of usual counting processes. Unlike the usual counting processes, there are also counting processes such that $\tau_{j+1} - \tau_j$ for any given j can only be defined on a proper subset of the sample space, and has no marginal distribution. Such counting processes differ essentially from any known stochastic process in the existing literature.

Denote by $(\Omega, \mathscr{A}, \mathbb{P})$ the probability space of a random experiment, where $\mathbb{P}$ is the probability measure on the measurable space $(\Omega, \mathscr{A})$, and $\mathscr{A}$ the σ-algebra of subsets of Ω. The sample space Ω consists of sequences $(E_j)_{j\geq 1}$ of "events". Let $m > 1$ be a fixed integer. For a given $\omega \in \Omega$, denote by $\{M_i(\omega) : 1 \leq i \leq m\}$ a partition of the set of positive integers $\mathbb{N}$, where

$$M_i(\omega) = \{j \in \mathbb{N} : (\tau_{j+1} - \tau_j)(\omega) = T_i(\omega)\}$$

https://doi.org/10.32388/OUKGO3

and T_i is a random variable with distribution F_i. If $i \neq k$ then $F_i \neq F_k$. Similarly, for a given $j \in \mathbb{N}$, denote by $\{\Omega_i(j) : 1 \leq i \leq m\}$ a partition of Ω, where

$$\Omega_i(j) = \{\omega \in \Omega : (\tau_{j+1} - \tau_j)(\omega) = T_i(\omega)\}$$

and $\mathbb{P}[\Omega_i(j)] > 0$ for all $1 \leq i \leq m$.

Definition 1.1. An integer-valued stochastic process $N(t) \geq 1$ on Ω, which has the above properties, is a counting process with multiple (or m-fold) randomness.

For $m = 1$, $N(t)$ degenerates into a usual counting process. Evidently, when $m > 1$, $\tau_{j+1} - \tau_j$ for any j is not defined on the whole sample space and hence not eligible to have a marginal distribution. The distribution of $\tau_{j+1} - \tau_j$ is not determined by any joint distribution of random variables on Ω; it is determined uniquely by the random experiment in question. The random variables in the sequence $(\tau_{j+1} - \tau_j)_{j \geq 1}$ may or may not be statistically independent. Furthermore, immediately from Definition 1.1, $(\tau_{j+1} - \tau_j)_{j \geq 1}$ is not a stationary sequence.

Counting processes with multiple randomness appear naturally in applications. Unfortunately, they are all mistaken for usual counting processes, as stochastic processes characterized by Definition 1.1 have never been identified. We shall use examples in queuing theory to demonstrate the existence of these new counting processes, and illustrate their properties. By identifying this new type of stochastic processes in queuing models, some long-standing inconsistencies, such as those concerning output processes of stable queues in steady state and product-form equilibrium distributions of queuing networks, can be readily resolved.

In section 2, we shall see that departures from a stable queue in statistic equilibrium constitute a counting process with two-fold randomness (i.e., $m = 2$). In section 3 and section 4, the output process of an M/M/1 queue and Jackson networks of queues are revisited, respectively. The article concludes in Section 5.

2 Departures from GI/GI/1 Queue

Consider the departure process from a stable, general single-server queue (i.e., a GI/GI/1 system) in statistical equilibrium. A GI/GI/1 queue has an infinite waiting room and a work-conserving server with a finite service capacity defined by the maximum rate at which the server can perform work. The meaning of "work-conserving" is that the server will not stand idle when there is unfinished work in the system. Customers arrive at this system according to a renewal process, and are served one at a time. Times between successive arrivals are independent and identically distributed (i.i.d.) random variables with a finite mathematical expectation, and so are service times of customers. Inter-arrival times and service times are also mutually independent. For such a

queue, a customer leaves the system if and only if the customer has been served. The mean inter-arrival time is greater than the mean service time. Hence the queue is stable, and assumed to be in steady state. For the purpose of this study, it is not necessary to assume a specific queuing discipline.

According to the literature, times between consecutive departures from a stable queue in statistical equilibrium follow a marginal distribution [1]. However, the existence of such a marginal distribution is only an unjustified assumption taken for granted without verification. As we shall see in this section, for systems modeled by a GI/GI/1 queue, inter-departure times between customers from the queue do not have a marginal distribution, even if the queue is stable and in steady state.

Let c_j represent the jth customer served, and denote by E_j the event "departure of the jth customer from the system." Because the queue is already in steady state, we shall focus on $(E_j)_{j\geq 1}$. Let τ_j be the departure time of c_j, and Q_j the queue size, i.e., the total number of customers in the queue (including the customer in service) immediately after the jth departure. Denote by X_j and Y_j the times between the jth and the $(j+1)$th departures when $Q_j = 0$ and $Q_j > 0$, respectively.

$$X_j = I_j + S_{j+1} \tag{2.1}$$

$$Y_j = S_{j+1} \tag{2.2}$$

where I_j is the idle time spent by the server waiting for the arrival of c_{j+1}, and S_{j+1} the service time of c_{j+1}.

By definition, a random variable is a measurable real-valued function; its domain can be the whole sample space or a subset of the sample space. To define a random variable U on the whole sample space Ω, a value must be assigned to U at *each* $\omega \in \Omega$. Similarly, to define a random variable on a subset of Ω, a value must be assigned to this random variable at each sample point in the subset. For a random variable on the whole sample space, such as U, its (marginal) distribution is defined by

$$P_U(B) = \mathbb{P}(U \in B)$$

where B is an arbitrary Borel set of the real line. When it is necessary to emphasize the connection between U and $\mathbb{P}$ on $(\Omega, \mathscr{A})$, the right-hand side of the above equation can be used to express the distribution of U directly.

Similarly, components of a random vector (or terms of a stochastic sequence) are random variables on Ω, such that all the components (or terms) take their values at the *same* $\omega \in \Omega$. If a random vector has been defined, then the joint distribution of its components is fixed, and the marginal distribution of each component is determined by the joint distribution. Based on a given random vector, some new random variables may be constructed on subsets of Ω.

For example, let (U, V) be a random vector. The joint distribution is $P_{U,V}$, which determines P_U and P_V, the marginal distributions of U and V. Based on this random vector, a random variable W may be constructed on a subset Θ of a positive probability, such that $W(\omega) = V(\omega)$ for $\omega \in \Theta$, where $\Theta \subset \Omega$

is specified by some values taken by U. The random variable W follows a conditional distribution P_W determined by $P_{U,V}$.

So far, two types of random variables have been mentioned; they are either defined on the whole sample space such as U, or follow a conditional distribution such as W, which is defined on some subset of the sample space. There also exist random variables different from those mentioned above. Random variables of this type are not components of a random vector, and their distributions are not determined by a joint distribution. As we shall see, times between successive departures from the GI/GI/1 queue are random variables of this type. They are defined on some proper subsets of Ω. Their distributions and properties are determined by a *chronological order of events experienced by customers*. This chronological order is not determined by properties of the queuing model; it is determined by physical systems modeled by the queue. For a work-conserving system, events experienced by every customer occur naturally as follows:

- First, a customer arrives at the queue.
- Upon arrival,
  - the customer receives service immediately if the server is idle;
  - otherwise the customer has to wait in line.
- Finally, after being served, the customer departs.

Consequently, for an arbitrary j, if the server becomes idle immediately after c_j leaves, the time between the departures of c_j and c_{j+1} is the sum of an idle time and a service time, as shown by Eq.(2.1); otherwise the inter-departure time is a service time, see Eq.(2.2). Because of the above chronological order, the sample space Ω has an interesting structure. Write

$$\Phi_j = \{\omega \in \Omega : Q_j(\omega) = 0\}$$

$$\Psi_j = \{\omega \in \Omega : Q_j(\omega) > 0\}$$

$$N(\omega) = \{i \in \mathbb{N} : Q_i(\omega) = 0\}$$

and

$$N'(\omega) = \{i \in \mathbb{N} : Q_i(\omega) > 0\}.$$

For each $j \geq 1$, $\mathbb{P}(\Phi_j) = 1 - \rho > 0$, $\mathbb{P}(\Psi_j) = \rho > 0$, and Φ_j and Ψ_j constitute a partition of Ω. Similarly, for each $\omega \in \Omega$, $N(\omega)$ and $N'(\omega)$ form a partition of $\mathbb{N}$. Clearly, $\Phi_j \cap \Psi_i \neq \emptyset$ if $i \neq j$ and $N(\omega) \cap N'(\omega') \neq \emptyset$ if ω differs from ω'.

In the literature, stability and some properties of the arrival process of the GI/GI/1 queue are established by using the strong law of large numbers. Results of this kind are true for every $\omega \in \Omega$, with the exception at most of a set $\mathcal{N}$ of sample points such that $\mathcal{N} \in \mathscr{A}$ and $\mathbb{P}(\mathcal{N}) = 0$. In other words, such results hold with probability one or almost surely. Write $\Omega_+ = \Omega \setminus \mathcal{N}$. Then we may also say that a statement is true for every $\omega \in \Omega_+$, if it is true almost surely. Clearly, for each $j \in \mathbb{N}$, Φ_j and Ψ_j are also a partition of Ω_+, and for each

$\omega \in \Omega_+$, $N(\omega)$ and $N'(\omega)$ are a partition of $\mathbb{N}$. By exploring the structure of Ω, we can capture some important properties of times between consecutive departures from the GI/GI/1 queue.

Lemma 2.1. *If a GI/GI/1 queue is stable and already in statistical equilibrium, and if the server has a finite service capacity, then for any given $j \in \mathbb{N}$,*

(a) $\tau_{j+1} - \tau_j$ cannot be expressed by any single, fixed random variable, i.e., no random variable on Ω can describe $\tau_{j+1} - \tau_j$, and hence $\tau_{j+1} - \tau_j$ has no marginal (i.e., unconditional) distribution;

(b) $(Q_j, \tau_{j+1} - \tau_j)$ is not a random vector, and hence $\tau_{j+1} - \tau_j$ has no distributions conditional on values taken by Q_j.

Corollary 2.2. *The terms of $(\tau_{j+1} - \tau_j)_{j \geq 1}$ are not random variables on Ω.*

To understand Lemma 2.1 and its corollary, it may be helpful for us to consider the following question: If $\tau_{j+1} - \tau_j$ could be expressed by a random variable on Ω, say Z_j, what would be the value of Z_j at ω corresponding to $Q_j(\omega)$? At any given ω, the value of Q_j equals either zero or a positive integer. Whatever value Q_j takes at ω, it is impossible to assign any value to Z_j at ω corresponding to $Q_j(\omega)$. In contrast to $\tau_{j+1} - \tau_j$, the service time S_j, the number of customers Q_j, and times between consecutive arrivals are all random variables on Ω. There is some subtlety here, however. For example, when playing the role of an inter-departure time Y_j defined on Ψ_j, S_{j+1} is no longer a random variable on Ω.

Proof. For any j, Φ_j and Ψ_j constitute a partition of Ω. Hence for any $\omega \in \Omega$, either $\omega \in \Phi_j$ and

$$\left.\begin{array}{l} Q_j(\omega) = 0 \\ \tau_{j+1}(\omega) - \tau_j(\omega) = X_j(\omega) \end{array}\right\} \tag{2.3}$$

or $\omega \in \Psi_j$ and

$$\left.\begin{array}{l} Q_j(\omega) > 0 \\ \tau_{j+1}(\omega) - \tau_j(\omega) = Y_j(\omega). \end{array}\right\} \tag{2.4}$$

As shown by Eq.(2.1) and Eq.(2.2), X_j and Y_j are well-defined random variables; their distributions, P_{X_j} and P_{Y_j}, are determined by the chronological order of events experienced by customers, as implied by Eq.(2.3) and Eq.(2.4).

$$P_{X_j}(B) = \mathbb{P}(X_j \in B | \Phi_j), \quad P_{Y_j}(B) = \mathbb{P}(Y_j \in B | \Psi_j)$$

where B is an arbitrary Borel set of the real line.

To prove (a), it is sufficient for us to show that, for an arbitrarily given j, $\tau_{j+1} - \tau_j$ can only be described by X_j or Y_j, i.e., any single, fixed random variable on Ω cannot express $\tau_{j+1} - \tau_j$. We first prove that the above statement is true for every $\omega \in \Omega_+$. Because Φ_j and Ψ_j are a partition of Ω_+, either Eq.(2.3) or Eq.(2.4) must hold exclusively for any $\omega \in \Omega_+$.

Suppose to the contrary that there is a random variable Z_j on Ω_+, such that $Z_j = \tau_{j+1} - \tau_j$. Consequently,

$$\{Q_j = 0, X_j \in B\} = \{Q_j = 0, Z_j \in B\} \tag{2.5}$$

$$\{Q_j > 0, Y_j \in B\} = \{Q_j > 0, Z_j \in B\}. \tag{2.6}$$

Because $\{Q_j = 0\}$ is independent of events concerning future departures after c_j leaves, such as $\{X_j \in B\}$ and $\{Z_j \in B\}$, Eq.(2.5) implies

$$\mathbb{P}(Q_j = 0)P_{X_j}(B) = \mathbb{P}(Q_j = 0)P_{Z_j}(B)$$

where P_{Z_j} is the distribution of Z_j. Similarly, $\{Q_j > 0\}$ is independent of $\{Y_j \in B\}$ and $\{Z_j \in B\}$, so Eq.(2.6) implies

$$\mathbb{P}(Q_j > 0)P_{Y_j}(B) = \mathbb{P}(Q_j > 0)P_{Z_j}(B).$$

Because $\mathbb{P}(Q_j = 0) = \mathbb{P}(\Phi_j) = 1 - \rho > 0$ and $\mathbb{P}(Q_j > 0) = \mathbb{P}(\Psi_j) = \rho > 0$ for all j, treating $\tau_{j+1} - \tau_j$ as Z_j on Ω_+ leads to

$$P_{X_j}(B) = P_{Y_j}(B) = P_{Z_j}(B).$$

This is absurd. The absurdity shows that Z_j does not exist on Ω_+, and cannot be defined on the whole sample space Ω. Actually $\{Z_j \in B\}$ is not an event in $\mathscr{A}$. Therefore, $\tau_{j+1} - \tau_j$ cannot be described by any single, fixed random variable on Ω, and is not eligible to have a marginal distribution. This proves (a).

By definition, each component of a random vector (or each term of a stochastic sequence) is a random variable on Ω, and all the components (or terms) take values at the same $\omega \in \Omega$. By (a) proved above, $\tau_{j+1} - \tau_j$ and Q_j cannot form a random vector, although Q_j is a random variable on Ω. Consequently, $\tau_{j+1} - \tau_j$ is not eligible to have distributions conditional on values taken by Q_j. Similarly, $\tau_2 - \tau_1, \tau_3 - \tau_2, \cdots$ cannot form a sequence of random variables on Ω. This completes the proof of the lemma and its corollary. □

The conditions required by Lemma 2.1 exclude two conceivable scenarios for $\tau_{j+1} - \tau_j$ to be a single, fixed random variable. One is the worst scenario, and the other is an ideal scenario. The worst scenario is a queue with $\mathbb{P}(\Phi_j) = 0$ for all j, i.e., the queue is unstable, because $\mathbb{P}(\Phi_j) = 0$ for all j implies $N(\omega) = \emptyset$ for all $\omega \in \Omega_+$, which means $\tau_{j+1} - \tau_j = S_{j+1}$ on Ω_+ for all j. That is, the server is always busy almost surely, as the idle time I_j vanishes identically on Ω_+, and hence the queue will never be empty.

The ideal scenario is a queue with $\mathbb{P}(\Psi_j) = 0$ for all j, i.e., the queue is always empty almost surely, because $\mathbb{P}(\Psi_j) = 0$ for all j implies $N'(\omega) = \emptyset$ for all $\omega \in \Omega_+$, which means $\tau_{j+1} - \tau_j = I_j$ on Ω_+ for all j. That is, the server has an infinite service capacity, as the service time S_j vanishes identically, and hence I_j becomes an inter-arrival time with probability one.

The scenarios above illustrate two extreme situations. For a stable queue in statistical equilibrium with $\mathbb{P}(\Phi_j) > 0$ and $\mathbb{P}(\Psi_j) > 0$ for any j, the departures

from the queue constitute a counting process with two-fold randomness, such that

$$\Omega_1(j) = \Phi_j,\ \Omega_2(j) = \Psi_j,\ M_1(\omega) = N(\omega),\ M_2(\omega) = N'(\omega)$$

and

$$T_1 = X_j,\ T_2 = Y_j.$$

Accordingly, $(\tau_{j+1} - \tau_j)_{j\geq 1}$ is a sequence with two-fold randomness in the following sense, which makes $(\tau_{j+1} - \tau_j)_{j\geq 1}$ essentially different from any sequence of random variables on Ω.

(i) At each $\omega \in \Omega_+$, the sequence $(\tau_{j+1} - \tau_j)_{j\geq 1}$ consists of two subsequences, such that the division of $(\tau_{j+1} - \tau_j)_{j\geq 1}$ into the subsequences is random.

$$[\tau_{j+1}(\omega) - \tau_j(\omega)]_{j\geq 1} = [\tau_{j+1}(\omega) - \tau_j(\omega)]_{j\in N(\omega)} \cup [\tau_{j+1}(\omega) - \tau_j(\omega)]_{j\in N'(\omega)}.$$

(ii) For a fixed $\omega \in \Omega_+$, if $j \in N(\omega)$, then $\tau_{j+1} - \tau_j$ equals $X_j(\omega)$; otherwise $\tau_{j+1} - \tau_j$ takes $Y_j(\omega)$ as its value.

Having a finite service capacity, the server is either busy or idle with a positive probability, and hence realistic inter-departure times always fall into two categories. In statistical equilibrium, the probability for the server to be busy or idle is fixed, and the inter-departure times are given by either X_j or Y_j according to values taken by Q_j, see Eq.(2.1) and Eq.(2.2). In the literature, the distribution of X_j or Y_j is interpreted as the distribution of $\tau_{j+1} - \tau_j$ conditional on values taken by Q_j. According to such interpretation, a marginal distribution of the inter-departure times could be constructed based on the distributions of X_j and Y_j.

However, the above interpretation is incorrect. By Lemma 2.1, for each $j \in \mathbb{N}$, $\tau_{j+1} - \tau_j$ cannot be expressed as a random variable on Ω, and hence is not eligible to have a marginal distribution; actually $\tau_{j+1} - \tau_j$ is a term of a sequence with two-fold randomness; it is wrong to interpret the distribution of X_j or Y_j as the distribution of $\tau_{j+1} - \tau_j$ conditional on values taken by Q_j, for $(Q_j, \tau_{j+1} - \tau_j)$ is not a random vector.

3 Departures from M/M/1 Queue and Burke's Theorem

Consider a stable M/M/1 queue in steady state, which is the simplest instance of the GI/GI/1 queue. Customers arrive at this system according to a Poison process; service times are mutually independent and follow a common exponential distribution. According to Burke's theorem [2], times between successive departures from this queue in steady state are mutually independent, following a marginal distribution identical to the distribution of inter-arrival times. Let λ be the parameter of this distribution. Clearly, Burke's theorem contradicts Lemma 2.1 and Corollary 2.2. In the following, we shall see that

Burke's original proof given in [2] and Reich's proof based on time reversibility given in [3] are both fundamentally flawed, and simulation results claimed to be in agreement with Burke's theorem are misinterpreted. In other words, Burke's theorem is wrong.

3.1 Flaw in Burke's Proof

In the proof given in [2], "the length of an arbitrary inter-departure interval" is considered. In the notation of this present article, such an interval is expressed as (τ_j, τ_{j+1}) for an arbitrary j. Denote by $Q(t)$ the state of the queue (i.e., the queue size) at time t. Burke's proof begins with calculating the probability of an event $\{Q(t) = k, \tau_{j+1} - \tau_j > t\}$, where t is an instant after τ_j, and τ_j is taken to be the instant of the last previous departure. Note that comparing $\tau_{j+1} - \tau_j$ with t amounts to choosing τ_j as the origin on the time axis.

Burke's proof implies an assumption: $\tau_{j+1} - \tau_j$ is a single, fixed random variable and can take values together with $Q(t)$ at *every* $\omega \in \Omega$. Under this assumption, $[Q(t), \tau_{j+1} - \tau_j]$ is treated as a random vector, and $\tau_{j+1} - \tau_j$ is treated as a random variable Z_j on Ω. In Burke's proof, a set of differential equations governing the probability of $\{Q(t) = k, Z_j > t\}$ is established based on the above assumption. Solving the equations yields [2]

$$\mathbb{P}[Q(t) = k, Z_j > t] = \mathbb{P}[(Q(t) = k]\mathbb{P}(Z_j > t),\ k = 0, 1, \cdots$$

where $\mathbb{P}[Q(t) = k]$ is the equilibrium probability of $\{Q(t) = k\}$, which actually does not depend on t, and

$$\mathbb{P}(Z_j > t) = e^{-\lambda t}.$$

According to the calculation above, $\{Q(t) = k\}$ and $\{Z_j > t\}$ are mutually independent for any $k \geq 0$, and inter-departure times are distributed as inter-arrival times. However, the assumption underlying Burke's proof is false, as it leads to contradictions.

To see this, consider first $Q(t) = 0$, i.e., the server is idle at time t. Because τ_j is the instant of the last previous departure, and taken to be the origin on the time axis, the arrival instant of c_{j+1} must be in the interval (t, τ_{j+1}), and no departure occurs in the open interval $(0, \tau_{j+1})$. Moreover, $\{Q(t) = 0, Z_j > t\}$ implies $\{Q(0) = 0, Z_j > t\}$, i.e.,

$$\{Q(t) = 0, Z_j > t\} \subset \{Q(0) = 0, Z_j > t\}. \tag{3.1}$$

Otherwise there would be at least one departure in the interval $(0, \tau_{j+1})$, which is impossible. On both sides of Expression (3.1), Z_j is the same random variable on Ω. According to Burke's proof, Z_j follows the exponential distribution with parameter λ. However, as required by the chronological order of events experienced by customers, whenever $Q(0) = 0$,

$$Z_j = I_j + S_{j+1}$$

where S_{j+1} is a service time, and the idle time I_j is distributed as an inter-arrival time. Clearly, $I_j + S_{j+1}$ does not follow the distribution of inter-arrival times. We see a contradiction.

Now consider $Q(t) = k > 0$, i.e., the server is busy at time t. As we can readily see, either

$$\{Q(t) = k, Z_j > t\} \subset \{Q(0) = 0, Z_j > t\}$$

or

$$\{Q(t) = k, Z_j > t\} \subset \{Q(0) = k', Z_j > t\}$$

where $0 < k' \leq k$. The former expression leads to $Z_j = I_j + S_{j+1}$, and we see the same contradiction as shown above. In the latter expression, $Z_j = S_{j+1}$, which is a service time. According to the assumption implied by Burke's proof, in the two expressions above, Z_j is the same random variable on Ω; we see a contradiction again. The contradictions are due to the assumption that $Q(t)$ and $\tau_{j+1} - \tau_j$, where $\tau_{j+1} - \tau_j$ is treated as a single, fixed random variable, can take values together at each $\omega \in \Omega$. Because of this false assumption, $\tau_{j+1} - \tau_j$ is mistaken for a random variable Z_j on Ω with a marginal distribution resulting from

$$\sum_{k=0}^{\infty} \mathbb{P}[Q(t) = k, Z_j > t] = e^{-\lambda t}.$$

However, such a single, fixed random variable Z_j is not defined on Ω, and fails to describe realistic inter-departure times.

As shown in Section 2 (Lemma 2.1), for any given j, there are two kinds of realistic inter-departure times, X_j and Y_j, see Eq.(2.1) and Eq.(2.2), which are already defined on Φ_j and Ψ_j, respectively, and cannot be expressed by any random variable on Ω. Such inter-departure times are terms of a sequence with two-fold randomness, and do not have a marginal distribution; their distributions cannot be determined by joint distributions of random vectors defined on Ω. The calculation leading to Burke's theorem is invalid; it ignores completely the dependence of departures on the state of the server. Such dependence is part of the constraints imposed by physical systems to be studied based on the queuing model.

3.2 Flaw in Reich's Proof

As a birth-death process in steady state, $Q(t)$ is a time-reversible Markov process. Denote by $Q^*(t)$ the time-reversed process of $Q(t)$, such that points of time are ordered in the reversed direction. The reversed process $Q^*(t)$ is also a birth-death process; this results in a different proof of Burke's theorem. The proof is given in [3] and goes like this: In steady state, $Q(t)$ and $Q^*(t)$ have the same probabilistic structure; the points of time at which $Q(t)$ increases by 1 form a Poisson process at rate λ, and hence the time points at which $Q^*(t)$ increases by 1 also form a Poisson process at rate λ; customers depart at the latter points of time, so departures from the M/M/1 queue constitute a Poisson process with rate λ. See also [4].

The proof above is questionable. Although the time points at which $Q(t)$ increases by 1 constitute a Poisson process, the time points τ_j at which $Q^*(t)$ increases by 1 cannot form a Poisson process. The length $|\tau_j - \tau_{j+1}|$ between two successive points at which $Q^*(t)$ increases by 1 is exactly the same length as $\tau_{j+1} - \tau_j$ between two successive points at which $Q(t)$ decreases by 1. As we can readily see below, for a birth-death process in whatever direction of time, times between consecutive deaths and times between consecutive births do not necessarily follow the same distribution.

In an open interval between two consecutive deaths, the state of a birth-death process may change m times, where $m \geq 0$. For $Q(t)$, which represents the number of customers in the M/M/1 queue at time t, "death" and "birth" refer to "departure" and "arrival", respectively. Consequently, intervals between consecutive deaths are inter-departure intervals, and a change of $Q(t)$ in (τ_j, τ_{j+1}) is due to an arrival in (τ_j, τ_{j+1}). It is sufficient to consider the following two cases.

Case 1: $m = 0$. That is, $Q(t)$ remains unchanged in (τ_j, τ_{j+1}), which implies $Q(t) > 0$ for $\tau_j \leq t < \tau_{j+1}$.

$$Q(t) = \begin{cases} k+2 & t < \tau_j \\ k+1 & \tau_j \leq t < \tau_{j+1} \\ k & t = \tau_{j+1}. \end{cases}$$

Because $Q(t)$ decreases by 1 if and only if one customer has been served, the inter-departure time $\tau_{j+1} - \tau_j$ is a service time.

Case 2: $m \geq 1$. That is, $Q(t)$ changes at least once in (τ_j, τ_{j+1}). In this case, either $Q(t) > 0$ for $\tau_j \leq t < \tau_{j+1}$, or $Q(t) = 0$ for $\tau_j \leq t < t_1$ where $t_1 < \tau_{j+1}$, and $Q(t) > 0$ for $t_1 \leq t < \tau_{j+1}$.

$$Q(t) = \begin{cases} k+1 & t < \tau_j \\ k & \tau_j \leq t < t_1 \\ k+1 & t_1 \leq t < t_2 \\ k+2 & t_2 \leq t < t_3 \\ \cdots & \cdots \\ k+m & t_m \leq t < \tau_{j+1} \\ k+m-1 & t = \tau_{j+1}. \end{cases}$$

If $k > 0$, $\tau_{j+1} - \tau_j$ is still a service time. However, if $k = 0$, $\tau_{j+1} - \tau_j$ consists of an idle time of the server (i.e., $t_1 - \tau_j$) and a service time (i.e., $\tau_{j+1} - t_1$). In both cases above, $\tau_{j+1} - \tau_j$ does not follow the distribution of inter-arrival times.

Similarly, for $Q^*(t)$, where points of time are ordered in the reversed direction, if $m = 0$,

$$Q^*(t) = \begin{cases} k & t = \tau_{j+1} \\ k+1 & \tau_{j+1} < t \leq \tau_j \\ k+2 & t > \tau_j \end{cases}$$

and $|\tau_j - \tau_{j+1}| = S_{j+1}$. When $m \geq 1$,

$$Q^*(t) = \begin{cases} k+m-1 & t = \tau_{j+1} \\ k+m & \tau_{j+1} < t \leq t_m \\ \cdots & \cdots \\ k+2 & t_3 < t \leq t_2 \\ k+1 & t_2 < t \leq t_1 \\ k & t_1 < t \leq \tau_j \\ k+1 & t > \tau_j \end{cases}$$

and $|\tau_j - \tau_{j+1}| = S_{j+1}$ if $k > 0$; otherwise $|\tau_j - \tau_{j+1}| = I_j + S_{j+1}$. As shown above, in steady state, times between consecutive departures from the M/M/1 queue always follow two different distributions, neither of which is identical to the distribution of inter-arrival times.

It is well known that, for a given t, $Q(t)$ is independent of future arrivals after t. Time reversibility is often used to argue that $Q(t)$ is also independent of past departures before t. Based on an analogy between $Q(t)$ and $Q^*(t)$, the argument goes as follows: $Q(t)$ is independent of future arrivals after t; departures are "arrivals" when looking backwards in time; because $Q(t)$ and $Q^*(t)$ are statistically identical, $Q(t)$ is independent of past departures before t.

However, the analogy between $Q(t)$ and $Q^*(t)$ is not appropriate, and the argument based on the analogy is wrong. Although $Q(t)$ is independent of arrivals after t, both arrivals and departures prior to t determine $Q(t)$: Increase in $Q(t)$ is due to arrivals before t, and decrease in $Q(t)$ is due to departures before t. For a stable M/M/1 queue in steady state, $Q(t)$ depends on departures before t necessarily.

Any queuing model for solving problems in the real world must satisfy the constraints imposed by physical systems to be studied based on the model. The reversed process $Q^*(t)$ is merely a pure mathematical entity constructed without considering the chronological order of events experienced by customers, and hence irrelevant to the departure process from the M/M/1 queue. By Corollary 2.2, $(\tau_{j+1} - \tau_j)_{j \geq 1}$ is even not a sequence of random variables on Ω. As shown in Section 2, $(\tau_{j+1} - \tau_j)_{j \geq 1}$ is a sequence with two-fold randomness; its terms do not have a marginal distribution. Time reversibility cannot change $(\tau_{j+1} - \tau_j)_{j \geq 1}$ into a sequence of i.i.d. random variables on Ω.

3.3 Misinterpreted Simulation Results

Some simulation results are claimed to be in agreement with Burke's theorem. However, as shown below, such results are misinterpreted. Let $\mu < \infty$ represent the parameter of the service-time distribution. The mean inter-arrival time and the mean service time are $1/\lambda$ and $1/\mu$, respectively. In steady state,

$$\mathbb{P}(Q_j = 0) = 1 - \rho$$

and

$$\mathbb{P}(Q_j > 0) = \rho$$

for all j, where $\rho = \lambda/\mu$. Clearly, $0 < \rho < 1$. When $Q_j = 0$, $\tau_{j+1} - \tau_j$ is given by X_j, see Eq.(2.1), and its probability density function (pdf) for the M/M/1 queue is

$$f_X(t) = \frac{\lambda\mu}{\mu - \lambda}(e^{-\lambda t} - e^{-\mu t}).$$

If $Q_j > 0$, $\tau_{j+1} - \tau_j$ is given by Y_j, see Eq.(2.2), and its pdf for the M/M/1 queue is $f_Y(t) = \mu e^{-\mu t}$.

If we let the service rate μ approach infinity, then $\rho \to 0$, $\mathbb{P}(Q_j = 0) \to 1$, $\mathbb{P}(Q_j > 0) \to 0$, and for any given t,

$$\lim_{\mu\to\infty} f_X(t) = \lambda e^{-\lambda t}, \quad \lim_{\mu\to\infty} f_Y(t) = 0.$$

As $\mu \to \infty$, S_j approaches zero, and $\tau_{j+1} - \tau_j$ tends to the exponentially distributed inter-arrival time for all j. This corresponds to the ideal scenario discussed in Section 2. Only in this idealized situation, the departures from the M/M/1 queue constitute a Poisson process at rate λ.

Let K_t represent a time interval $(t, t + dt]$ of an infinitesimal length dt. For a stable M/M/1 queue with $\mu < \infty$ in steady state, a simple calculation yields

$$\mathbb{P}(Q_j = 0, X_j \in K_t) + \mathbb{P}(Q_j > 0, Y_j \in K_t) = \lambda e^{-\lambda t} dt. \tag{3.2}$$

The above equation is indeed in agreement with the simulation, but it cannot be interpreted as

$$\mathbb{P}(Z_j \in K_t) = \mathbb{P}(Q_j = 0, Z_j \in K_t) + \mathbb{P}(Q_j > 0, Z_j \in K_t) \tag{3.3}$$

where Z_j is a random variable representing $\tau_{j+1} - \tau_j$ at every $\omega \in \Omega$.

By Lemma 2.1, $\tau_{j+1} - \tau_j$ cannot be expressed by any random variable on Ω, and $(Q_j, \tau_{j+1} - \tau_j)$ is not a random vector. It is wrong to consider Eq.(3.2) identical to Eq.(3.3), for $\{Z_j \in K_t\} \notin \mathscr{A}$. That is, Z_j is not a random variable on Ω (see Section 2). In fact, Eq.(3.2) is merely the probability of an event given below.

$$H_{j,t} = \{Q_j = 0, X_j \in K_t\} \cup \{Q_j > 0, Y_j \in K_t\}.$$

Simulation studies are based on the strong law of large numbers. In simulations, the strong law is usually applied to a sequence of i.i.d. random variables on Ω. However, it is not legitimate to apply the strong law to $(\tau_{j+1} - \tau_j)_{j\geq 1}$, which is not a sequence of random variables on Ω; its terms do not have a marginal distribution (see Corollary 2.2). Actually $(\tau_{j+1} - \tau_j)_{j\geq 1}$ is a sequence with two-fold randomness. Nevertheless, if we consider $\mathcal{I}(H_{j,t})$, the indicator of $H_{j,t}$, it is not difficult to see that $\mathcal{I}(H_{1,t}), \mathcal{I}(H_{2,t}), \cdots$ constitute a sequence of i.i.d. random variables on Ω for any given t. By the strong law,

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} \mathcal{I}(H_{j,t})}{n} = \mathbb{E}[\mathcal{I}(H_{1,t})]$$

with probability one, where

$$\mathbb{E}[\mathcal{I}(H_{1,t})] = \mathbb{P}(Q_1 = 0, X_1 \in K_t) + \mathbb{P}(Q_1 > 0, Y_1 \in K_t) = \lambda e^{-\lambda t} dt.$$

The above analysis not only explains why Eq.(3.2) is consistent with the simulation, but it also explains why Eq.(3.2) differs from Eq.(3.3); the latter equation does not hold.

4 Queuing Networks and Jackson's Theorem

Consider a Jackson network in which there are J queues denoted by $Q_m, m = 1, 2, \cdots, J$. At Q_m there are s_m servers, and the mean service time is $1/\mu_m$. In Jackson's theorem [5], these queues are considered stable if $\lambda_m < s_m\mu_m$, where λ_m is the total arrival rate of customers at Q_m, determined by the following equations.

$$\lambda_m = \gamma_m + \sum_{n=1}^{J} \lambda_n P_{nm},\ m = 1, 2, \cdots, J. \tag{4.1}$$

In Eq.(4.1), γ_m is the arrival rate of customers at Q_m from outside of the system, and P_{nm} is the probability for a customer to join Q_m immediately after leaving Q_n. So the arrival rate of customers at Q_m from Q_n is $\lambda_n P_{nm}$. All the proofs of Jackson's theorem, including Jackson's original proof and the proof with time reversibility, rely on Eq.(4.1).

According to Jackson's interpretation (e.g, [5, 6]), after the network has been in operation for an infinitely long time, it behaves *as if* the numbers in the queues were independent random variables with their joint distribution exhibiting a product form. However, this interpretation is inconsistent with the notion of statistical independence. As we shall see below, the condition in Jackson's theorem, i.e., $\lambda_m < s_m\mu_m, m = 1, 2, \cdots, J$, does not imply stability of *every* queue in the network, because Eq.(4.1) relies on an unjustified assumption (see below), which holds only in an unrealistic scenario. Consequently, the assumption makes Jackson's theorem irrelevant to physical systems modeled by Jackson networks of queues, although the solution of Eq.(4.1) may not be difficult to find.

Consider two queues Q_1 and Q_2 in tandem, where Q_1 is a stable M/M/1 queue in steady state. Customers arrive first at Q_1 according to a Poisson process with a rate λ. After being served at Q_1, they join Q_2 immediately; Q_2 also has an infinite waiting room. Service times of a customer spent at Q_1 and Q_2 are mutually independent, following exponential distributions with finite service rates $\mu_1 > \lambda$ and $\mu_2 > \lambda$. In addition, service times at Q_2 are not only mutually independent but also independent of arrivals both at Q_1 and at Q_2.

According to Burke's theorem, departures from Q_1 constitute a Poisson process with the rate λ. By the interpretation based on Burke's theorem, Q_2 behaves as if it were a stable M/M/1 queue isolated from Q_1. Such interpretation allows Q_1 and Q_2 to be viewed as a Jackson network of queues [6], which

will be used here as a counterexample to show that Jackson's theorem does not hold, and its proofs are all flawed. By considering this simple network, we shall not only find out what is wrong in Jackson's theorem in a straightforward way, but also see why time reversibility cannot explain away the inconsistent results in the literature.

According to Jackson's theorem [5], the numbers of customers in Q_1 and Q_2 are independent, and follow a product-form joint distribution after the network has been in operation for an infinitely long time. For these two queues, Eq.(4.1) is simply

$$\lambda_1 = \lambda_2 = \lambda \tag{4.2}$$

which can also be obtained by using Burke's theorem.

In the literature, it is claimed that Q_2 is a stable queue; the claimed stability of Q_2 follows from Eq.(4.2). However, because times between consecutive departures (i.e., $\tau_{j+1} - \tau_j$) from Q_1 are times between consecutive arrivals at Q_2, and because $\tau_{j+1} - \tau_j$ cannot be described by any single, fixed random variable, the number of customers in Q_2 cannot be described by any single, fixed random variable either, even if Q_1 is stable and statistical equilibrium obtains with respect to the number of customers in Q_1.

The unjustified assumption underlying Eq.(4.1) and Eq.(4.2) is the following: times between successive departures from a stable queue in steady state follow a marginal distribution. This assumption is the basis to define the arrival rate at a queue for customers coming from inside of the system. By Lemma 2.1, for the network of Q_1 and Q_2 in tandem, the assumption does not hold, unless the server at Q_1 has an infinite service capacity; only in this unrealistic scenario, treating Q_2 as a stable M/M/1 queue isolated from Q_1 will not lead to the inconsistent results. Because service capacities in the real world must be finite, it is illegitimate to use λ, the arrival rate at Q_1, to characterize the arrivals at Q_2. As terms of a sequence with two-fold randomness, $\tau_{j+1} - \tau_j$ do not have a marginal distribution; time reversibility cannot change this fact, see also Subsection 3.2.

In general, so long as inter-arrival times between customers at a queue are times between consecutive departures from another queue, any single, fixed random variable cannot describe such inter-arrival times. Consequently, any single, fixed random vector cannot describe the behavior of a Jackson network of queues, regardless of whether the structure of the network is simple or complex; in a Jackson network, with or without feedback paths, at least one queue is not stable. That is, statistical equilibrium with respect to the numbers of customers in all the queues in the network as a whole does not exist. Therefore, Jackson's theorem is false, and no Jackson network is stable.

By definition [1], if the number of customers in a queue remains finite after the queue has been in operation for an infinitely long time, the queue is sub-stable. A stable queue is of course sub-stable. But a sub-stable queue may not necessarily be stable. If a queue is sub-stable but not stable, the queue is properly sub-stable. If a queue is properly sub-stable, the number of customers in the queue is always finite, but its distribution will not converge to a limit.

That is, the behavior of the queue cannot be described by any single, fixed random variable. The meaning of "not stable" is not "unstable"; the latter means "not sub-stable". The number of customers in an unstable queue will become infinitely large as time approaches infinity. As we have seen, in Jackson's theorem, properly sub-stable queues are mistaken for stable queues.

Now consider a system of work-conserving, single-server queues in series. Each queue has a finite service capacity and an infinite waiting room. At each queue, service times are generally distributed, mutually independent, and independent of its arrivals. All customers arrive from outside of the system at the first queue, which is a stable GI/GI/1 queue in statistical equilibrium. After receiving service at a queue that is not the last queue in the system, a customer goes immediately to the next queue; all customers leave the system from the last queue after being served there.

Except the first queue, other queues in the system, called the downstream queues, are not stable; they are merely properly sub-stable, and this system of queues in series as a whole is not stable, in the sense that its behavior cannot be described by any single, fixed random vector. As shown above, for the downstream queues, the two different notions, proper sub-stability and stability, are confused in the literature; the confusion leads to mistaking properly sub-stable queues for stable queues. Such proper sub-stability due to two-fold randomness exhibited in the departure processes is entirely ignored.

The above analysis can be generalized in several ways straightforwardly. For example, it may apply to a queuing network with a more general topological structure; we may also allow each queue in the network to have multiple work-conserving servers with finite service capacities, and the waiting room may not necessarily be infinite; external arrivals at a queue may not necessarily form a renewal process; service times may have different distributions at different queues in the network; external arrivals and service times may also be dependent.

5 Conclusion

In this article, counting processes with multiple randomness, which differ essentially from known stochastic processes in the existing literature, are introduced. With examples in queuing theory, the existence of these new stochastic processes is demonstrated, and their properties are illustrated. By identifying counting processes with two-fold randomness in queuing models, the long-standing inconsistencies concerning Burke's theorem and Jackson's theorem are resolved.

References

[1] R. M. Loynes, The stability of a queue with non-independent inter-arrival and service times, *Proc. Camb. Philos. Soc.*, 58(1962), 497-520, DOI 10.1017/S0305004100036781.

[2] P. J. Burke, The output of a queuing system, *Operations Research*, 4(1956), 699-704, DOI 10.1287/OPRE.4.6.699.

[3] E. Reich, Waiting times when queues are in tandem, *Ann. Math. Statist.*, 28(1957), 768-773, DOI 10.1214/AOMS/1177706889.

[4] F. P. Kelly, **Reversibility and Stochastic Networks**, John Wiley & Sons, New York, 1979.

[5] J. R. Jackson, Networks of waiting lines, *Operations Research*, 5(1957), 518-521, DOI 10.1287/OPRE.5.4.518.

[6] J. R. Jackson, Job-shop like queueing systems, *Management Science*, 10(1963), 131-142, DOI 10.1287/MNSC.1040.0268.