Probability in the Engineering and Informational Sciences, **31**, 2017, 1–42.
doi:10.1017/S0269964816000371

APPROXIMATING AND STABILIZING DYNAMIC RATE JACKSON NETWORKS WITH ABANDONMENT

JAMOL PENDER
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York, USA
E-mail: jjp274@cornell.edu

WILLIAM A. MASSEY
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey, USA
E-mail: wmassey@princeton.edu

In this paper, we generalize the *Gaussian Variance Approximation* (GVA), developed by Massey and Pender [16], to Jackson networks with abandonment. We approximate the queue length process with a multivariate Gaussian distribution and thus, we are able to estimate the mean and covariance matrix of the entire network with more accuracy than the associated fluid and diffusion limits of Mandelbaum, Massey, and Reiman [14]. We also show how the GVA method can be used to construct staffing schedules that approximately stabilize salient performance measures such as the probability of delay and the abandonment probabilities for the entire network. Unlike the work of Feldman et al. [5] which uses Monte Carlo simulation to stabilize the delay probabilities, our method does not require simulation and only requires the numerical integration of $\frac{1}{2}(N^2 + 3N)$ differential equations for an N-dimensional network, which is more computationally efficient. Lastly, to confirm our approximations are accurate, we perform several numerical experiments for a wide range of parameter settings.

Keywords: queueing theory, applied probability, computational probability, simulation, probabilistic networks, stochastic modeling

1. INTRODUCTION

Large-scale service systems such as healthcare centers, data centers, and call centers have very similar dynamics of customer flow. The arrival of patients, customers, or data typically varies from day to day and even within the day itself. This behavior suggests that any stochastic queueing model must incorporate time-varying arrival rates. Moreover, these systems have a large number of customers and these customers have access to many parallel servers. In the healthcare setting, the servers can be the beds or nurses. In a data center setting, the servers are the computers and machines performing the necessary work.

 0269-9648/16 $25.00

Furthermore, if customers feel as if their wait for service is excessively long, they may choose to abandon the system, or start another process in the meantime. Lastly, customers may have to interact with several agents or types of services before they can leave the system.

A fundamental Markovian time-varying queueing network model that exhibits many of these features for service systems is a Jackson network with abandonment. This model assumes that each node of the Jackson network is a multi-server queue with non-homogeneous Poisson external arrivals, exponentially distributed service as well as customer abandonment times, and random routing through the network. One way to analyze the Jackson network with abandonment is to use Monte Carlo simulation. However, Monte Carlo simulation is computationally intensive, especially when a network has a large number of nodes or stations. Thus, we are motivated by this computational problem to develop alternative methods and numerical algorithms that estimate the non-stationary dynamics of the Jackson network without using Monte Carlo simulation.

One method that has been used to describe the dynamics of Jackson networks is to analyze the transition probabilities of the network. However, even in the one server setting with one node and constant coefficients, an analysis of the transition probabilities is non-trivial as they involve infinite sums of modified Bessel functions. Moreover, with time-varying arrivals, service, abandonment, and in the network setting, a full understanding of the transition probabilities requires the numerical integration of an infinite number of differential-difference equations. Thus, it is important to develop new ways of analyzing the moment behavior of the Jackson network with abandonment that avoids using the transition probabilities.

A more practical method to analyze the Jackson network is to use asymptotic techniques. In the work of Mandelbaum, Massey, and Reiman [14], the authors prove both a functional strong law of large numbers and a functional central limit theorem for the Jackson network with abandonment. They show using strong approximations that the functional strong law of large numbers limit converges to a deterministic dynamical system or fluid model. Moreover, the functional central limit theorem converges to a Gaussian diffusion under mild conditions. The mean and covariance of the Gaussian diffusion can be combined with the fluid model that describes the mean and covariance behavior of the original stochastic network model when the number of servers and the arrival rate are large.

The theory provided by the limit theorems is appropriate for models where the arrival rate and the number of servers are large. It is well known from examples given in Mandelbaum et al. [15], Ko and Gautam [8], and Massey and Pender [16] that the theory may yield less precise approximations when the arrival rate and number of servers is small or during the times when the number of servers is equal to or close to the mean queue length. The latter is known as the *lingering* condition. However, the work of Massey and Pender [16,17] proposes a new technique called the *Gaussian Variance Approximation* (GVA) to correct some of the inaccuracies of the fluid and diffusion limits in the one node case. This extends a method first formulated by Ko and Gautam [8] for multi-server queues with retrials. In this paper, we generalize the GVA to the Jackson network with abandonment model. We show that the GVA improves the estimation of the mean and covariance matrix of the network. Accurate values of these statistics for the network are not only necessary for managers to understand the dynamics of their network. They are also needed to accurately staff the networks properly to meet performance targets.

In addition to our improved estimates of the mean and covariance behavior of the network, we also show that it is possible to construct staffing algorithms for stabilizing various performance measures that are of importance to the queueing community. In fact,

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

we demonstrate that we can stabilize the probability of delay and abandonment probabilities of each node in the network using approximate values for the transition rates of each node. These performance measures have been stabilized in one-dimensional queueing processes, see for instance, Jennings et al. [6], Liu and Whitt [12]. In the context of multi-dimensional networks the work of Defraeye and Van Nieuwenhuyse [2] and Yom-tov and Mandelbaum [31] consider retrial queueing models and propose non-stationary staffing schedules to stabilize delay probabilities. Moreover, recent work by Liu and Whitt [11,13] considers a feed forward network a retrial queue and attempts to stabilize performance using infinite server queue approximations. However, no such work covers all of these performance measures in the full Jackson network setting, thus making this the first paper to stabilize these performance measures in the network setting although in the Markovian setting. Moreover, these methods not based on simulation or the infinite server queue length process, but are based on using our Gaussian approximations and simple ratios of the transition rates of the queue length processes, which is a new and novel idea in the literature.

1.1. Contributions

Our contributions in this work are the following.

1. We obtain accurate estimates for the mean and covariance matrix of the $(M_t/M_t/c_t + M_t)^N$ queueing network in critical and non-critically loaded regions.
2. We give closed form approximations for the probability of delay and derive a staffing schedule that stabilizes the delay probabilities at each node of the $(M_t/M_t/c_t + M_t)^N$ queueing network.
3. We develop a new approach to stabilizing performance measures of queueing networks using the transition rates of the queueing network. This new approach is applied to stabilize the abandonment probabilities of each node of the Jackson network.
4. We provide several numerical examples to demonstrate the effectiveness of our approximations and the how we stabilize the performance of the network.

1.2. Organization of the Paper

The rest of the paper continues as follows. In Section 2, we review our queueing model and the associated fluid and diffusions limits derived in [14]. We also provide explicit expressions for the functional Kolmogorov forward equations for the mean and covariance matrix for the Jackson network and we show how to simulate non-stationary Jackson networks as well as integrate dynamical systems. In Section 3, we describe our new methods for estimating the mean and covariance dynamics of the Jackson network with abandonment. In Section 4, we also show how to use our new method of approximating the moment dynamics to construct stabilizing algorithms for the probability of delay at positive target values. In Section 5, we show how to use the transition rates of the Jackson network in order to stabilize the abandonment probabilities of the network at each node. In Section 6, we conclude and identify opportunities for future work. In the Appendix, we provide several numerical examples to demonstrate the effectiveness of our approximations and stabilizing methods. We also extend the Deterministic Mean Approximation (DMA) and GVA approximations to the case of loss networks that have state-dependent arrival rate functions.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

1.3. Notation

The paper will use the following notation:

- $\lambda_i(t)$ is the external arrival rate to node i at time t.
- $\beta_i(t)$ is the abandonment rate for node i at time t.
- $\mu_i(t)$ is the service rate for node i at time t.
- $\tau_{ij}(t)$ is the abandonment routing probability from node i to node j at time t.
- $\gamma_{ij}(t)$ is the service routing probability from node i to node j at time t.
- $\tau_i(t)$ is the abandonment departure probability from node i at time t.
- $\gamma_i(t)$ is the service departure probability from node i at time t.
- $c_i(t)$ is the number of servers for node i at time t.
- $x \wedge y = \min(x, y)$.
- $(x - y)^+ = \max(0, x - y)$.
- $x \circ y =$ Hadamard or componentwise product of x and y.
- $x \otimes y =$ Kronecker product of two vectors x and y.
- $\mathbf{v}_i =$ jump vectors as explained in Section 2 of [14].
- $\Delta(\mu) =$ diagonalization of vector μ.
- $\{x < y\}$ denotes an *indicator function* that equals. one if the statement is true that is, if $x < y$, and zero if the statement is false.
- P^{s} denotes the matrix of service departure routing probabilities.
- P^{a} denotes the matrix of abandonment routing probabilities.
- $\varphi(x)$ is the probability density function (pdf) of the standard normal distribution.
- $\Phi(x) = 1 = \overline{\Phi}(x)$ is the cumulative distribution function (cdf) of the standard normal distribution.
- e_i is a vector of length N where the ith entry is equal to one and all other entries are zero.

We also require the following conditions:

$$\tau_i(t) + \sum_{j=1}^{N} \tau_{ij}(t) = 1 \quad \text{and} \quad \gamma_i(t) + \sum_{j=1}^{N} \gamma_{ij}(t) = 1. \tag{1.1}$$

These conditions ensure that the outflow from each node matches the number of departures from the system and the inflow to other nodes.

2. JACKSON NETWORK WITH ABANDONMENT

Markovian Jackson networks with abandonment are the standard models to analyze when one wants to model large-scale service systems where each node or location has multiple servers and impatient customers. In the context of telecommunication systems, call centers with interactive voice response (IVR) can be modeled as Jackson networks with abandonment, see for example Khudyakov, Feigin, and Mandelbaum [7]. In the context of healthcare systems, hospital networks where a patient may have to interact with the hospital administration, nurses, doctors, medical machinery, and hospital staff can also be modeled as a

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Jackson network with abandonment, see for example Yom-Tov and Mandelbaum [31] or Vericourt and Jennings [28].

The Jackson network with customer abandonment is also a special case of a Markovian service network. Markovian service networks were first considered in the work of Mandelbaum et al. [14], where fluid and diffusion limit theorems were proved using strong approximations. In Mandelbaum et al. [14] it is shown that each queue length process can be written in terms of time-changed Poisson processes and can be represented by the following stochastic integral equation for all i such that $1 \leq i \leq n$

$$\begin{aligned} Q_i(t) = Q_i(0) &+ \Pi_i^{\mathrm{a}}\left(\int_0^t \lambda_i(s)ds\right) - \sum_{j=1}^{N} \Pi_{ij}^{\mathrm{r}}\left(\int_0^t (Q_i(s) - c_i(s))^+ \cdot \beta_i(s) \cdot \tau_{ij}(s)ds\right) \\ &+ \sum_{j=1}^{N} \Pi_{ji}^{\mathrm{r}}\left(\int_0^t (Q_j(s) - c_j(s))^+ \cdot \beta_j(s) \cdot \tau_{ji}(s)ds\right) \\ &- \sum_{j=1}^{N} \Pi_{ij}^{d}\left(\int_0^t (Q_i(s) \wedge c_i(s)) \cdot \mu_i(s) \cdot \gamma_{ij}(s)ds\right) \\ &+ \sum_{j=1}^{N} \Pi_{ji}^{d}\left(\int_0^t (Q_j(s) \wedge c_j(s)) \cdot \mu_j(s) \cdot \gamma_{ji}(s)ds\right) \\ &- \Pi_i^{\mathrm{r}}\left(\int_0^t (Q_i(s) - c_i(s))^+ \cdot \beta_i(s) \cdot \tau_i(s)ds\right) \\ &- \Pi_i^{d}\left(\int_0^t (Q_i(s) \wedge c_i(s)) \cdot \mu_i(s) \cdot \gamma_i(s)ds\right), \end{aligned} \tag{2.1}$$

where each of the Π_i, Π_{ij}'s are each independent unit rate Poisson processes. Time changes of these Poisson processes give each process a probabilistic interpretation in terms of the queue length processes. A deterministic time change for each Π_i^{a} transforms it into a non-homogeneous Poisson arrival process with rate $\lambda_i(t)$. This arrival process counts the number of exogenous customer arrivals to node i of the network that have arrived in the interval $(0, t]$. If we subject Π_i^d to a random time change, it counts the number of service departures from node i in the interval $(0, t]$. Moreover, a random time change for Π_{ij}^d represents the number of customers that receive service from node i and move to node j. Similarly, we have that Π_i^{r} represents the number of customers that have abandoned from node i during the interval $(0, t]$. Lastly, the process Π_{ij}^{r} counts the number of customers that do not receive service and abandon from node i and moves to node j.

It is important to also note that the parameters that define the time changed Poisson processes $(\mu, \beta, \tau, \gamma)$ are also themselves non-stationary and therefore have time dependence. In terms of the service and abandonment processes, this means that the service rate of a customer in service at time t is served at rate $\mu(t)$ and customers that are not in service abandon the system at rate $\beta(t)$. In the context of service completions, time dependent rates can be used to model the fact that servers might slow down when they are tired during a particular part of the day (after lunch for example). In the context of abandonments, customers might be more likely to abandon during certain periods of the day. Lastly, in terms of routing probabilities, patients might be routed to different parts of a hospital depending on what time they might arrive and what condition they have. This is likely to change over time as doctors and nurses do not sit in just one place during the entire day.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Finally, one should also note that these time changes depend on the number of servers $c_i(t)$, which is a time-varying parameter for each node of the network. Although there is no problem with displacement of customers when the number of servers is increased, when the number of servers is decreased, then we must be concerned removing customers that might have been in service. There are several approaches to this issue. If one uses uniformization or the simulation approach given in [16], then this is irrelevant since the simulation only depends on rates. However, if the simulation is implemented in another fashion, then this concern is quite negligible. Given that all of the servers are identical, one method is to take the customer with the smallest remaining service time and have that customer either wait for the next server to complete their service or to hold the server with the minimum remaining service until the customer is finished. Since we are dealing with exponential random variables, taking the minimum over the remaining service times (which is also exponential) has a very small mean and variance and can be considered negligible in entire simulation. However, this is irrelevant for us as we use the approach in [16].

Remark 2.1: From this point on, we will suppress time dependence for the queueing network parameters and processes. This is just a notational convenience, however, all parameters in subsequent sections have time dependence.

2.1. Fluid and Diffusion Limits

The traditional way to analyze the moment behavior of queueing networks is to use asymptotic methods. In Mandelbaum et al. [14], general limit theorems were developed for *Markovian service networks*. Using these results for Jackson networks with abandonment, one can construct an associated, *uniformly accelerated* queueing process à la Halfin-Whitt, where the new arrival rate function is $\eta \cdot \lambda$ and the new number of servers is $\eta \cdot c$ for some scale factor $\eta > 0$. Moreover, by taking the following pointwise limits yields the *fluid* models of Mandelbaum et al. [14].

THEOREM 2.2: *If* $\lim_{\eta\to\infty} \frac{1}{\eta} Q^\eta(0) = q(0)$ *almost surely, then we have that*

$$\lim_{\eta\to\infty} \frac{1}{\eta} Q^\eta(t) = q(t) \quad \text{a.s.}, \tag{2.2}$$

where the deterministic process $q(t) \equiv \{q_1(t), q_2(t), \ldots, q_N(t) | t \geq 0\}$, *the fluid mean, is governed by the* N*-dimensional dynamical system*

$$\begin{aligned} \dot{q}_i = \lambda_i - \mu_i \cdot (q_i \wedge c_i) - \beta_i \cdot (q_i - c_i)^+ \\ + \sum_{j=1}^N (q_j - c_j)^+ \cdot \beta_j \cdot \tau_{ji} + \sum_{j=1}^N (q_j \wedge c_j) \cdot \mu_j \cdot \gamma_{ji} \end{aligned} \tag{2.3}$$

for all $1 \leq i \leq n$.

In addition to the fluid limits, if we normalize the queue length by the fluid limit and rescale by $\sqrt{\eta}$, then we get the following diffusion limits for the queue length process:

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

THEOREM 2.3: *If* $\lim_{\eta\to\infty} \frac{1}{\sqrt{\eta}} Q^{\eta}(0) - \sqrt{\eta} q(0) \stackrel{d}{=} \hat{Q}(0)$, *then we have that*

$$\lim_{\eta\to\infty} \sqrt{\eta} \cdot \left(\frac{1}{\eta} Q^{\eta(t)} - q(t) \right) \stackrel{d}{=} \hat{Q}(t), \tag{2.4}$$

where $\hat{Q}$ *is the same* N*-dimensional diffusion process in derived in Mandelbaum et al. [14]. Moreover, the diffusion limit suggests that if the set*

$$A^N \equiv \bigcup_{i=1}^{N} \{ t \mid q_i(t) = c_i(t) \} \tag{2.5}$$

has Lebesgue measure zero, then $\hat{Q}(t) \equiv \{\hat{Q}_1(t), \hat{Q}_2(t), \ldots, \hat{Q}_N(t) | t \geq 0\}$ *is an* N*-dimensional Gaussian diffusion process whose covariance* $\mathrm{Cov}[\hat{Q}, \hat{Q}] \equiv \mathrm{Cov}[\hat{Q}]$ *combines with the fluid mean to form a* $(N^2 + N)$*-dimensional dynamical system for the covariance matrix given by Eq. (2.3),*

$$\begin{aligned} \dot{v}_i = {} & \lambda_i + \mu_i \cdot (q_i \wedge c_i) + \beta_i \cdot (q_i - c_i)^+ - 2 \cdot \mu_i \cdot v_i \cdot \{q_i \leq c_i\} - 2 \cdot \beta_i \cdot v_i \cdot \{q_i > c_i\} \\ & + \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot (q_j \wedge c_j) + \sum_{j=1}^{N} \beta_j \cdot \tau_{ji} \cdot (q_j - c_j)^+ + \sum_{j=1}^{N} 2 \cdot \mu_j \cdot \gamma_{ji} \cdot v_{ij} \cdot \{q_j \leq c_j\} \\ & + \sum_{j=1}^{N} 2 \cdot \beta_j \cdot \tau_{ji} \cdot v_{ij} \cdot \{q_j > c_j\}, \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} \dot{v}_{ij} = {} & -\mu_i \cdot v_{ij} \cdot \{q_i \leq c_i\} - \beta_i \cdot v_{ij} \cdot \{q_i > c_i\} - \sum_{j=1}^{N} \mu_i \cdot \gamma_{ij} \cdot (q_i \wedge c_i) \\ & - \sum_{j=1}^{N} \beta_i \cdot \tau_{ij} \cdot (q_i - c_i)^+ - \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot (q_j \wedge c_j) - \sum_{j=1}^{N} \beta_j \cdot \tau_{ji} \cdot (q_j - c_j)^+ \\ & + \sum_{j=1}^{N} \mu_i \cdot \gamma_{ij} \cdot v_i \cdot \{q_i \leq c_i\} + \sum_{j=1}^{N} \beta_i \cdot \tau_{ij} \cdot v_i \cdot \{q_i > c_i\} \\ & - \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot (v_{ij} - v_j) \cdot \{q_j \leq c_j\} - \sum_{j=1}^{N} \beta_j \cdot \tau_{ji} \cdot (v_{ij} - v_j) \cdot \{q_j > c_j\} \end{aligned} \tag{2.7}$$

for all $i \neq j$.

2.2. Simulation of Jackson Network and Dynamical Systems

In this section, we give a description of how to simulate the Jackson network with abandonment using uniformization. In Figure 4, which is adapted from Massey and Pender [16], we illustrate how to simulate a Markovian multi-server queue using Monte Carlo simulation methods outlined in Ross [26]. The inner loop labeled *update* is given in Figure 5 and shows how to uniform random variables to construct the sample path behavior of the multi-server queue. For the Jackson network with abandonment we can use a similar transition rate

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

approach to generate the sample path dynamics of the entire network. In order to determine what Poisson process will jump next, we use a uniform random variable to determine the probability of each jump. Then we iterate this procedure until the final time is reached. In order to gain some intuition and insights about the dynamics of the Jackson network with abandonment, we use the simulation methodology outlined in Figures 4 and 5 and we simulate a two-dimensional Jackson network with abandonment. In fact, the numerical example that we present here will serve as the basis for comparing our new approximations in this paper, however, we provide more numerical examples in the Appendix to give the reader confidence that our approximations work in a variety of parameter settings. For our main numerical example we simulate the mean and variance of a two node Jackson network with abandonment with the parameters given in Table 1. In all of our simulations, the step size or Δt is equal to 10^{-4}, the number of averaged sample paths is equal to 10,000, and final time that we use is $T = 40$. In terms of numerically integrating ordinary differential equations like the ones obtained from the fluid and diffusion limits, we also consider a time step Δt that is equal to 10^{-4} and final time that is equal to $T = 40$. In order to integrate the differential equations, we also use the traditional Euler numerical scheme, which can be found in any standard textbook on differential equations. See the Appendix of [16] for more information on the Euler scheme.

Using the parameters of Table 1, in Figure 1 we compare the fluid mean with its simulated counterpart. In Figure 1, we see that the fluid mean of the queue does moderately well at approximating the non-stationary dynamics of the mean stochastic behavior of both queues. We observe that the accuracy of the fluid mean is also at its best when the stochastic

TABLE 1. Two Node Jackson Network Model Parameters

Parameter	Value	Parameter	Value
λ_1	$10 + 5\sin(t)$	λ_2	$10 + 2\sin(t)$
μ_1	1	μ_2	1
c_1	10	c_2	20
β_1	0.25	β_2	0.5
τ_1	0.25	τ_2	0.25
γ_1	0.25	γ_2	0.25
τ_{11}	0	τ_{22}	0
γ_{11}	0	γ_{22}	0
τ_{12}	0.75	τ_{21}	0
γ_{12}	0.75	γ_{21}	0

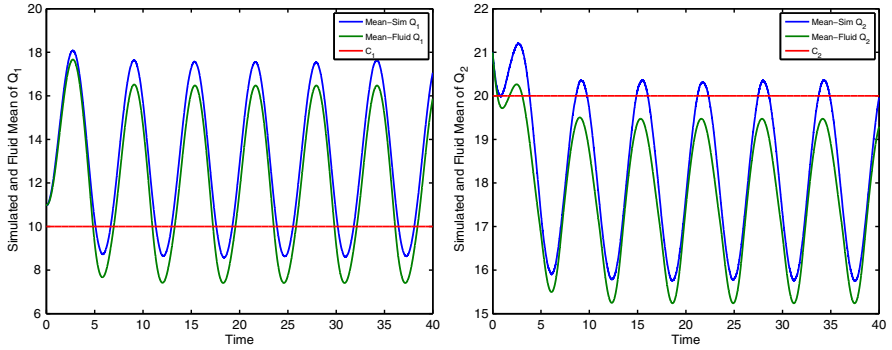

FIGURE 1. Simulated mean versus fluid limit mean of Q_1 (left). Simulated mean versus fluid limit mean of Q_2 (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

mean queue lengths are not near a local maximum, minimum, or when the number of servers crosses is equal to the fluid mean. Part of this phenomenon can be explained by the stochastic nature of the simulation. Since the queue length processes are stochastic in nature, they will experience larger queue lengths than those that are predicted by the fluid approximation. This explains partly why the simulated mean dominates the fluid mean.

In Figure 2, it is observed that the variance of the diffusion limit is not approximating the dynamics of the stochastic variance very well for each of the queue length processes. In fact, we notice that the diffusion limit approximation is the least accurate when the fluid limit mean is approximately equal to the number of servers. This is also precisely where the Gaussian limit theorems of [14] break down since there is a discontinuity in the variance and can be seen in the kink in the diffusion limit neat the times $t = \{5, 11, 17, 24, 30\}$. We should mention that it may appear that the largest difference between the variance is at the local maximum and minimum of the diffusion variance on the left of Figure 2, however, the largest discrepancy occurs at the kinks in the diffusion variance. This is precisely where the fluid mean and the number of servers are equal to each other. However, on the right side of Figure 2 we see that the largest discrepancy of the simulated and diffusion variance occurs at the local maximum of the each of the variances. This is precisely where the fluid mean of the second queue is closest to the number of servers and the lingering condition can begin go take effect.

On the left side of Figure 3, we plot the diffusion limit approximation of the covariance between the two queue length processes with its simulated counterpart. We also see that

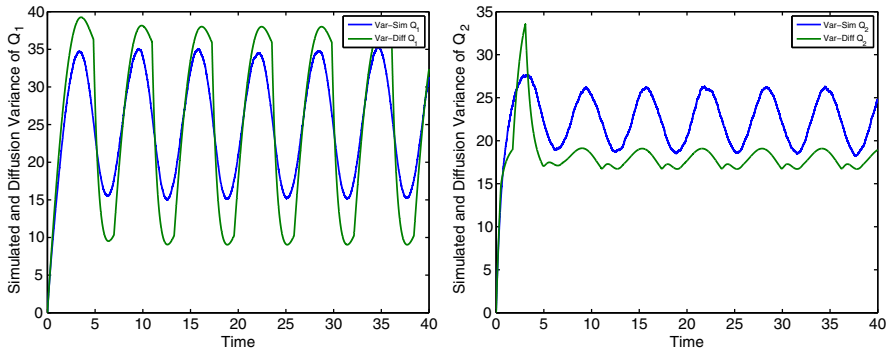

FIGURE **2.** Simulated variance versus diffusion variance of Q_1 (left). Simulated variance versus diffusion variance of Q_2 (right).

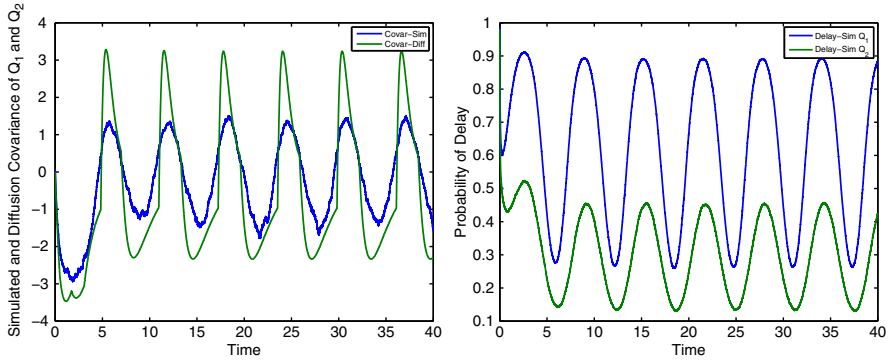

FIGURE **3.** Simulated covariance versus diffusion covariance (left). Simulated probability of delay (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

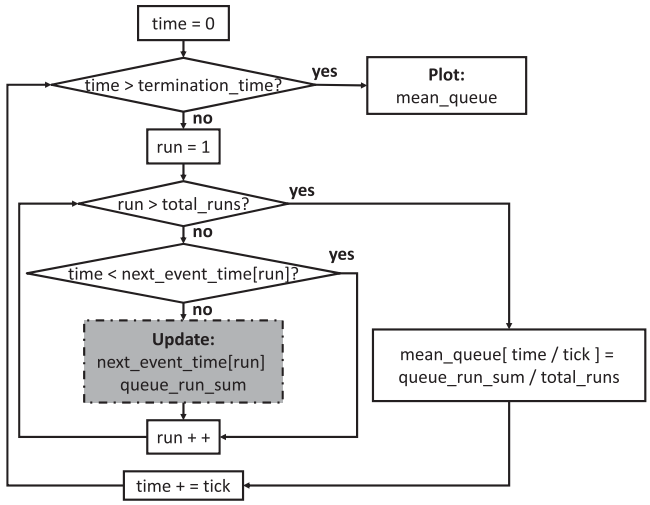

FIGURE 4. Algorithm for queueing network simulation.

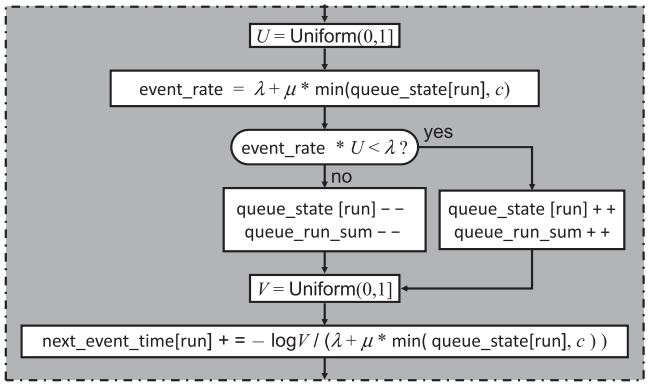

FIGURE 5. Update algorithm for queueing network simulation.

the diffusion limit does not estimate the dynamics of the covariance of the two-dimensional Jackson network well. Thus, we conclude from our simulations that we need significant improvements in the fluid and diffusion limits in order to accurately approximate the stochastic queueing processes in this example. On the right side of Figure 3, we also plot the delay probabilities of the two queue lengths. Since the number of servers are constant throughout time in this example, we see that the delay probability increases when the mean queue length for each queue increases. Moreover, when the queue lengths achieve their local maxima, then the probability of delay also achieves a local maxima since the number of servers is a constant function of time.

In all of the above figures, we observe that the mean and variance approximations need some improvement. In the sequel, we will generalize the GVA of [16] to provide refinements for the mean and covariance differential equations. However, in order to construct the GVA, we need to derive the functional version of the Kolmogorov forward equations for the $(M_t/M_t/c_t + M_t)^N$ queueing network.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

2.3. Functional Kolmogorov Forward Equations

The functional Kolmogorov forward equations are the first step in order to develop the GVA method for our Jackson network with abandonment. Following the same procedure given in Massey and Pender [16], we have the following representation for the marginal distribution of the functional Kolmogorov forward equations for the Jackson network with abandonment:

THEOREM 2.4: *Suppose that f is a polynomial function that maps from $\mathbb{R} \to \mathbb{R}$, then we have the following functional Kolmogorov forward equations for the Jackson network with abandonment*

$$\begin{aligned}\overset{\bullet}{E}[f(Q_i)] &= \lambda_i \cdot E[(f(Q+1) - f(Q))] + E[\delta_i(Q) \cdot (f(Q-1) - f(Q))] \\ &+ \sum_{j=1}^{N} E[D_{ij}(Q) \cdot (f(Q-1) - f(Q))] + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q) \cdot (f(Q+1) - f(Q))\right],\end{aligned} \tag{2.8}$$

where we have the following expressions for the rate functions:

$$\delta_i(Q) = \mu_i \cdot \tau_i \cdot (Q_i \wedge c_i) + \beta_i \cdot \gamma_i \cdot (Q_i - c_i)^+, \tag{2.9}$$

$$D_{ij}(Q) = \mu_i \cdot \tau_{ij} \cdot (Q_i \wedge c_i) + \beta_i \cdot \gamma_{ij} \cdot (Q_i - c_i)^+, \tag{2.10}$$

$$\tilde{D}_{ij}(Q) = \mu_j \cdot \tau_{ji} \cdot (Q_j \wedge c_j) + \beta_j \cdot \gamma_{ji} \cdot (Q_j - c_j)^+. \tag{2.11}$$

PROOF: See Engblom and Pender [4]. ■

Moreover, if one specializes to functions such as $\{Q_i, Q_i \cdot Q_j - E[Q_i] \cdot E[Q_j], (Q_i - E[Q_i])^2\}$, one gets the following expressions for the mean, covariance, and variance functions

COROLLARY 2.5: *The functional Kolmogorov forward equations for the mean, variance, and covariance of the Jackson network with abandonment have the following expressions:*

$$\overset{\bullet}{E}[Q_i] = \lambda_i - E[\delta_i(Q)] - \sum_{j=1}^{N} E[D_{ij}(Q)] + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right], \tag{2.12}$$

$$\begin{aligned}\overset{\bullet}{\mathrm{Var}}[Q_i] &= \lambda_i + E[\delta_i(Q)] + \sum_{j=1}^{N} E[D_{ij}(Q)] + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right] \\ &- 2 \cdot \mathrm{Cov}[Q_i, \delta_i(Q)] - 2 \cdot \sum_{j=1}^{N} \mathrm{Cov}[Q_i, \tilde{D}_{ij}(Q)] + 2 \cdot \sum_{j=1}^{N} \mathrm{Cov}[Q_i, D_{ij}(Q)],\end{aligned} \tag{2.13}$$

and

$$\begin{aligned}\overset{\bullet}{\mathrm{Cov}}[Q_i, Q_l] &= -\sum_{j=1}^{N} E[D_{ij}(Q)] - \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right] \\ &- \mathrm{Cov}[Q_i, \delta_l(Q)] - \mathrm{Cov}[Q_l, \delta_i(Q)] \\ &+ \sum_{j=1}^{N} \mathrm{Cov}[Q_i, D_{ij}(Q) - \tilde{D}_{ij}(Q)] - \sum_{j=1}^{N} \mathrm{Cov}[Q_l, D_{ij}(Q) - \tilde{D}_{ij}(Q)].\end{aligned} \tag{2.14}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Remark 2.6: It is worth noting that the functional forward equations are not-autonomous. This means that one needs to know a priori the distribution of the queue length process in order to compute the unknown expectations and covariance terms. The only exception to this is when $P^{\mathrm{s}} = P^{\mathrm{a}}$ and $\mu_i = \beta_i$, when the queueing network has the same dynamics as an infinite server queueing network. In this case the dynamics are much easier to analyze via a Poisson random measure approach and the work of Massey and Whitt [18] provides an in depth analysis for this case.

2.3.1. Matrix Version of Functional Forward Equations. The equations for the functional forward equations are very quite complicated since they involve many sum of different stations. However, below we present a more compact version of the dynamical system equations for the mean and covariance matrix of the Jackson network. For the mean we have that

$$\begin{aligned}\dot{E[Q]} &= \frac{d}{dt}E[Q] \\ &= \boldsymbol{\lambda} + \boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\boldsymbol{P}^s - \mathbb{I}) + \boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\boldsymbol{P}^a - \mathbb{I}) \\ &= \boldsymbol{\lambda} - \boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\mathbb{I} - \boldsymbol{P}^s) - \boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\mathbb{I} - \boldsymbol{P}^a)\,,\end{aligned}$$

and for the covariance we have that

$$\begin{aligned}\dot{\mathrm{Cov}}[Q,Q] &= \frac{d}{dt}\left(E[Q \otimes Q] - E[Q] \otimes E[Q]\right) \\ &= -\mathrm{Cov}[Q, Q \wedge \boldsymbol{c}] \cdot \mathbb{I} \otimes (\Delta(\boldsymbol{\mu}) \cdot (\mathbb{I} - \boldsymbol{P}^s)) - \mathrm{Cov}[Q \wedge \boldsymbol{c}, Q] \cdot (\Delta(\boldsymbol{\mu}) \cdot (\mathbb{I} - \boldsymbol{P}^s)) \otimes \mathbb{I} \\ &\quad - \mathrm{Cov}[Q, (Q - \boldsymbol{c})^+] \cdot \mathbb{I} \otimes (\Delta(\boldsymbol{\beta}) \cdot (\mathbb{I} - \boldsymbol{P}^a)) - \mathrm{Cov}[(Q - \boldsymbol{c})^+, Q] \cdot (\Delta(\boldsymbol{\beta}) \cdot (\mathbb{I} - \boldsymbol{P}^a)) \otimes \mathbb{I} \\ &\quad + \Delta\left(\boldsymbol{\lambda} + \boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\mathbb{I} - \boldsymbol{P}^s) + \boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\mathbb{I} - \boldsymbol{P}^a)\right) \\ &\quad - \Delta\left(\boldsymbol{\mu} \circ E[(Q \wedge \boldsymbol{c})] \cdot (\boldsymbol{P}^s \oplus \boldsymbol{P}^s)\right) - \Delta\left(\boldsymbol{\beta} \circ E[(Q - \boldsymbol{c})^+] \cdot (\boldsymbol{P}^a \oplus \boldsymbol{P}^a)\right).\end{aligned}$$

The matrix representation is extremely useful for numerically computing the mean and covariance by using simple matrix operations. In the next section, we will show how to combine the functional Kolmogorov forward equations with a multivariate Gaussian distribution to construct better approximations for the mean and covariance matrix of the queueing network.

3. DETERMINISTIC AND GAUSSIAN APPROXIMATIONS

In this section, we derive two approximations for the Jackson network with abandonment. These approximations are combine the functional forward equations with polynomial chaos expansions of stochastic processes. Polynomial chaos expansions have been used a variety of fields to estimate the the statistical properties of complex stochastic processes. The main idea is to expand a stochastic process in terms of simple random variables, that is,

$$Q(t,\xi) = \sum_{i=0}^{\infty} a_i(t) \cdot \psi_i(\xi) \tag{3.1}$$

where $a_i(t)$ are deterministic functions and $\psi_i(\xi)$ are multi-dimensional polynomials that are orthogonal with respect to a probability measure $w(\xi)$. One should note that this expansion

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

includes an infinite number of terms. However, as is common in many polynomial chaos expansions, one must truncated the expansion and use only a finite number of terms i.e.,

$$Q^n(t,\xi) = \sum_{i=0}^{n} a_i(t) \cdot \psi_i(\xi) \approx Q(t,\xi). \tag{3.2}$$

Hermite polynomial chaos expansions were first proposed by Wiener in [29] and they exploit Gaussian random variables. When combined with the Cameron–Martin theorem [1], the Hermite polynomial chaos can approximate any square integrable functional and $Q^n(t,\xi)$ converges to $Q(t,\xi)$ as $n \to \infty$ in the L_2 sense. Thus, using the polynomial chaos, one can expand any square integrable stochastic process in terms of orthogonal polynomials and more specifically the Hermite polynomials. We choose the Hermite polynomials for several reasons. The first reason is that we are inspired by the Gaussian heavy traffic limits of Mandelbaum et al. [14] for the Jackson network and therefore the probability measure that we choose for the orthogonal polynomial expansion is the Gaussian measure. This implies that the orthogonal polynomials that are relevant are the Hermite polynomials. The second reason is that the Hermite polynomials have special derivative properties such as Stein's lemma [27] and they are useful for smoothing out functions that are not differentiable or continuous, see for example Massey et al. [16] and Xiu et al. [30]. This is especially important in queueing where many of the functions that arise are maximum, minimum, or indicator functions. Lastly, since we want to improve the mean and variance dynamics of the queueing process, we are confident that this method works since it converges in the mean square sense. As a result, in the sequel, we will show that two terms of the expansion is enough to estimate the mean, covariance matrix, and several performance measures of our Jackson network quite well.

Remark 3.1: It is important to also mention that the polynomial chaos approach can be extended to other type of orthogonal polynomials like the Laguerre, which are orthogonal to the gamma distribution on the positive real line or the Poisson–Charlier polynomials which are discrete and are orthogonal to the Poisson distribution on the non-negative integers; see for example Engblom and Pender [4] or Pender [19–23] in the one-dimensional setting. For more work on polynomial chaos expansions, the reader should see Xiu and Karniadakis [30].

3.1. Deterministic Mean Approximation

Our first approximation take advantage of the fact that the first term in the polynomials chaos expansion is a deterministic function. Thus, we will use these deterministic functions to approximate the distribution of the queueing process. We call this the DMA since we assume $\{q(t)|t \geq 0\}$ is a deterministic process that approximates the queueing process. By applying the DMA to our Jackson network with abandonment, we arrive at our first theorem.

THEOREM 3.2: *If we substitute the approximate distribution $Q \equiv q$ for the queue length process of the Jackson network with abandonment, we get the following differential equations for the mean*

$$\begin{aligned} \dot{q}_i = \lambda_i - \mu_i \cdot (q_i \wedge c_i) - \beta_i \cdot (q_i - c_i)^+ \\ + \sum_{j=1}^{N} (q_j - c_j)^+ \cdot \beta_j \cdot \tau_{ji} + \sum_{j=1}^{N} (q_j \wedge c_j) \cdot \mu_j \cdot \gamma_{ji}. \end{aligned} \tag{3.3}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

PROOF: Since the deterministic mean approximation is not stochastic, we have that

$$\dot{E}[Q_i] = \lambda_i - E[\delta_i(Q)] - \sum_{j=1}^{N} E[D_{ij}(Q)] + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right] \tag{3.4}$$

$$\begin{aligned}\dot{E}[q_i] &= \lambda_i - \mu_i \cdot E[(q_i \wedge c_i)] - \beta_i \cdot E\left[(q_i - c_i)^+\right] \\ &\quad - \sum_{j=1}^{N} E\left[(q_j - c_j)^+\right] \cdot \beta_j \cdot \tau_{ji} + \sum_{j=1}^{N} E[(q_j \wedge c_j)] \cdot \mu_j \cdot \gamma_{ji}\end{aligned} \tag{3.5}$$

$$\begin{aligned}\dot{q}_i &= \lambda_i - \mu_i \cdot (q_i \wedge c_i) - \beta_i \cdot (q_i - c_i)^+ \\ &\quad + \sum_{j=1}^{N} (q_j - c_j)^+ \cdot \beta_j \cdot \tau_{ji} + \sum_{j=1}^{N} (q_j \wedge c_j) \cdot \mu_j \cdot \gamma_{ji}.\end{aligned} \tag{3.6}$$

■

This approximation, however, yields the same equation as the fluid limit results of [14]. This is not surprising in that the fluid limit can be interpreted as the best deterministic function that approximates the queueing network. Moreover, DMA in the network setting can be viewed as an one-dimensional projection onto the deterministic function $q(t)$. For the covariance, the DMA approximation yields a value of zero since the DMA implicitly assumes that the covariance and other cumulant moments are equal to zero. If we want to appropriately model other moments such as the covariance, then we must add an additional term to the approximation in order to add some randomness into the approximation.

3.2. Gaussian Variance Approximation

Now that we have seen that the fluid and diffusion approximations need some improvement we will show how we can generalize the work of [8] and [16] to improve the estimates of the mean and covariance matrix for the Jackson network with abandonment. In the one node case, [16] approximates the queue length process, by a Gaussian(q,v) random variable at each time point t. However, we since our network is N-dimensional, we will use a multivariate Gaussian distribution at each time point for the network case. To construct the GVA in the network setting, we first let $X_1, X_2, \ldots, X_N$ be N-independent Gaussian(0,1) random variables. Now we define the quantity $Z_1 = X_1$ and define recursively

$$Z_i = Z_{i-1} \cdot \cos\theta_{i-1} + X_i \cdot \sin\theta_{i-1}. \tag{3.7}$$

Using this recursive definition for the Z_i terms, it is easily seen that the Z_i random variables are distributed as Gaussian(0,1) random variables themselves, however, it needs to emphasized that they are not independent. Now using the definition of the Z_i random variables, we have the following construction of the GVA for each individual node of the network as

$$Q_i = q_i + \sqrt{v_i} \cdot Z_i. \tag{3.8}$$

This construction implies that if $i < j$ that

$$E[Q_i] = q_i, \tag{3.9}$$

$$\mathrm{Var}[Q_i] = v_i, \tag{3.10}$$

$$\mathrm{Cov}[Q_i, Q_j] \equiv v_{ij} = \sqrt{v_i v_j} \cdot \cos\theta_i \cdots \cos\theta_{j-1}. \tag{3.11}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Remark 3.3: The iterated cosine functions represent the correlation between two queue length processes. Moreover, this representation implies when the θ_i terms are non-zero that the queueing processes are *not* assumed to be independent of each other. Lastly, this method is equivalent to assuming that the queue length process is a multivariate Gaussian and is one of many such representations. However, we believe that our representation in terms of independent Gaussian allows us to calculate the closed form expressions for the rate functions in a systematic fashion without the use of Gaussian integrals. It should also be emphasized that the trigonometric representation is no different from the standard correlation representation. For instance in the two-dimensional case, $\cos(\theta) = \rho$ and $\sin(\theta) = \sqrt{1-\rho^2}$. There is no difference in representation; however, we find the trigonometric representation to be more useful although it is not standard.

Moreover, our independent Gaussian representation also allows us to derive closed form expressions for the θ_i terms as well. Using the adjacent nodes for the network, we arrive at the following formula for the θ_i terms, which are viewed as the inverse cosine of the correlations between the queue length nodes, that is,

$$\theta_i = \cos^{-1}\left(\frac{v_{ii+1}}{\sqrt{v_i \cdot v_{i+1}}}\right). \tag{3.12}$$

This representation of the correlation terms is useful for numerically integrating the differential equations for the mean and variance, but is not necessary to do so.

THEOREM 3.4: *If we substitute the above approximate distribution in Eq. (3.8) for each of the queue length processes of the Jackson network with abandonment, we get the following differential equations for the mean, covariance, and variance:*

$$\begin{aligned}\dot{q_i} &= \lambda_i - \mu_i \cdot \left(q_i - \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right) - \beta_i \cdot \left(\sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right) \\ &+ \sum_{j=1}^{N} \left(\sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right) \cdot \beta_j \cdot \tau_{ji} \\ &+ \sum_{j=1}^{N} \left(q_j - \sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right) \cdot \mu_j \cdot \varphi_{ji},\end{aligned} \tag{3.13}$$

$$\begin{aligned}\dot{v_i} &= \lambda_i + \mu_i \cdot \left(q_i - \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right) + \beta_i \cdot \left(\sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right) \\ &- 2 \cdot \mu_i \cdot v_i \cdot \Phi(\chi_i) - 2 \cdot \beta_i \cdot v_i \cdot \overline{\Phi}(\chi_i) + \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot \left(q_j - \sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right) \\ &+ \sum_{j=1}^{N} \beta_j \cdot \tau_{ji} \cdot \left(\sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right) \\ &+ \sum_{j=1}^{N} 2 \cdot \mu_j \cdot \gamma_{ji} \cdot v_{ij} \cdot \Phi(\chi_j) + \sum_{j=1}^{N} 2 \cdot \beta_j \cdot \tau_{ji} \cdot v_{ij} \cdot \overline{\Phi}(\chi_j),\end{aligned} \tag{3.14}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

and

$$\dot{v_{ij}} = -\mu_i \cdot v_{ij} \cdot \Phi(\chi_i) - \beta_i \cdot v_{ij} \cdot \overline{\Phi}(\chi_i) + \sum_{j=1}^{N} \mu_i \cdot \gamma_{ij} \cdot v_i \cdot \Phi(\chi_i) \tag{3.15}$$

$$+ \sum_{j=1}^{N} \beta_i \cdot \tau_{ij} \cdot v_i \cdot \overline{\Phi}(\chi_i) - \sum_{j=1}^{N} \mu_i \cdot \gamma_{ij} \cdot \left(q_i - \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right)$$

$$- \sum_{j=1}^{N} \beta_i \cdot \tau_{ij} \cdot \left(\sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right)$$

$$- \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot \left(q_j - \sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right)$$

$$- \sum_{j=1}^{N} \beta_j \cdot \tau_{ji} \cdot \left(\sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right)$$

$$- \sum_{j=1}^{N} \mu_j \cdot \gamma_{ji} \cdot (v_{ij} - v_j) \cdot \Phi(\chi_j) - \sum_{j=1}^{N} \beta_j \cdot \tau_{ji} \cdot (v_{ij} - v_j) \cdot \overline{\Phi}(\chi_j),$$

where

$$\chi_i = \frac{c_i - q_i}{\sqrt{v_i}}. \tag{3.16}$$

In order to prove the main theorem, we will show that it suffices to derive the following closed form expressions for the expectation and covariance terms of present in the functional forward equations of the queueing network

$$\begin{aligned}
E[(Q_i \wedge c_i)] &= q_i - \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right), \\
E[(Q_i - c_i)^+] &= \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right), \\
\mathrm{Cov}\left[Q_j, (Q_i \wedge c_i)\right] &= v_{ji} \cdot \Phi(\chi_i), \\
\mathrm{Cov}[Q_j, (Q_i - c_i)^+] &= v_{ji} \cdot \overline{\Phi}(\chi_i), \\
\mathrm{Cov}[Q_j, Q_i] &= v_{ji}.
\end{aligned}$$

However, before we prove the main result, we need to prove a simple lemma about the Z_i random variables that we use for constructing the GVA algorithm. This simple lemma will be useful for the remainder of the calculations.

LEMMA 3.5: *For all $1 \leq i \leq n$, Z_i is a Gaussian(0,1) random variable.*

PROOF: We will use induction to proof this result. For Z_1, this is true by our definition of the Z_i. Now if Z_{i-1} is distributed as a Gaussian(0,1) random variable, then by our construction we have that

$$Z_i = \cos(\theta_i) \cdot Z_{i-1} + \sin(\theta_i) \cdot X_i. \tag{3.17}$$

Then, by the property of Gausisan random variables, we have that Z_i is a Gaussian(0,1) random variable since $\cos^2(\theta_i) + \sin^2(\theta_i) = 1$ and both Z_{i-1} and X_i are mean zero. ■

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

PROOF: See the Appendix. ■

We should mention that the GVA equations may seem quite intimidating at first, but a careful understanding of them makes it easier to digest the difference between them and the fluid and diffusion differential equations. If one views the GVA equations as smoothed or infinitely differentiable versions of the fluid and diffusion limits, they become much easier to understand. One result of the GVA is to "smooth" out any of the discontinuities that are present in the fluid and diffusion limits. In fact terms like $\mathrm{Cov}[Q_i, (Q_i - c_i)^+]$ are equal to indicator functions such as $\{q_i < c_i\}$ under the fluid and diffusion limit theorems since the covariance operation can be viewed as a directional derivative. However, terms like $\mathrm{Cov}[Q_i, (Q_i - c_i)^+]$ are equal to Gaussian cdfs such as $\Phi((c_i - q_i)/\sqrt{v_i}))$, which are infinitely differentiable and no longer discontinuous. Thus, numerically integrating the GVA equations is no harder than integrating the fluid and diffusion limits using Gaussian pdfs and cdfs instead of indicator, maximum, and minimum functions.

To gather more insight on the differences between GVA and the fluid and diffusion limits, in Figure 6, we compare the DMA and GVA means with their simulated counterparts. On the left and right of Figure 6, we see that the GVA is better at approximating the dynamics at all times for the first queue. On the right of Figure 6, we plot the $\log_{10}$ relative error of the approximations. It is clear that the GVA method is doing much better than the DMA or fluid limit. In Figure 7, we observe similar dynamics for the second queue length process. On

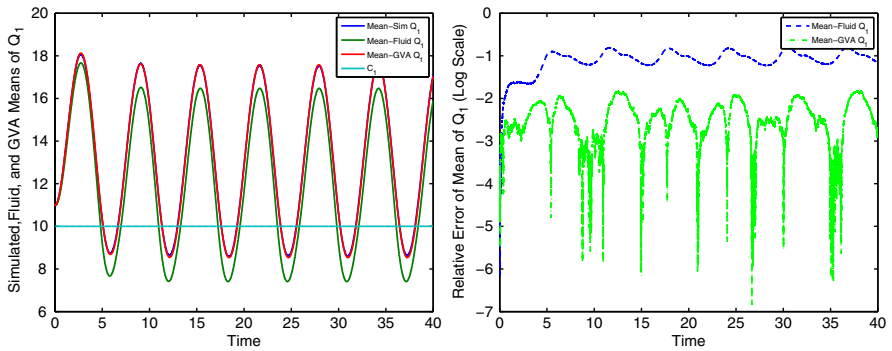

FIGURE **6.** Simulated mean versus DMA and GVA mean of Q_1 (left). Relative error of fluid limit versus GVA of Q_1 (right).

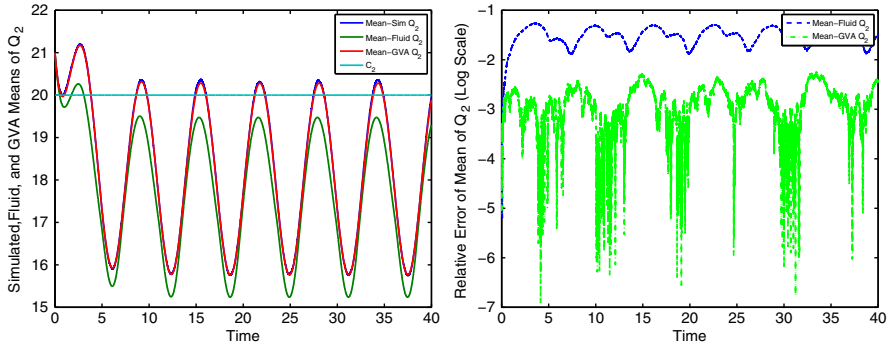

FIGURE **7.** Simulated mean versus DMA and GVA mean of Q_2 (left). Relative error of fluid limit versus GVA of Q_2 (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

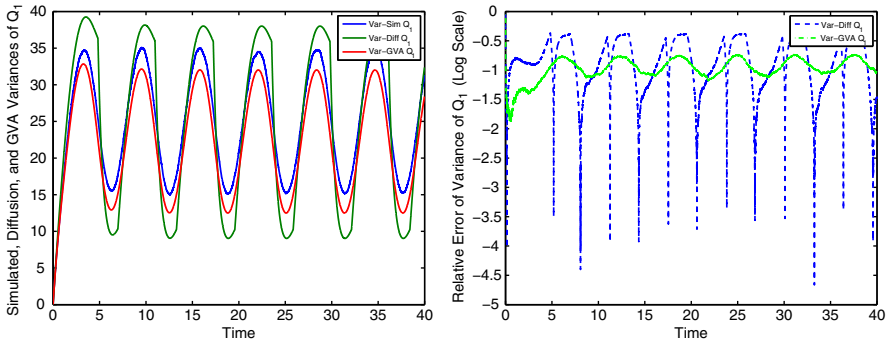

FIGURE **8.** Simulated variance versus GVA variance of Q_1 (left). Relative error of diffusion limit versus GVA of Q_1 (right).

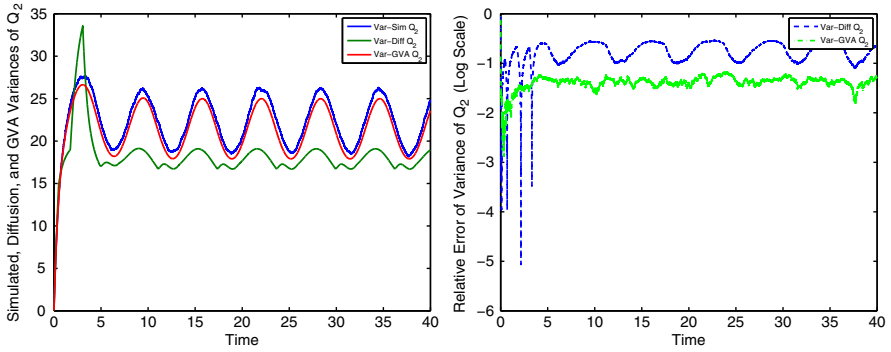

FIGURE **9.** Simulated variance versus GVA variance of Q_2 (left). Relative error of diffusion limit versus GVA of Q_2 (right).

the left side of Figures 8 and 9 we see that the GVA substantially improves the estimation of the dynamics of the variances of both queue lengths. On the right side of Figures 8 and 9 we see in the relative error plots that the GVA method is significantly improving the variance dynamics when compared with the diffusion limit. Lastly, in Figure 10 we see that GVA does a good job of estimating the covariance between the two queue length processes. It

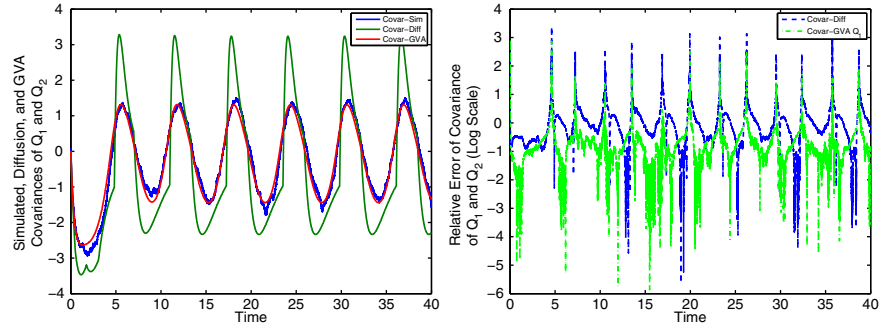

FIGURE **10.** Simulated covariance versus diffusion and GVA Covariance (left). Relative error of diffusion limit versus GVA of covariance (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

does a much better job than the diffusion limit as seen by the right side of Figure 10. Thus, in all of the figures, we observe that the GVA method is better than the fluid and diffusion limit theorems at estimating the mean and variance dynamics of the Jackson network with abandonment.

4. STABILIZING THE PROBABILITY OF DELAY

In this section, we explore the additional usefulness of using the GVA method. In addition to estimating the mean and covariance of the Jackson network with good accuracy, we can also use the GVA method to approximate the probability of delay for each station in the network. The fact that the GVA method can accurately estimate the probability of delay, which is an important performance measure is very useful. One reason is that it can allow managers to assess the impact of parameter changes on the customer performance at each node of the network making it unnecessary to simulate the network to assess the impact of perturbing one or many parameters.

4.1. Probability of Delay GVA

Similar to the one-dimensional case of the GVA algorithm we can approximate the probability of delay using GVA for each node of the Jackson network. The probability of delay is for each node of the network is the probability that the queue length exceeds or is equal to the number of servers, that is,

$$\mathbb{P}(Q_i \geq c_i) = \mathbb{P}(Z_i \geq \chi_i) \tag{4.1}$$

$$= E[\{Z_i \geq \chi_i\}] \tag{4.2}$$

$$= \overline{\Phi}(\chi_i). \tag{4.3}$$

Thus, by using the Gaussian tail cdf function, we can approximate the probability of delay for each node of the network. In fact in Figure 11, we see that we can provide good estimates for the probability of delay for our Jackson network example. Thus, our GVA method is generalizable to higher-dimensional examples of queueing networks is not just useful in the one-dimensional setting. Moreover, we can show that by inverting our probability of delay approximations, we can construct delay stabilizing staffing schedules for the entire network.

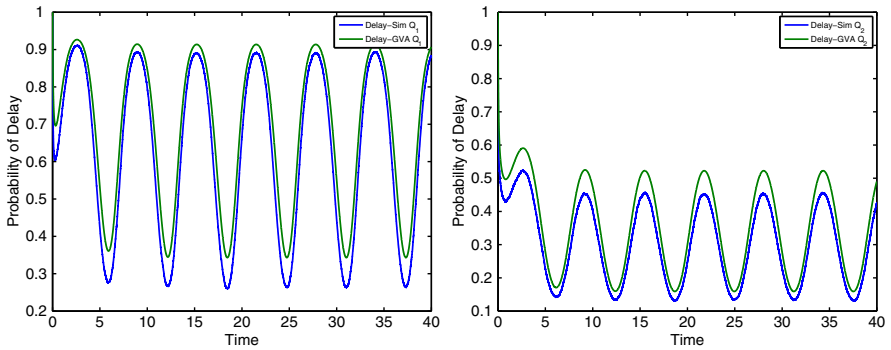

FIGURE **11.** Simulated and GVA probability of delay of Q_1 (left). Simulated and GVA probability of delay of Q_2 (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

4.2. GVA Stable Staffing Algorithm

The closed form expression for the probability of delay for each node of the network is an important object for constructing a staffing algorithm that stabilizes the probability of delay. Recall that the probability of delay approximation uses the tail cdf of the Gaussian distribution. Thus, using the properties of the cdf, one can invert the tail cdf to get

$$\overline{\Phi}(\chi_i) = \epsilon_i, \tag{4.4}$$

$$\overline{\Phi}^{-1}\left(\overline{\Phi}(\chi_i)\right) = \overline{\Phi}^{-1}(\epsilon_i), \tag{4.5}$$

$$\chi_i = \overline{\Phi}^{-1}(\epsilon_i), \tag{4.6}$$

$$c_i = q_i + \sqrt{v_i} \cdot \overline{\Phi}^{-1}(\epsilon_i). \tag{4.7}$$

Thus, by inverting the probability of delay approximation using the GVA method we see that we should use the following staffing function:

$$c_i = \left\lceil q_i + \sqrt{v_i} \cdot \overline{\Phi}^{-1}(\epsilon_i) \right\rceil \tag{4.8}$$

in order to stabilize station i of the queueing process with approximately ϵ_i probability of delay. Although the staffing procedure is no different than the one used by the fluid and diffusion approximation, the actual staffing functions are different because the dynamics for the mean and variance of the GVA method are better estimates of the true mean and variance obtained via simulation. Since GVA does a better job of estimating the mean and variance dynamics than the fluid and diffusion limits, the GVA should also produce a better staffing schedule for stabilizing the delay probabilities at their true target values.

This method of stabilization is deterministic and, thus is very powerful since it does not require the use of simulation, which is computationally expensive. Unlike Feldman et al. [5] there is no need to actually simulate the queueing system in order to update the staffing schedule. Our method simply requires the numerical solution of $\frac{1}{2}(N^2 + 3N)$ differential equations, which is computationally fast and does not require much computational effort especially in the large-scale setting or when there are a large number of nodes in the network.

4.3. Stabilization of the Feldman Example

Now using the fluid and diffusion limits and the GVA method, we will illustrate that the GVA method can not only stabilize the delay probabilities, but we also show that it stabilizes the delay probabilities better than the fluid and diffusion limits. In our numerical example, we use simulate a tandem Jackson network where the first queue has the same parameters as Figure 4 of Feldman et al. [5]. As for the second queue, the arrivals are identical to the service and abandonment departures of the first queue and the service rate and abandonment rate parameters are identical to that of the first queue. In Figure 12, we see that fluid and diffusion limits do a decent but not great job of stabilizing the delay probabilities at target delay values.

One can also observe that when the target is nearest to $\epsilon = 0.5$, the inaccuracy of the stabilizing method is the largest. One reason for the largest inaccuracy occurring near the target of $\epsilon = 0.5$ is that the queueing process is critically loaded at this point i.e $q(t) \approx c(t)$. It is at this precise point that the fluid and diffusion limits break down. However, we see in Figure 14 that the GVA method does a good job of stabilizing at the target delay values. We expect that the GVA should do better since it actually approximates the mean and variance quite well unlike the fluid and diffusion limits. In Figure 17, we also show using the log

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

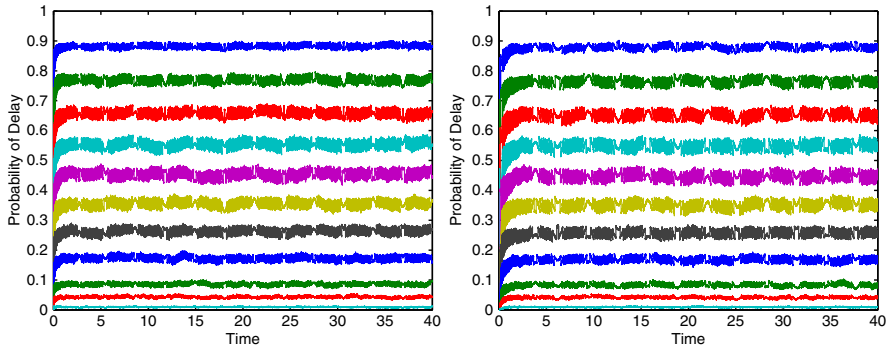

FIGURE **12.** Stable delay probabilities of Q_1 using fluid and diffusion staffing algorithm (left). Stable delay probabilities of Q_2 using fluid and diffusion staffing algorithm (right).

relative error that the GVA method outperforms the fluid and diffusion limits at stabilizing the delay probabilities. Thus, we see that the GVA method is better at stabilizing the queueing process at target values. To convince the reader further that this is not a special example, we demonstrate the outperformance of the GVA method in several additional numerical examples in the Appendix.

Remark 4.1: Moreover, this inaccuracy highlights the difference between the mean and the median. The probability of delay is more of a measure of the median, while when we staff at the level $\epsilon = 0.5$, we are staffing at the mean. One way to correct for this is to use the skewness of the queueing process, however, we do not consider skewness in this paper.

In Figures 13 and 15, we compute the staffing schedules for each queue using the fluid and diffusion algorithms. For each figure, we plot the staffing schedule that corresponds to the target delay probability of each queue length. The first observation that the staffing schedule is monotonically decreasing as we increase the probability of delay. This is natural because as we add more servers, the probability of delay should decrease to zero since the queue becomes more like an infinite server, which has no delay. The second observation that we make is that both the fluid and diffusion limits and the GVA have similar values and do not differ by much. This is further investigated in Figure 16, where we compare the staffing schedule for the fluid and diffusion limit theorems and the GVA method. We see

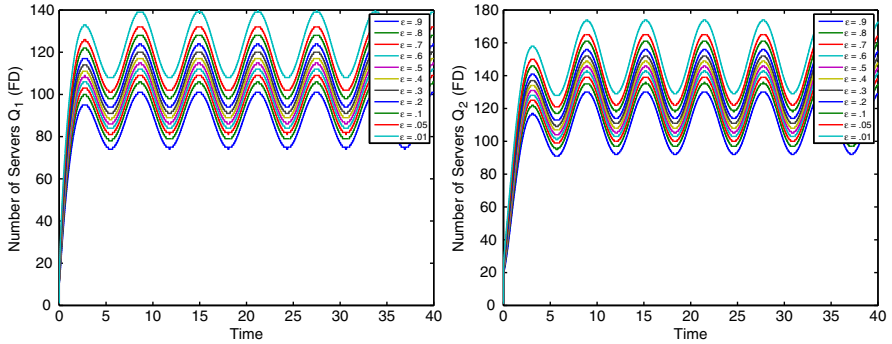

FIGURE **13.** Stable staffing schedules of Q_1 using fluid and diffusion (left). Stable staffing schedules of of Q_2 using fluid and diffusion (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

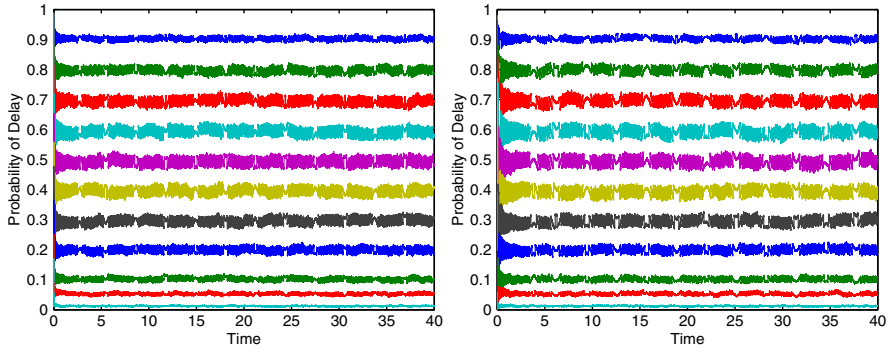

FIGURE **14.** Stable delay probabilities of Q_1 using GVA (left). Stable delay probabilities of Q_2 using GVA (right).

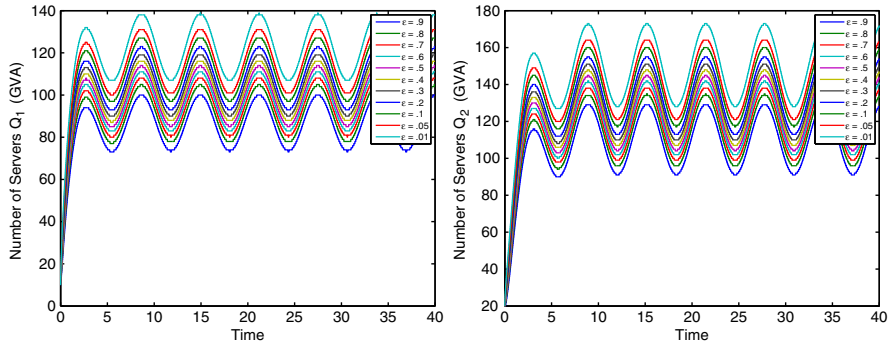

FIGURE **15.** Stable staffing schedules of Q_1 using (GVA) (left). Stable staffing schedules of of Q_2 using (GVA) (right).

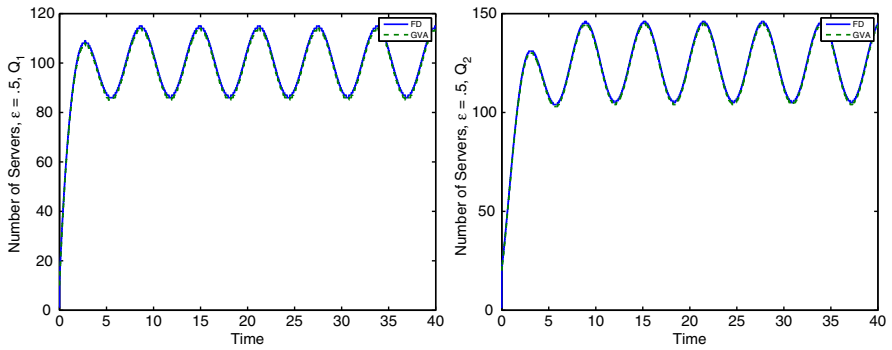

FIGURE **16.** Staffing schedule for Q_1 FD versus GVA (left). Staffing schedule for Q_2 FD versus GVA (right).

that the difference is about one server. This is consistent with the fact that the fluctuations due to adding or removing a server that are seen in Figures 12 and 14 is about 0.03 and the difference between the delay probabilities in Figures 12 and 14 is roughly 0.03 as well. Thus, the GVA is adding some value in producing a better staffing schedule than the fluid and diffusion limits even though it differs by one server. However, we will show in the Appendix

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

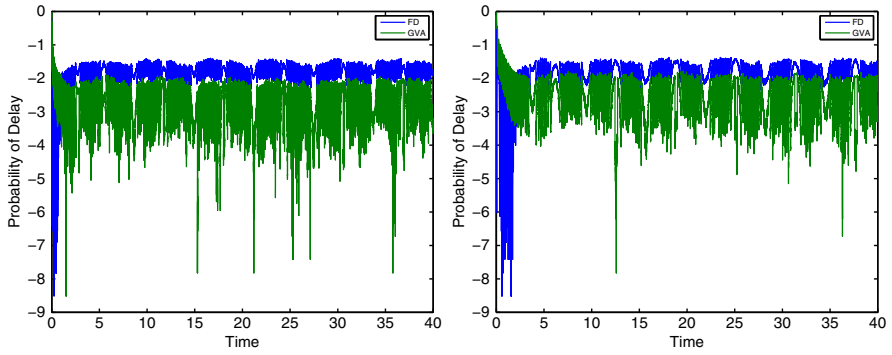

FIGURE **17.** Log relative error comparison when $\epsilon = 0.5$ of Q_1 (left). log relative error comparison when $\epsilon = 0.5$ of Q_2 (right).

where the difference between GVA and fluid and diffusion staffing schedules is much larger and GVA is once again better at producing more stable staffing functions.

5. STABILIZING THE ABANDONMENT PROBABILITIES

In addition to stabilizing the delay probabilities at positive target values, we also can also stabilize the abandonment probabilities using a new technique. Before we describe the new technique for stabilizing the abandonment probabilities, we provide a brief overview of how to simulate the non-stationary Erlang-A queue since it motivates our idea to construct approximate staffing functions for stabilizing the time dependent abandonment probabilities. The dynamics of the multi-server queue with abandonment can be split into three different type of events. The first event is for arrivals to the queueing process where the queue length is increased by one. The the second is for service departures where the queue length is decreased by one unless the queue is empty. Lastly, the third event is for customer abandonment from the queue where the queue length is also decreased by one unless no customers are waiting for service. Thus, the transition probabilities of the Erlang-A queue have the following representation

$$\begin{aligned}
\mathbb{P}(\Delta Q(t+\Delta t) = 1) &= \lambda \cdot \Delta t + o(\Delta t),\\
\mathbb{P}(\Delta Q(t+\Delta t) = -1) &= \mu \cdot (Q \wedge c) \cdot \Delta t + o(\Delta t),\\
\mathbb{P}(\Delta Q(t+\Delta t) = -1) &= \beta \cdot (Q-c)^+ \cdot \Delta t + o(\Delta t),\\
\mathbb{P}(\Delta Q(t+\Delta t) = 0) &= 1 - \lambda \cdot \Delta t - \mu \cdot (Q \wedge c) \cdot \Delta t - \beta \cdot (Q-c)^+ \cdot \Delta t + o(\Delta t).
\end{aligned}$$

It is assumed that the time interval Δt is sufficiently small so that the possibility of multiple events occurring in the time interval can be ignored. Moreover, in the above and following section we suppress time dependence to ease notation, however, it should be assumed that all parameters can depend on time.

5.1. A Fundamental Observation

From simulating the queueing process, we are able make the fundamental observation that we can approximate the probability that a customer will abandon at any time t. We define the probability of abandonment to be the probability of abandonment for an arrive that

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

enters at time t. This is the same as the *virtual abandonment probability* that is defined in Section 6.2 of [5]. In order to approximate this probability, we notice that we must calculate the following ratio, which is given by the transition probabilities of the queueing process

$$\mathbb{P}(\text{Abandon}) \approx \frac{\beta \cdot (Q-c)^+ \cdot \Delta t}{\mu \cdot (Q \wedge c) \cdot \Delta t + \beta \cdot (Q-c)^+ \cdot \Delta t}. \tag{5.1}$$

This ratio is the rate at which customers abandon from the queue divided by the total rate of departures from the queue regardless of whether they are service departures or customer abandonments. It is important to realize that this ratio is consistent with the uniformization approach of simulating the queueing process. In the uniformization approach, one samples a standard uniform random variable and chooses an abandonment departure or a service departure based on whether the uniform random variable is less than or greater than the above ratio. Thus, it is natural to consider the above ratio as a good approximation for the probability of abandonment. This interpretation should not be thought of as a pathwise approximation, but should be thought of as an approximation for the mean behavior of a large number of simulation runs. We will use this observation in conjunction with the moment estimates of the queue length process of [16] in order to derive our staffing functions that stabilize the abandonment probabilities. In fact, we can write the approximation in a simpler form that is more useful numerically since the staffing level c only appears once.

$$\mathbb{P}(\text{Ab}) \approx \frac{\beta \cdot E[(Q-c)^+]}{\mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+]} \tag{5.2}$$

$$= \frac{\beta \cdot E[(Q-c)^+]}{\mu \cdot (E[Q] - E[(Q-c)^+]) + \beta \cdot E[(Q-c)^+]} \tag{5.3}$$

$$= \frac{\beta \cdot E[(Q-c)^+]}{\mu \cdot E[Q] + (\beta - \mu) \cdot E[(Q-c)^+]} \tag{5.4}$$

$$= \frac{\beta}{\mu \cdot \frac{E[Q]}{E[(Q-c)^+]} - \mu + \beta}. \tag{5.5}$$

5.2. Probability of Abandonment (DMA)

Our fundamental observation is that when simulating a Markovian queueing system one can approximate the probability of abandonment as the following ratio of event times:

$$\mathbb{P}(\text{Ab}) \approx \frac{\beta}{\mu \cdot \frac{E[Q]}{E[(Q-c)^+]} - \mu + \beta}. \tag{5.6}$$

Using the approximation given by DMA, we can approximate the probability of abandonment by the following expression

$$\mathbb{P}(\text{Ab}) \approx \frac{\beta}{\mu \cdot \frac{E[Q]}{E[(Q-c)^+]} - \mu + \beta} \approx \frac{\beta}{\frac{\mu \cdot q}{(q-c)^+} - \mu + \beta}. \tag{5.7}$$

By splitting the probability into understaffed and overstaffed cases the probability of abandonment becomes

$$\mathbb{P}(\text{Ab}) = \begin{cases} \dfrac{\beta}{\dfrac{\mu \cdot q}{(q-c)} - \mu + \beta} & \text{if } q > c, \\ 0 & \text{if } q \leq c. \end{cases} \tag{5.8}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Moreover, since the probability of abandonment depends on the level of staffing c and is a non-increasing function of c, we can invert it to construct a staffing schedule that stabilizes the probability of abandonment. By inverting the approximation, it suggests solving the following equation to find the staffing schedule that stabilizes the probability of abandonment with a target level of ϵ

$$c^* = \inf \left\{ c \geq 0, c \in \mathbb{N} \;\middle|\; \frac{\beta}{\frac{\mu \cdot q}{(q-c)^+} - \mu + \beta} \leq \epsilon \right\}. \tag{5.9}$$

Solving for the optimal solution, we get that

$$c^* = \left\lceil q - \frac{\mu \cdot q}{\beta/\epsilon + \mu - \beta} \right\rceil. \tag{5.10}$$

Thus, it is apparently clear from the stabilizing solution that if ϵ is close to zero, then we will staff with the mean number of servers. Moreover, if ϵ is near one, then we staff the system with no servers. We will show later that when ϵ is near zero, we will need a further refinement to get proper stabilization.

We see that on the left of Figure 18 that the DMA does well at stabilizing the abandonment probabilities for target values greater than 0.1. However, we see that the DMA

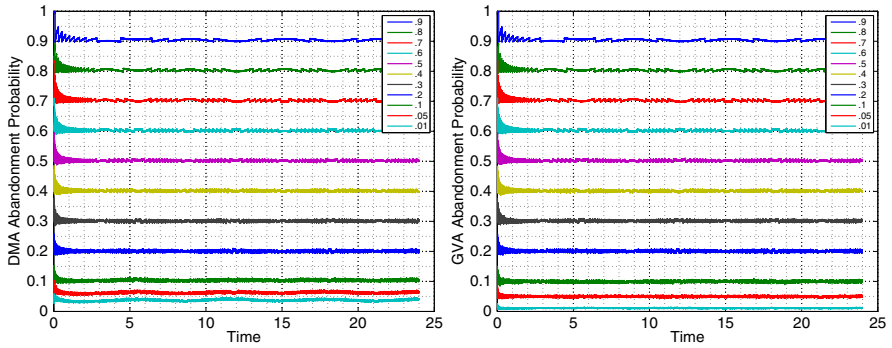

FIGURE **18.** Stable abandonment probabilities using DMA (left). Stable abandonment probabilities using GVA (right). $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = 0.5$

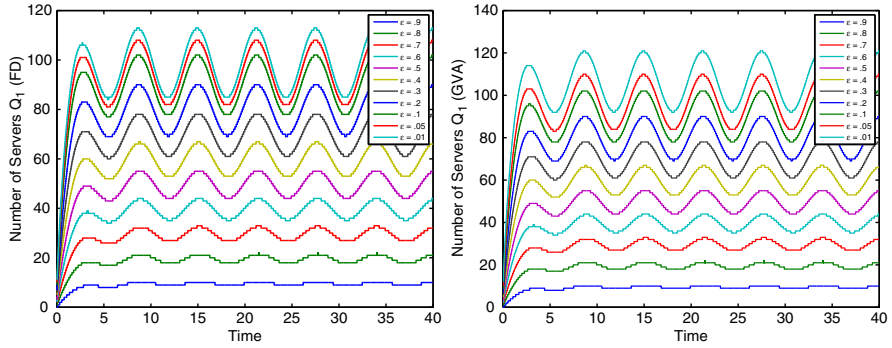

FIGURE **19.** Stable abandonment staffing schedule DMA (left). Stable abandonment staffing schedule GVA (right). $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = 0.5$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

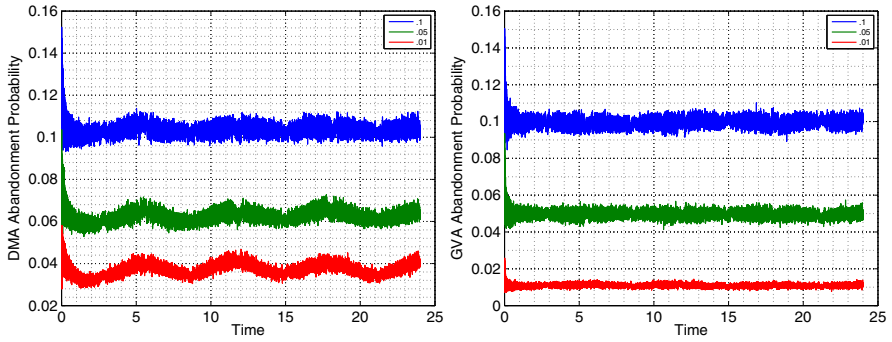

FIGURE **20.** Probability of abandonment (DMA) (left). Probability of abandonment (GVA) (right) $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = 0.5 \quad \epsilon = \{0.1, 0.05, 0.01\}$

does not stabilize the probability of abandonment very well for target values less than 0.1. On the left of Figure 20, we highlight the levels of $\{\epsilon = 0.1, 0.05, 0.1\}$ and we see that DMA is not stabilizing the abandonment probabilities well. This can be explained by Eq. (5.8) because, when we get to the target levels of $\{\epsilon = 0.1, 0.05, 0.1\}$, we must staff the queueing system with more servers than the mean of the queueing process, which has an abandonment probability of zero.

Consequently, when we are trying to stabilize at the target levels of $\{\epsilon = 0.1, 0.05, 0.1\}$ we are operating in a quality driven regime. Thus, the DMA fails because when $c \geq q$, our approximation of the probability of abandonment is equal to 0. Thus, the failure of the DMA in important quality drive regimes suggests that we might need a more refined approximation for the probability of abandonment that uses more information than just a the mean of the queue length process. We should also mention that the failure of DMA is similar to the failure of the mean offered load to stabilize abandonment probability as noted by Liu and Whitt [12] in their DIS–OL approximation. This inaccuracy motivated their DIS–MOL approximation and will motivate us refining our deterministic estimates of the queue length function in order to provide better functions for stabilizing the probability of abandonment.

5.3. Probability of Abandonment (GVA)

Similar to the DMA method, we take advantage of our fundamental observation is that when simulating a Markovian queueing system one can calculate the probability of abandonment as the following ratio of event times

$$\mathbb{P}(\text{Ab}) = \frac{\beta \cdot E[(Q-c)^+]}{\mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+]}. \tag{5.11}$$

Now we exploit the GVA approximation and substitute it into the ratio of event times as an approximation of the queue length distribution. This yields the following expression for the probability of abandonment:

$$\begin{aligned} \mathbb{P}(\text{Ab}) &\approx \frac{\beta \cdot E[(Q-c)^+]}{\mu \cdot E[Q \wedge c] + \beta \cdot E[(Q-c)^+]} \\ &= \frac{\beta}{\frac{\mu \cdot q}{\sqrt{v} \cdot (\varphi(\chi) - \chi \cdot \overline{\Phi}(\chi))} - \mu + \beta}. \end{aligned} \tag{5.12}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Moreover, since the probability of abandonment depends on the level of staffing c, we can invert it to construct a stabilizing staffing schedule for the probability of abandonment. This suggests the following procedure to find a staffing schedule that stabilizes the probability of abandonment with a target level of ϵ

$$c^* = \inf\left\{ c \geq 0, c \in \mathbb{N} \;\middle|\; \frac{\beta}{\frac{\mu \cdot q}{\sqrt{v} \cdot (\varphi(\chi) - \chi \cdot \overline{\Phi}(\chi))} - \mu + \beta} \leq \epsilon \right\}. \tag{5.13}$$

This staffing can be seen in the right of Figure 19 using GVA. The subsequent staffing for DMA is given on the left of Figure 19.

Before we move to the network case, we give a summary of the method in the one-dimensional setting to make sure the reader has a good understanding of the method using the transition rates of the queueing process.

- First approximate the probability of abandonment with the ratio of the rate of abandonments and the total number of departures from the queue. This ratio approximates the fraction of customers that will abandon at time t.
- Use DMA or GVA to estimate the expectation terms that appear in the ratio.
- Set the ratio of the rate of abandonments and the total number of departures from the queue be equal to the target value of ϵ.
- Given the mean for DMA, the mean and variance for GVA and the service and abandonment rates, we then solve for the staffing level c that achieves this ratio of ϵ, which is the solution to a fixed point equation.
- We then make the solution to the fixed point equation equal the staffing level for the next time increment and repeat this procedure until the final time point.

5.4. Extension to the Network Case

In addition to the one-dimensional Erlang-A model, we can also extend our stabilization method to Jackson networks as well. To do this, we observe that

$$\tau_t^i + \sum_{j=1}^{N} \tau_t^{ij} = 1 \quad \text{and} \quad \gamma_t^i + \sum_{j=1}^{N} \gamma_t^{ij} = 1 \tag{5.14}$$

and therefore, we have that the probability of abandonment at the ith station can be approximated as

$$\mathbb{P}(\text{Ab}) \approx \frac{\beta_i \cdot E[(Q_i - c_i)^+]}{\mu_i \cdot E[Q_i \wedge c_i] + \beta_i \cdot E[(Q_i - c_i)^+]}. \tag{5.15}$$

Now using this representation and using the DMA, we arrive at the following procedure that approximately stabilizes the probability of abandonment with a target level of ϵ_i

$$c_i^* = \inf\left\{ c_i \geq 0, c_i \in \mathbb{N} \middle| \frac{\beta_i}{\frac{\mu_i \cdot q_i}{(q_i - c_i)^+} - \mu_i + \beta_i} \leq \epsilon_i \right\}. \tag{5.16}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

Moreover, by repeating the same procedure and using GVA, we arrive at the following procedure that approximately stabilizes the probability of abandonment with a target level of ϵ_i

$$c_i^* = \inf\left\{ c_i \geq 0, c_i \in \mathbb{N} \;\middle|\; \frac{\beta_i}{\frac{\mu_i \cdot q_i}{\sqrt{v_i}\cdot\left(\varphi(\chi_i)-\chi_i\cdot\overline{\Phi}(\chi_i)\right)} - \mu_i + \beta_i} \leq \epsilon_i \right\}. \tag{5.17}$$

Now we have two different methods to stabilize the abandonment probabilities in Jackson networks. Although not presented here in the main paper, we give several numerical examples in the Appendix to illustrate that the DMA and GVA are both effective at stabilizing the abandonment probabilities at positive target levels, however, as in the one-dimensional setting, the GVA is better when the target levels of $\{\epsilon = 0.1, 0.05, 0.1\}$ are to be stabilized.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we generalize the GVA to compute the mean and covariance matrix of dynamic rate Jackson networks with abandonment. We show through our simulation experiments that the GVA method is better at approximating the moment dynamics of the queue length processes. Perhaps more important, we also show that the GVA method can be used as a staffing tool for managers of service systems. We show that by inverting our approximation for the probability of delay, we can stabilize the delay probabilities of the entire network at positive target values. Moreover, we show that the GVA method a better stabilizing algorithm than naively using the fluid and diffusion approximations.

To address the abandonment probabilities, we develop a new and novel method by using the transition rates of the queueing network to approximate the abandonment probabilities. We also invert this approximation to construct a staffing schedule to stabilize the abandonment probabilities at positive targets. We also show numerically that this procedure stabilizes the abandonment probabilities with good performance. We should also mention that this method of stabilization is applies to any Markov process as long as can compute closed form expressions for the rate functions. Therefore, it is also possible to apply our methodology to queues with state dependent abandonment like in the work of Dong, Feldman, P., and Yom-Tov [3].

There are several directions for future research. The first direction is to extend the Gaussian skewness approximation (GSA) method [17] to Jackson networks with abandonment. This requires a better understanding of the quadratic formula for multi-dimensional polynomials as well as a greater understanding of multi-dimensional Hermite polynomials and multi-dimensional Wiener chaos theory. In addition, it requires that one analyze the third cumulant moment tensor, which has not be explored in the queueing theory literature. Lastly, it is of interest to extend the GVA and GSA methods to non-Markovian Jackson networks, where the arrival, service, and abandonment distributions can be of phase-type like in the work of Pender and Ko [9,10,24] or Pender and Phung-Duc [25]. This extension would allow us to model more realistic service and abandonment distributions, which may not be exponentially distributed. We hope to explore these extensions in later papers.

Acknowledgements

The first author (J. P.) gratefully acknowledges the support of Cornell University (ORIE) and the gracious hospitality of Columbia University's IEOR Department where some of this work was written. The second author (W. A. M.) was partially supported by National Science Foundation grants DMS-0807440 and CMMI-1436334.

References

1. Cameron, R.H. & Martin, W.T. (1947). The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. *Annals of Mathematics* 385–392.
2. Defraeye, M. & Van Nieuwenhuyse, I. (2016). Staffing and scheduling under nonstationary demand for service: A literature review. *Omega* 58: 4–25.
3. Dong, J., Feldman, P., & Yom-Tov, G. (2013). Slowdown services: Staffing service systems with load-dependent service rate. Available at SSRN 2317410.
4. Engblom, S. & Pender, J. (2014). Approximations for the moments of nonstationary and state dependent birth-death queues. Arxiv preprint arXiv:1406.6164.
5. Feldman, Z., Mandelbaum, A., Massey, W.A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science* 54(2): 324–338.
6. Jennings, O.B., Mandelbaum, A., Massey, W.A., & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science* 42(10): 1383–1394.
7. Khudyakov, P., Feigin, P.D., & Mandelbaum, A. (2010). Designing a call center with an IVR (interactive voice response). *Queueing Systems* 66(3): 215–237.
8. Ko, Y.M. & Gautam, N. (2013). Critically loaded time-varying multiserver queues: computational challenges and approximations. *INFORMS Journal on Computing* 25(2): 285–301.
9. Ko, Y.M. & Pender, J. Diffusion limits for the (map(t)/ph(t)/) n queueing network.
10. Ko, Y.M. & Pender, J. Strong approximations for time varying infinite-server queues with non-renewal arrival and service processes.
11. Liu, Y. & Whitt, W. Stabilizing performance in a service system with time-varying arrivals and customer feedback. Technical Report., Working paper.
12. Liu, Y. & Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* 60(6): 1551–1564.
13. Liu, Y. & Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 28(04): 419–449.
14. Mandelbaum, A., Massey, W.A., & Reiman, M.I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* 30(1–2): 149–201.
15. Mandelbaum, A., Massey, W.A., Reiman, M.I., Stolyar, A., & Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems* 21 (2–4): 149–171.
16. Massey, W. & Pender, J. (2011). Skewness variance approximation for dynamic rate multi-server queues with abandonment. *Performance Evaluation Review* 39: 74–74.
17. Massey, W. & Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* 75(2): 243–277.
18. Massey, W.A. & Whitt, W. (1993). Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems* 13(1–3): 183–250.
19. Pender, J. Time varying queues with abandonment via Laguerre polynomial expansions. Technical report., Cornell University, Cornell University.
20. Pender, J. (2014). Gram charlier expansion for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics* 74(4): 1238–1265.
21. Pender, J. (2014). A Poisson–Charlier approximation for nonstationary queues. *Operations Research Letters* 42(4): 293–298.
22. Pender, J. (2016). An analysis of nonstationary coupled queues. *Telecommunication Systems* 61(4): 823–838.
23. Pender, J. (2015). Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Informational Sciences* 29(01): 27–49.
24. Pender, J. & Ko, Y.M. Approximations for the queue length distributions of time-varying many-server queues.
25. Pender, J. & Phung-Duc, T. (2016). A law of large numbers for M/M/c/delayo-setup queues with nonstationary arrivals. In *Analytical and Stochastic Modelling Techniques and Applications: 23rd International Conference Proceedings*, Cardiff, UK, 24–26 August, 2016, vol. 9845, Springer.
26. Ross, S.M. (2006). *Simulation*. Amsterdam: Elsevier Academic Press.
27. Stein, C. (1986). Approximate computation of expectations. *Lecture Notes-Monograph Series* 7: i–164.
28. Véricourt, F.d. & Jennings, O.B. (2011). Nurse staffing in medical units: A queueing perspective. *Operations Research* 59(6): 1320–1331.
29. Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics* 60(4): 897–936.
30. Xiu, D., & Karniadakis, G.E. (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing* 24(2): 619–644.
31. Yom-Tov, G.B. & Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* 16(2): 283–299.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

APPENDIX

This Appendix has two main purposes. The first is to extend the DMA and GVA methods to the case when the queueing network is a loss network and customers are rejected from the queue is there is not enough space for them to join. The second purpose of the Appendix is to provide readers with ample numerical examples of networks to illustrate that the GVA has not only the ability to estimate the mean and covariance matrix of the Jackson network, but also to illustrate that the GVA is also useful for stabilizing the probability of delay and abandonment probabilities in various parameter settings. We begin first with the extension to non-stationary loss networks with customer abandonment.

A.1. Proof of Main Approximation

PROOF: Now we begin to prove the result. We first begin with the expectation of the max function. In order to compute the max function we simply use Lemma 3.5 along with the properties of expectations.

$$\begin{aligned}
E[(Q_i - c_i)^+] &= E\left[(q_i + \sqrt{v_i} \cdot Z_i - c_i)^+\right] \\
&= E\left[\left(q_i + \sqrt{v_i} \sum_{k=1}^{i} a_k \cdot X_k - c_i\right)^+\right] \\
&= \sqrt{v_i} \cdot E\left[\left(\sum_{k=1}^{i} a_k \cdot X_k - \chi_i\right)^+\right] \\
&= \sqrt{v_i} \cdot E\left[(X - \chi_i)^+\right] \\
&= \sqrt{v_i} \cdot \left(\phi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right),
\end{aligned}$$

where the a_k variables represent the value of the iterated cosines from the Z_k random variable and the squared sum of the a_k variables always sums to one. In order to compute the expectation of the min function, we use the following identity.

$$(Q_i \wedge c_i) = Q_i - (Q_i - c_i)^+. \tag{A.1}$$

Thus, the expectation of the min function has the following closed form expression under the GVA

$$\begin{aligned}
E[(Q_i \wedge c_i)] &= E\left[Q_i - (Q_i - c_i)^+\right] \\
&= E[Q_i] - E\left[(Q_i - c_i)^+\right] \\
&= q_i - \sqrt{v_i} \cdot \left(\phi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right).
\end{aligned}$$

Now that we have computed all of the expectation terms, it is necessary to compute the covariance terms. We compute the covariance terms in a similar manner, but leverage a multi-dimensional version of the Stein's lemma that is seen in [27]. For the covariance of the max function and the queue length process we have that

$$\begin{aligned}
\mathrm{Cov}[Q_j, (Q_i - c_i)^+] &= \mathrm{Cov}\left[q_j + \sqrt{v_j} \cdot Z_j, (Q_i - c_i)^+\right] \\
&= \mathrm{Cov}\left[q_j + \sqrt{v_j} \cdot Z_j, (q_i + \sqrt{v_i} \cdot Z_i - c_i)^+\right] \\
&= \mathrm{Cov}\left[q_j + \sqrt{v_j} \sum_{k=1}^{j} a_k \cdot X_k, \left(q_i + \sqrt{v_i} \sum_{k=1}^{i} \tilde{a}_k \cdot X_k - c_i\right)^+\right]
\end{aligned}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

$$
\begin{aligned}
&= \mathrm{Cov}\left[\sqrt{v_j}\sum_{k=1}^{j} a_k \cdot X_k, \left(q_i + \sqrt{v_i}\sum_{k=1}^{i} \tilde{a}_k \cdot X_k - c_i\right)^+\right] \\
&= \sqrt{v_j \cdot v_i} \cdot \sum_{k=1}^{i} a_k \cdot \mathrm{Cov}\left[X_k, \left(\sum_{m=1}^{i} \tilde{a}_m \cdot X_m - \chi_i\right)^+\right] \\
&= \sqrt{v_j \cdot v_i} \cdot \sum_{k=1}^{i} a_k \cdot \tilde{a}_k \cdot E\left[\left\{\sum_{m=1}^{i} a_m \cdot X_m > \chi_i\right\}\right] \\
&= \sqrt{v_j \cdot v_i} \cdot \sum_{k=1}^{i} a_k \cdot \tilde{a}_k \cdot \overline{\Phi}(\chi_i) \\
&= \sqrt{v_j \cdot v_i} \cdot \overline{\Phi}(\chi_i) \cdot \sum_{k=1}^{i} a_k \cdot \tilde{a}_k \\
&= \sqrt{v_j \cdot v_i} \cdot \overline{\Phi}(\chi_i) \cdot \cos\theta_j \ldots \cos\theta_{i-1} \\
&= v_{ji} \cdot \overline{\Phi}(\chi_i).
\end{aligned}
$$

A similar result also holds for the covariance of the min function and the queue length process using the same identity of Eq. (A.1).

$$
\begin{aligned}
\mathrm{Cov}\left[Q_j, (Q_i \wedge c_i)\right] &= \mathrm{Cov}\left[Q_j, Q_i - (Q_i - c_i)^+\right] \\
&= \mathrm{Cov}\left[Q_j, Q_i\right] - \mathrm{Cov}\left[Q_j, (Q_i - c_i)^+\right] \\
&= v_{ji} - v_{ji} \cdot \overline{\Phi}(\chi_i) \\
&= v_{ij} \cdot \Phi(\chi_i).
\end{aligned}
$$

■

A.2. Non-stationary Loss Networks

This extension is slightly non-trivial since the rejection of customers is involves additional indicator functions. We show that by using a simple independence assumption for the indicator functions, we are able to use the same theory to approximate these loss networks.

A.2.1. Functional Forward Equations for Loss Jackson Networks.

$$
\begin{aligned}
\dot{E}\left[f(Q_i)\right] = {} & E\left[\alpha_i(Q) \cdot (f(Q + e_i) - f(Q))\right] + E\left[\delta_i(Q) \cdot (f(Q - e_i) - f(Q)\right] \\
& + \sum_{j=1}^{N} E\left[D_{ij}(Q) \cdot \big(f(Q - e_i + e_j) - f(Q)\big)\right] \\
& + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\left(f(Q + e_i - e_j) - f(Q)\right)\right]
\end{aligned}
\tag{A.2}
$$

for all appropriate real-valued functions $f : \mathbb{R}^n \to \mathbb{R}$ and where we have the following expressions for the rate functions

$$\alpha_i(Q) = \lambda_i \cdot \{Q_i < c_i + k_i\}, \tag{A.3}$$

$$\delta_i(Q) = \mu_i \cdot (Q_i \wedge c_i) + \beta_i \cdot (Q_i - c_i)^+, \tag{A.4}$$

$$\tilde{D}_{ij}(Q) = \mu_i \cdot \tau_{ij} \cdot (Q_i \wedge c_i) \cdot \{Q_j < c_j + k_j\} + \beta_i \cdot \gamma_{ij} \cdot (Q_i - c_i)^+ \cdot \{Q_j < c_j + k_j\}, \tag{A.5}$$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

and

$$D_{ij}(Q) = \mu_j \cdot \tau_{ji} \cdot (Q_j \wedge c_j) \cdot \{Q_i < c_i + k_i\} + \beta_j \cdot \gamma_{ji} \cdot (Q_j - c_j)^+ \cdot \{Q_i < c_i + k_i\}. \quad \textbf{(A.6)}$$

This implies that we have the following expressions for the mean, variance, and covariance matrix of the loss network

$$\dot{E}[Q_i] = E[\alpha_i(Q)] - E[\delta_i(Q)] - \sum_{j=1}^{N} E\left[D_{ij}(Q)\right] + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right], \quad \textbf{(A.7)}$$

$$\begin{aligned}\dot{\mathrm{Var}}[Q_i] &= E[\alpha_i(Q)] + E[\delta_i(Q)] + \sum_{j=1}^{N} E\left[D_{ij}(Q)\right] + \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right] \\ &\quad + 2 \cdot \mathrm{Cov}[Q_i, \alpha_i(Q)] - 2 \cdot \mathrm{Cov}[Q_i, \delta_i(Q)] \\ &\quad - 2 \cdot \sum_{j=1}^{N} \mathrm{Cov}[Q_i, \tilde{D}_{ij}(Q)] + 2 \cdot \sum_{j=1}^{N} \mathrm{Cov}[Q_i, D_{ij}(Q)],\end{aligned} \quad \textbf{(A.8)}$$

and

$$\begin{aligned}\dot{\mathrm{Cov}}[Q_i, Q_l] &= -\sum_{j=1}^{N} E\left[D_{ij}(Q)\right] - \sum_{j=1}^{N} E\left[\tilde{D}_{ij}(Q)\right] \\ &\quad + \mathrm{Cov}[Q_l, \alpha_i(Q)] + \mathrm{Cov}[Q_i, \alpha_l(Q)] - \mathrm{Cov}[Q_i, \delta_l(Q)] - \mathrm{Cov}[Q_l, \delta_i(Q)] \\ &\quad + \sum_{j=1}^{N} \mathrm{Cov}[Q_i, D_{ij}(Q) - \tilde{D}_{ij}(Q)] - \sum_{j=1}^{N} \mathrm{Cov}[Q_l, D_{ij}(Q) - \tilde{D}_{ij}(Q)].\end{aligned} \quad \textbf{(A.9)}$$

A.2.2. DMA for Loss Networks. Now that we have an understanding of the mean and variance of the functional forward equations for the loss network, we can now use DMA to approximate the mean. This approximation leads us to the following theorem.

THEOREM A.1: *If we substitute the approximate distribution $Q \equiv q$ for the queue length process of the non-stationary loss network with abandonment, we get the following differential equations for the mean:*

$$\begin{aligned}\dot{q}_i &= \lambda_i \cdot \{q_i < c_i + k_i\} - \mu_i \cdot \gamma_i \cdot (q_i \wedge c_i) - \beta_i \cdot \tau_i \cdot (q_i - c_i)^+ \\ &\quad - \sum_{j=1}^{N} (q_i - c_i)^+ \cdot \beta_i \cdot \tau_{ij} - \sum_{j=1}^{N} (q_i \wedge c_i) \cdot \mu_i \cdot \gamma_{ij} \\ &\quad + \sum_{j=1}^{N} (q_j - c_j)^+ \cdot \beta_j \cdot \tau_{ji} + \sum_{j=1}^{N} (q_j \wedge c_j) \cdot \mu_j \cdot \gamma_{ji}.\end{aligned} \quad \textbf{(A.10)}$$

PROOF: This follows directly from the proof of the Jackson network case. ■

A.2.3. GVA for Loss Networks. In addition, to the DMA, we can also use the GVA to approximate the mean and covariance matrix of the loss network. This approximation leads us to the following theorem.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

THEOREM A.2: *If we substitute the above approximate distribution in Eq. (3.8) for each of the queue length processes of the non-stationary loss network with abandonment, then we have the following closed form expressions for the rate functions of the functional forward equations*

$$E[\alpha_i(Q)] = \lambda_i \cdot \Phi(\psi_i), \tag{A.11}$$

$$E[\delta_i(Q)] = \mu_i \cdot \tau_i \cdot \left(q_i - \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right) + \beta_i \cdot \gamma_i \cdot \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right), \tag{A.12}$$

$$\begin{aligned} E[D_{ij}(Q)] \approx{}& \mu_i \cdot \tau_{ij} \cdot \left(q_i - \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right)\right) \cdot \Phi(\psi_j) \\ &+ \beta_i \cdot \gamma_{ij} \cdot \sqrt{v_i} \cdot \left(\varphi(\chi_i) - \chi_i \cdot \overline{\Phi}(\chi_i)\right) \cdot \Phi(\psi_j), \end{aligned} \tag{A.13}$$

$$\begin{aligned} E[\tilde{D}_{ij}(Q)] \approx{}& \mu_j \cdot \tau_{ji} \cdot \left(q_j - \sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right)\right) \cdot \Phi(\psi_i) \\ &+ \beta_j \cdot \gamma_{ji} \cdot \sqrt{v_j} \cdot \left(\varphi(\chi_j) - \chi_j \cdot \overline{\Phi}(\chi_j)\right) \cdot \Phi(\psi_i), \end{aligned} \tag{A.14}$$

$$\mathrm{Cov}[Q_l, \alpha_i(Q)] = -\frac{v_{il}}{\sqrt{v_i}} \cdot \varphi(\psi_i), \tag{A.15}$$

$$\mathrm{Cov}[Q_l, \delta_i(Q)] = \mu_i \cdot \tau_i \cdot v_{li} \cdot \Phi(\chi_i) + \beta_i \cdot \gamma_i \cdot v_{li} \cdot \overline{\Phi}(\chi_i), \tag{A.16}$$

$$\mathrm{Cov}[Q_l, D_{ij}(Q)] \approx \mu_i \cdot \tau_{ij} \cdot v_{li} \cdot \Phi(\chi_i) \cdot \Phi(\psi_j) + \beta_i \cdot \gamma_{ij} \cdot v_{li} \cdot \overline{\Phi}(\chi_i) \cdot \Phi(\psi_j), \tag{A.17}$$

$$\mathrm{Cov}[Q_l, \tilde{D}_{ij}(Q)] \approx \mu_j \cdot \tau_{ji} \cdot v_{lj} \cdot \Phi(\chi_j) \cdot \Phi(\psi_i) + \beta_j \cdot \gamma_{ji} \cdot v_{lj} \cdot \overline{\Phi}(\chi_j) \cdot \Phi(\psi_i), \tag{A.18}$$

and where we now define

$$\chi_i = \frac{c_i - q_i}{\sqrt{v_i}} \quad \textit{and} \quad \psi_i = \frac{c_i + k_i - q_i}{\sqrt{v_i}}. \tag{A.19}$$

PROOF: This also follows directly from the proof of the Jackson network case and the independence assumption of the indicator functions. ■

Remark A.3: The reason why we have the symbol $\approx$ instead of $=$ for Eqs. (A.13), (A.14), (A.17), and (A.18) is because this is exactly where we take use the independence assumption on pairwise stations. If we did not use this assumption, our approximations would involve infinite Hermite polynomial series expansions for N-dimensional Gaussian densities, which is very complicated, see for example [22].

A.3. Additional Numerics for Approximating Moments

In this section, we also provide more numerical support for using the GVA method. Our goal here is not to give more examples of stabilizing the network, but to provide more examples of the mean, variance, and probability of delay in different network settings. It is our hope that these examples will show that our method is good at estimating various performance measures are of importance to the service systems community.

A.3.1. High Arrival Rate Example. In our first example, we consider a two node Jackson network with the following parameters given in Table A.1. We choose to simulated a queueing system that is closer to the limiting Halfin–Whitt regime with a large number servers and a large arrival rate. In Figure A.1, we see the GVA method is better approximating the mean queue length for Q_1 and Q_2. We also see that the GVA method does a better job of approximating the variance of the queueing processes when compared to the fluid and diffusion limits. On the bottom of Figure A.1, we compare the delay probabilities with their simulated counterparts. We observe that the GVA method is also approximating the probability of delay dynamics quite well.

A.3.2. Discontinuous Arrival Rate Example. In our second example, we consider a two node Jackson network with the following parameters given in Table A.2. Instead of using sinusoidal arrival dynamics, we consider a piecewise constant arrival process. In all aspects of the approximations,

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

TABLE **A.1.** Two Node Jackson Network Model Parameters High Arrival Rate

Parameter	Value	Parameter	Value
λ_1	$50 + 25\sin(t)$	λ_2	$50 + 10\sin(t)$
μ_1	1	μ_2	1
β_1	0.25	β_2	0.5
c_1	50	c_2	100
τ_{11}	0.25	τ_{22}	0
γ_{11}	0.25	γ_{22}	0
τ_{12}	0.75	τ_{21}	0
γ_{12}	0.75	γ_{21}	0

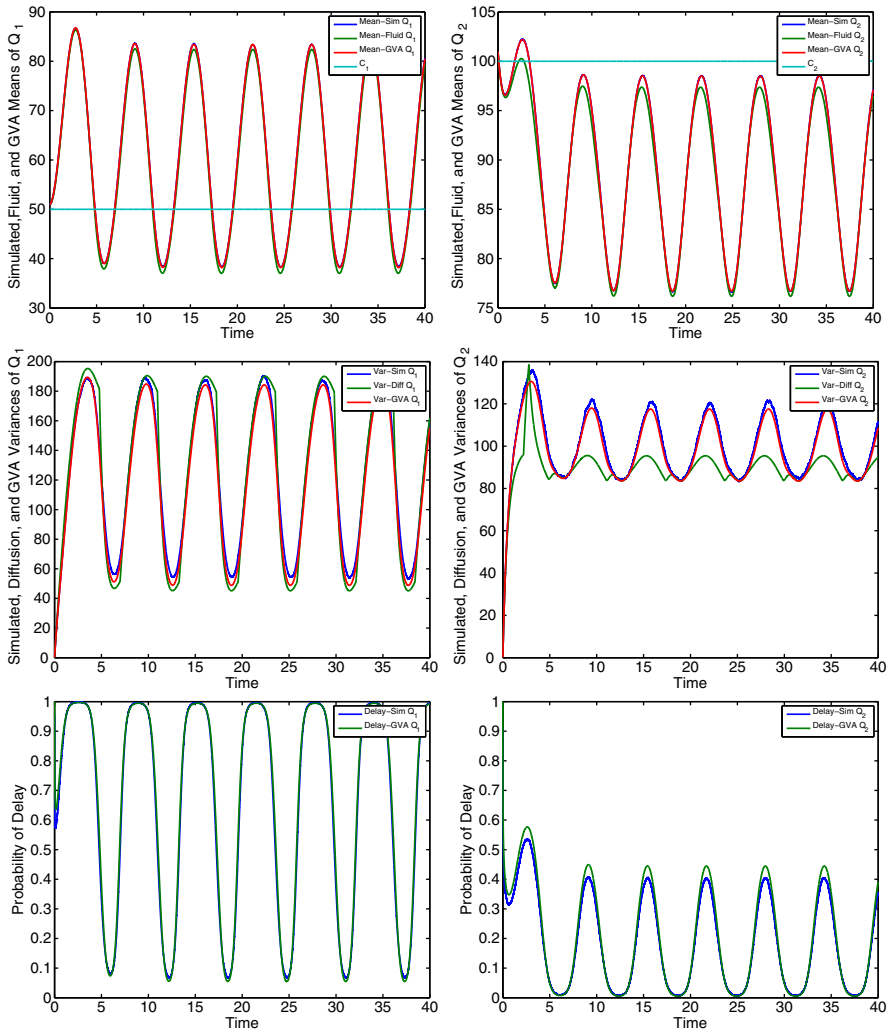

FIGURE **A.1.** Mean of Q_1 (top left). Mean of Q_2 (top right). Variance of Q_1 (middle left). Variance of Q_2 (middle right). Probability of delay of Q_1 (bottom left). Probability of delay of Q_2 (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

TABLE **A.2.** Two Node Jackson Network Model Parameters Discontinuous Arrival Rate

Parameter	Value	Parameter	Value
λ_1	$20 + 10\sin(t)$	λ_2	$30 + 15\sin(t)$
μ_1	1	μ_2	1
β_1	0.5	β_2	0.5
μ_{12}	1	μ_{21}	1
β_{12}	0.5	β_{21}	0.5
τ_{11}	0.25	τ_{22}	0.5
γ_{11}	0.25	γ_{22}	0.5
τ_{12}	0.75	τ_{21}	0.5
γ_{12}	0.75	γ_{21}	0.5

GVA seems to do quite well. The good performance of the GVA is best seen for the variance of the two queueing processes seen in the middle two graphs of Figure A.2.

A.3.3. Dynamic Staffing Example. In our third additional numerical example for the moment approximations, we consider a two node Jackson network with the following parameters given in Table A.3. Instead of using a constant number of servers, we consider deterministically changing the number servers according to the arrival process. In all aspects of the approximations, GVA seems to do quite well. The good performance of the GVA is best seen for the variance of the two queueing processes seen in the middle two graphs of Figure A.3.

A.3.4. All Parameters Non-stationary Example In our last numerical example for the moment approximations, we consider a two node Jackson network with the following parameters given in Table A.4. We present this numerical example since we want to present an example of when all of the parameters are time varying and can impact the queue length process. Once again, we see that in all aspects of the approximations that GVA seems to do quite well. The good performance of the GVA is best seen for the variance of the two queueing processes seen in the middle two graphs of Figure A.4.

A.4. Additional Numerics for Stabilizing Delay Probabilities

A.4.1. Impatient Customers Delay Example. In this section, we provide additional numerical examples for stabilizing the delay probabilities in the network setting. This time instead of the service rate and the abandonment rate being identical as in Figure 4 of Feldman et al. [5], we have that $\mu_1 = \mu_2 = 1$ and $\beta_1 = \beta_2 = 2$. This implies that customers in each queue are relatively *impatient* relative to the service time as they are not willing to wait longer than the mean service time. Moreover, in this example we see in Figure A.4 that the GVA method does a good job of stabilizing at the target delay values. In fact, this performance is much better than the fluid and diffusion limits. This can be explained since the GVA is better at approximating the variance behavior in all of the cases. We also compare the two methods when the probability of delay is equal to 0.5, which is the most important case since the queueing process is critically loaded. We also see at the bottom of Figure A.4 that the GVA method is outperforming the fluid and diffusion algorithm. Thus, we see that the GVA method is better at stabilizing the queueing process at target values in the *impatient* customer case.

A.4.2. Patient Customers Delay Example. The last numerical example that we present for the stabilizing the delay probabilities is similar to the previous example except that we set $\mu_1 = \mu_2 = 1$ and $\beta_1 = \beta_2 = 0.5$. This implies that customers in each queue are relatively *patient* relative to the service time as they are willing to wait longer than the mean service time. Lastly,

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

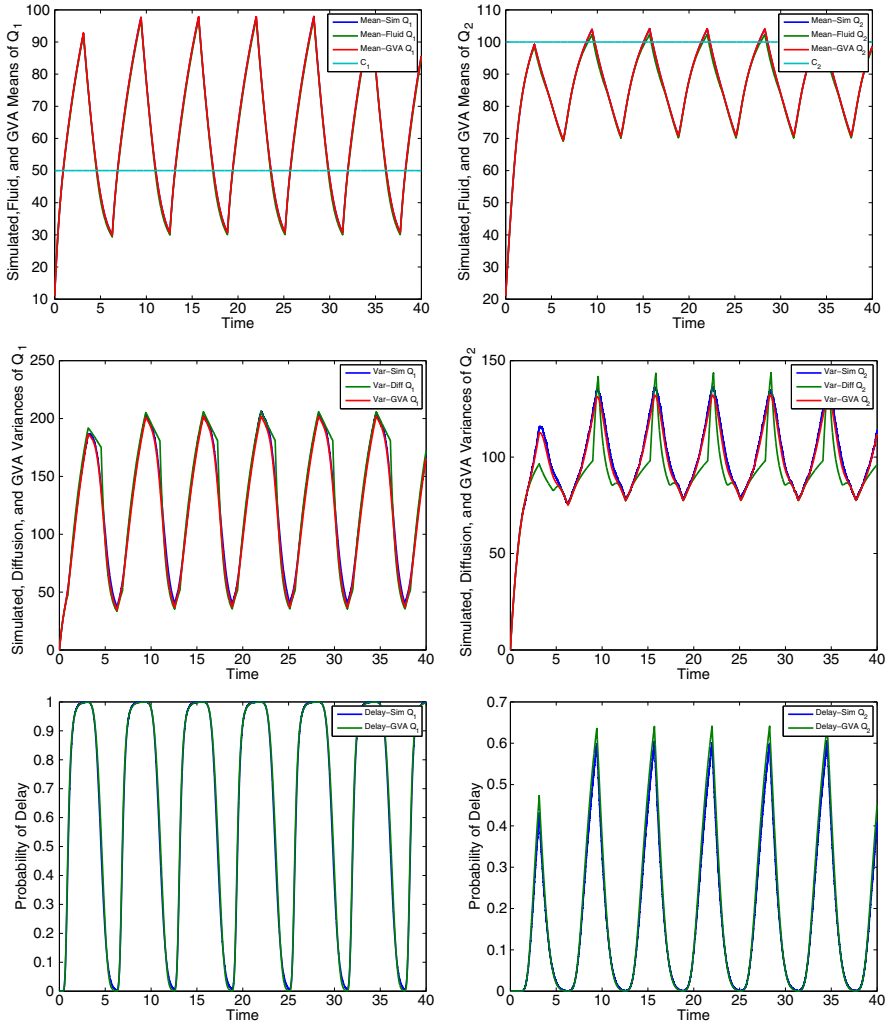

FIGURE **A.2.** Mean of Q_1 (top left). Mean of Q_2 (top right). Variance of Q_1 (middle left). Variance of Q_2 (middle right). Probability of Delay of Q_1 (bottom left). Probability of delay of Q_2 (right).

in this example where the service rate is greater than the abandonment rate, we see in Figure A.6 that the GVA method does a good job of stabilizing at the target delay values. Once again the performance of the GVA is much better than the fluid and diffusion limits. This can be explained since the GVA is better at approximating the variance behavior in all of the cases. Again, we also compare the two methods when the probability of delay is equal to 0.5. At the bottom of Figure A.6 we see that the GVA method is out performing the fluid and diffusion algorithm at this point. Thus, we see that the GVA method is better at stabilizing the queueing process at target values in the *patient* customer case and thus in all possible cases of performance.

A.5. Stabilizing Abandonment Probabilities

A.5.1. Impatient Customers Example. In this section, we provide additional numerical examples for stabilizing the abandonment probabilities in the network setting. This time instead of

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

TABLE **A.3.** Two Node Jackson Network Model Parameters Dynamic Staffing

Parameter	Value	Parameter	Value
λ_1	$50 + 25\sin(t)$	λ_2	$50 + 10\sin(t)$
μ_1	1	μ_2	1
β_1	0.25	β_2	0.5
c_1	$\lceil 1.5 \cdot \lambda \rceil$	c_2	$\lceil 1.5 \cdot \lambda \rceil$
τ_{11}	0.25	τ_{22}	0
γ_{11}	0.25	γ_{22}	0
τ_{12}	0.75	τ_{21}	0
γ_{12}	0.75	γ_{21}	0

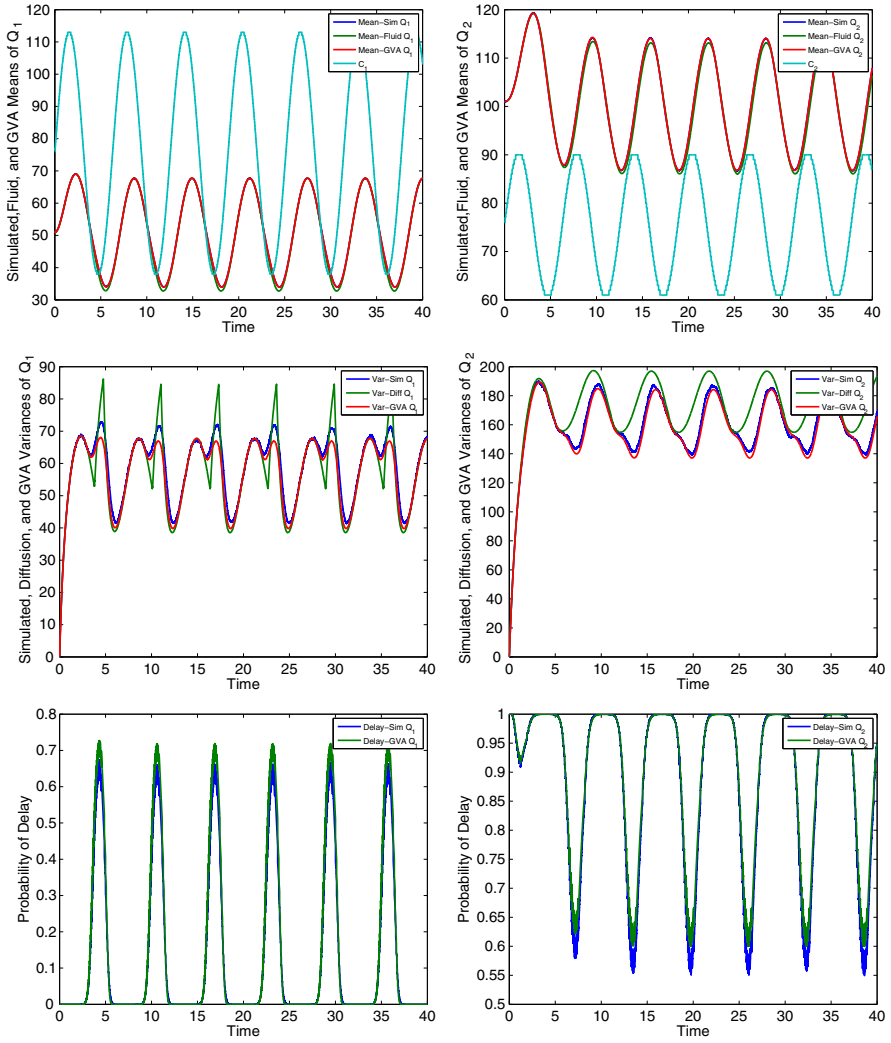

FIGURE **A.3.** Mean of Q_1 (top left). Mean of Q_2 (top right). Variance of Q_1 (middle left). Variance of Q_2 (middle right). Probability of delay of Q_1 (bottom left). Probability of delay of Q_2 (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

TABLE **A.4.** Two Node Jackson Network Model Parameters All Dynamic Parameters

Parameter	Value	Parameter	Value
λ_1	$20 + 5\sin(t)$	λ_2	$10 + 2\sin(t)$
μ_1	$1 + 0.2\sin(t)$	μ_2	$1 + 0.2\sin(t)$
β_1	$0.25 + 0.1\sin(t)$	β_2	$0.5 + 0.1\sin(t)$
c_1	$50 + 25\sin(t)$	c_2	$50 + 25\sin(t)$
τ_{11}	$0.25 - 0.1\sin(t)$	τ_{22}	0
γ_{11}	$0.25 - 0.1\sin(t)$	γ_{22}	0
τ_{12}	$0.75 + 0.1\sin(t)$	τ_{21}	0
γ_{12}	$0.75 + 0.1\sin(t)$	γ_{21}	0

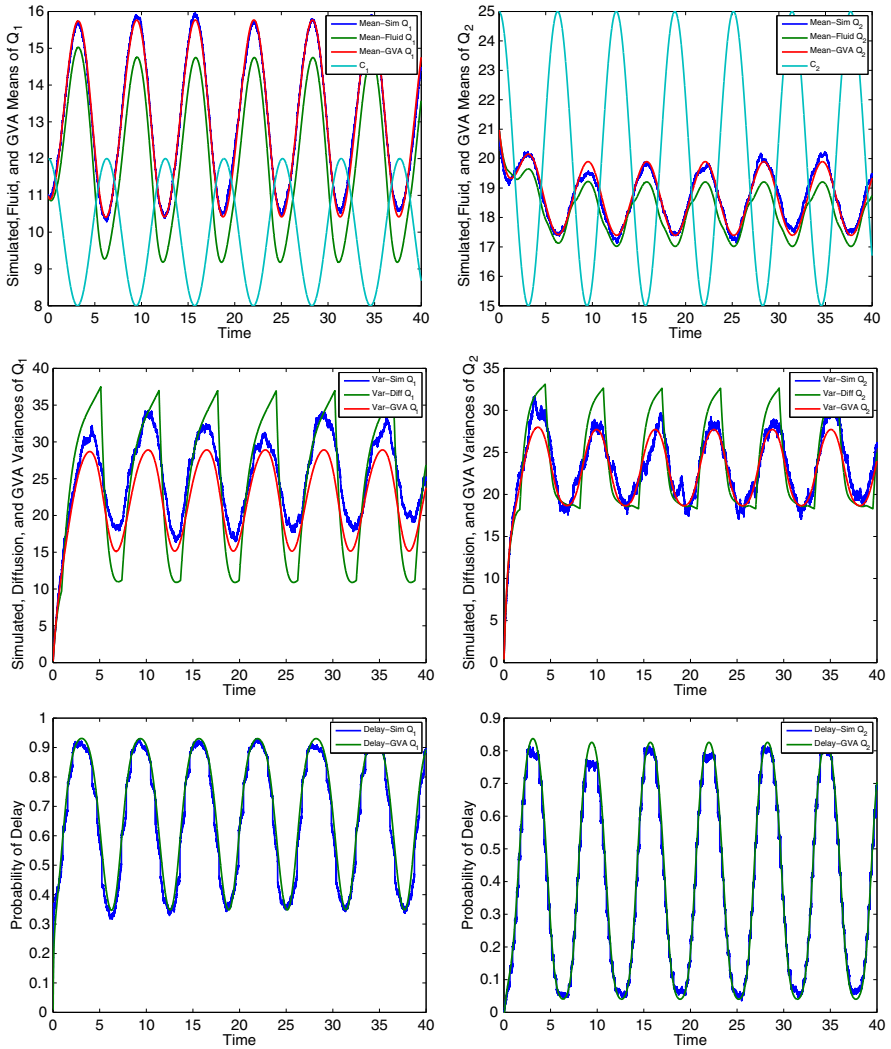

FIGURE **A.4.** Mean of Q_1 (top left). Mean of Q_2 (top right). Variance of Q_1 (middle left). Variance of Q_2 (middle right). Probability of delay of Q_1 (bottom left). Probability of delay of Q_2 (right).

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

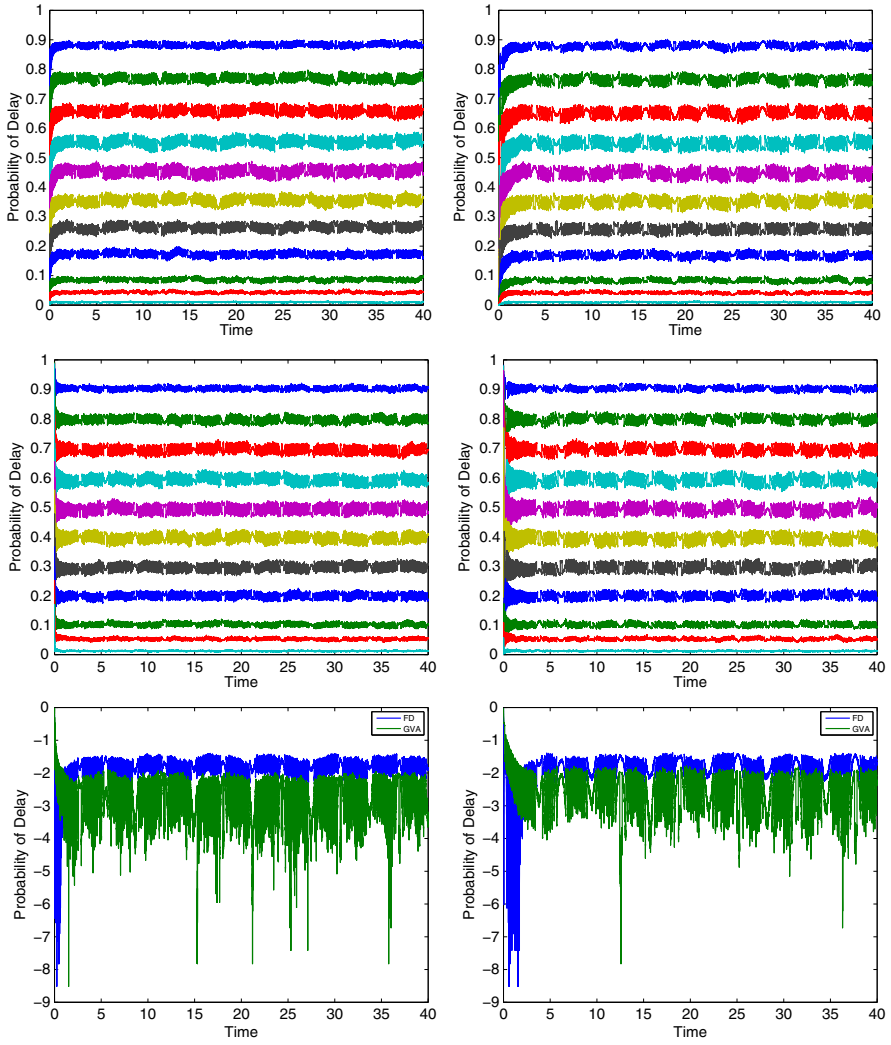

FIGURE **A.5.** Delay probabilities Q_1 FD (top left). Delay probabilities Q_2 FD (top right). Delay probabilities Q_1 GVA (middle left). Delay probabilities Q_2 GVA (middle right). Relative error Q_1 (bottom left). Relative error Q_2 (bottom right).

the service rate and the abandonment rate being identical as in Figure 4 of Feldman et al. [5], we have that $\mu_1 = \mu_2 = 1$ and $\beta_1 = \beta_2 = 2$. This implies that customers in each queue are relatively *impatient* relative to the service time as they are not willing to wait longer than the mean service time. In Figure A.7, we observe that GVA is better able to stabilize the abandonment probabilities. This is especially true for abandonment probabilities that are less than or equal to 0.1. Thus, we conclude that the GVA method is better at stabilizing the queueing process at target values in the *impatient* customer case in this example.

A.5.2. Patient Customers Example. In this section, we provide additional numerical examples for stabilizing the abandonment probabilities in the network setting. This time instead of the service rate and the abandonment rate being identical as in Figure 4 of Feldman et al. [5], we have that $\mu_1 = \mu_2 = 1$ and $\beta_1 = \beta_2 = 0.5$. This implies that customers in each queue are relatively

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

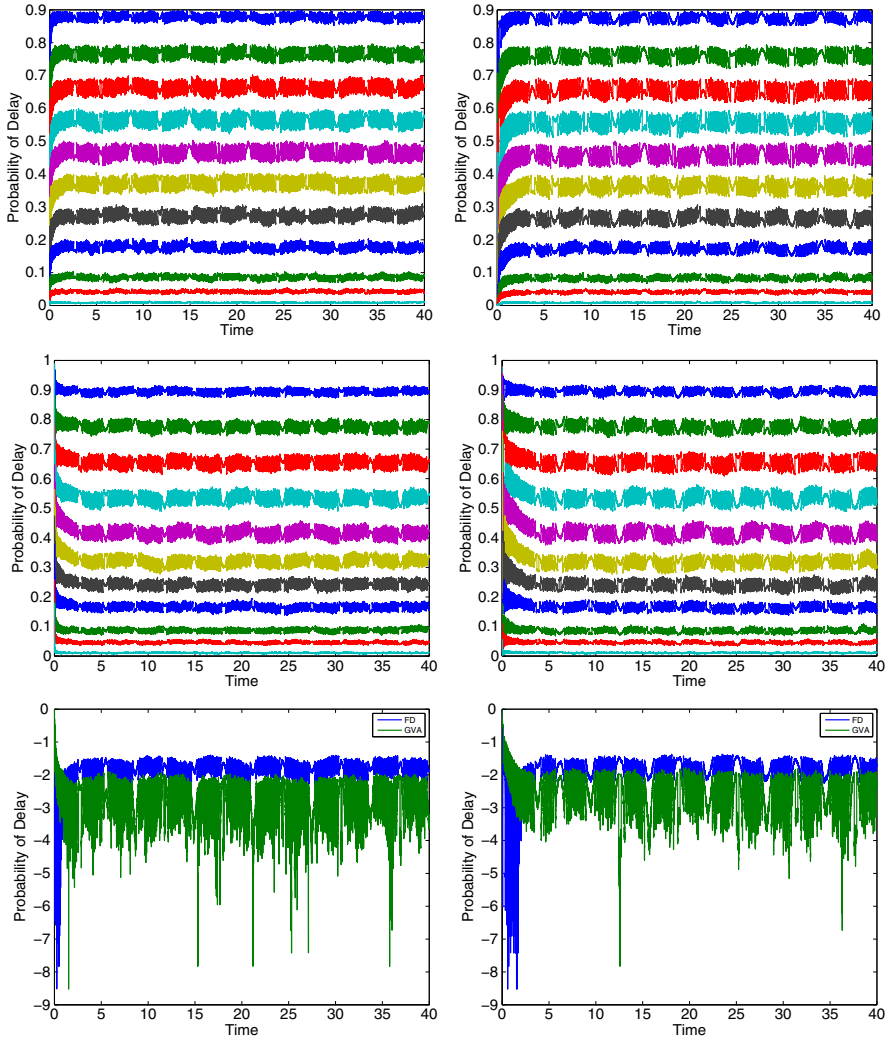

FIGURE **A.6.** Delay probabilities Q_1 FD (top left). DElay probabilities Q_2 FD (top right). Delay probabilities Q_1 GVA (middle left). Delay probabilities Q_2 GVA (middle right). Relative error Q_1 (bottom left). Relative error Q_2 (bottom right).

patient relative to the service time as they are willing to wait longer than the mean service time. In Figure A.8, we observe that the GVA is better able to stabilize the abandonment probabilities. This is especially true for abandonment probabilities that are less than or equal to 0.1. Thus, we conclude that the GVA method is better at stabilizing the queueing process at target values in the *patient* customer case in this example.

A.5.3. Non-stationary Abandonment Rate Example. In Figure A.9, we also stabilize the abandonment probabilities even when the abandonment rate is a function of time. In the example for Figure A.9, instead of having $\beta(t) = 1$ for all time, we set $\beta(t) = 1 + 0.5 \cdot \sin(t)$. In Figure A.9, it is apparent that our method also works well when the abandonment rate is also a function of time. Thus, we have confidence that our method works well when any of the model parameters are functions of time, which has not be explored in the current literature.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

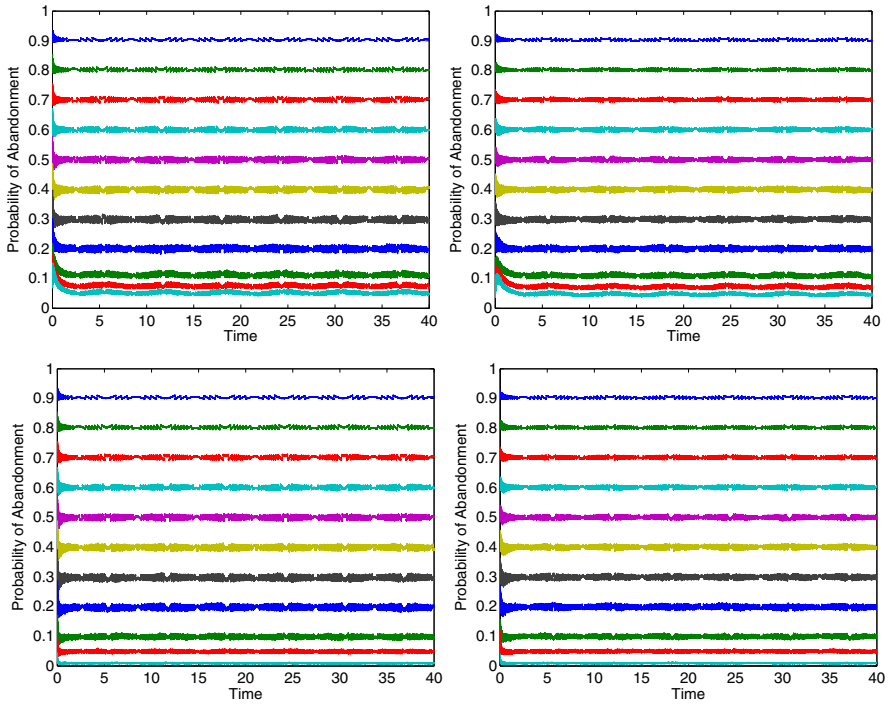

FIGURE **A.7.** Stable abandonment probabilities using DMA (left). Stable abandonment probabilities using GVA (right). $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = 0.5$

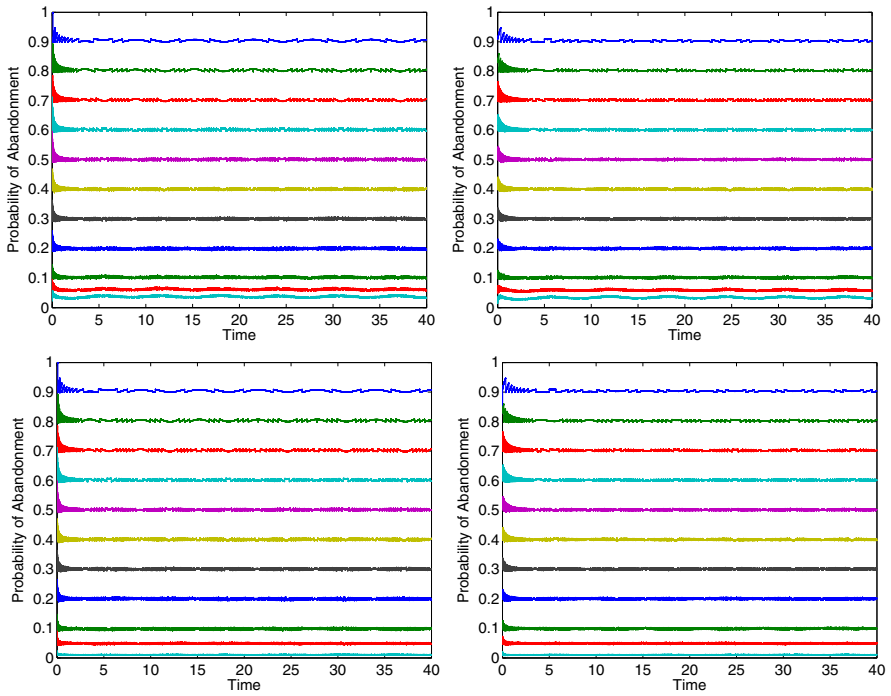

FIGURE **A.8.** Stable abandonment probabilities using DMA (left). Stable abandonment probabilities using GVA (right). $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = 0.5$.

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press

TABLE **A.5.** Two Node Jackson Feldman Extension

Parameter	Value	Parameter	Value
λ_1	$100 + 20\sin(t)$	λ_2	0
μ_1	1	μ_2	1
β_1	1	β_2	1
c_1	50	c_2	100
τ_{11}	0	τ_{22}	0
γ_{11}	0	γ_{22}	0
τ_{12}	1	τ_{21}	0
γ_{12}	1	γ_{21}	0

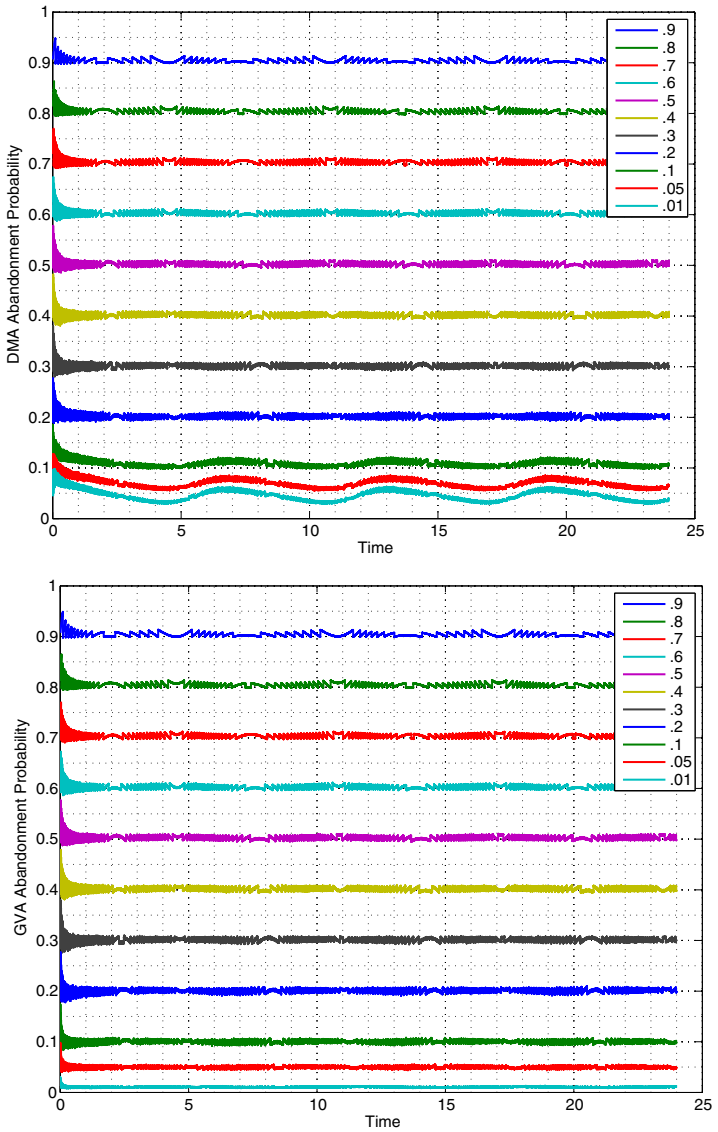

FIGURE **A.9.** Probability of abandonment (DMA) $\epsilon = 0.05$ (left), $\epsilon = 0.01$ (right) $\lambda(t) = 100 + 20 \cdot \sin(t)$, $\mu = 1$, $\beta = 1 + 0.5 \cdot \sin(t)$

https://doi.org/10.1017/S0269964816000371 Published online by Cambridge University Press