

A TECHNIQUE FOR COMPUTING SOJOURN TIMES IN LARGE NETWORKS OF INTERACTING QUEUES

V. ANANTHARAM AND M. BENCHEKROUN

*School of Electrical Engineering
Cornell University
Ithaca, New York 14850*

Consider a large number of interacting queues with symmetric interactions. In the asymptotic limit, the interactions between any fixed finite subcollection become negligible, and the overall effect of interactions can be replaced by an empirical rate. The evolution of each queue is given by a time inhomogeneous Markov process. This may be considered a technique for writing *dynamic* Erlang fixed-point equations. We explore this as a tool to approximate sojourn time distributions.

1. INTRODUCTION

In this paper we discuss the use of a statistical mechanics technique called *propagation of chaos* for computing the sojourn times of typical customers in large networks of interacting queues with symmetry. Propagation of chaos is a property enjoyed by certain statistical mechanical models of interacting particles [8]. Consider a system of n identical interacting particles, and focus attention on the first p of them.

Suppose that as the number of particles increases to infinity the initial distribution of the first p particles becomes asymptotically independent and the empirical distribution of particle states approaches a limit. In a model having

Research supported by the NSF under NCR 8710840, PYI award NCR 8857731, and IRI 9005849, by IBM under a Faculty Development award, by an AT&T Foundation award, and by Bellcore Inc.

the propagation of chaos property, one can write down a time inhomogeneous Markov chain, called the *one-particle chain*, such that the evolution of any given particle is described by this chain, started from the appropriate initial distribution. The rates appearing in this chain consist of rates associated with autonomous changes of the state of the particle and rates that represent the aggregate effects of interactions of the tagged particle with the large number of other particles. Further, the distinguished finite collection of particles will evolve independently, each according to the one-particle chain. The intuition is that the probability that the distinguished collection of particles interacts with one another becomes asymptotically negligible, and because the interaction is symmetric the interactions with the other particles can be replaced by empirical rates. The terminology comes because the chaotic initial condition (finite collections of particles have asymptotically independent initial conditions) propagates.

Propagation of chaos was first introduced into statistical mechanics by Kac [8]. In an attempt to derive the Boltzmann equation giving the evolution of the density in phase space of the molecules of a dilute gas of hard spheres from mechanics, Kac [8, Section 3] introduced a Markovian caricature. He had the idea that if the initial distribution of the molecules was independent, then for any fixed collection of them the joint distribution should be approximately independent at any given time as long as the total number of molecules is very large. Kac was able to demonstrate this in his Markovian caricature. Subsequently, propagation of chaos has been demonstrated in several models of physical interest. A survey of the subject is presented in Sznitman [11], to which we refer the interested reader.

Our purpose in this paper is to discuss the use of these ideas in studying large networks of interacting queues. In the context of queuing networks, propagation of chaos can be considered a technique for writing *dynamic* Erlang fixed-point equations. Erlang fixed-point equations are widely used in the approximate analysis of networks and have been the subject of several recent papers [5,9,10]. The traditional use of this technique is static in nature, reflecting the quantities at a single time. We hope that becoming aware of the possibility of justifying a dynamic analog of this approximation scheme will increase the range of its applicability.

As a starting point, we begin by familiarizing the reader with the intuition underlying the phenomenon of propagation of chaos by studying a simple example in Section 2. In Section 3 we prove the main theorem of propagation of chaos in a model of pairwise interaction, introduced first by Uchiyama [12], that generalizes the example of Section 2. The tools from the theory of weak convergence of Markov processes that are needed for this proof are summarized in the Appendix.

Then we turn our attention to the problem of computing sojourn time distributions of typical customers in large networks of interacting queues. In Section 4 we discuss a model for dynamic routing in circuit-switched networks that was introduced by Gibbens, Hunt, and Kelly [4]. We indicate how propagation

of chaos can be used to immediately write down the sojourn time distribution of a typical customer in this model. This calculation is typical of situations where all the motion of a typical customer between queues occurs at a fixed time—thus, one can immediately use this proof technique to deal with similar models.

The preceding calculation is in the transient regime; i.e., it depends on the initial conditions. In Section 5 we discuss the problem of determining stationary sojourn times. The key difficulty is that the ordinary differential equation (ODE) describing the evolution of empirical occupancy distributions may, in general, admit multiple equilibria. We prove that if this ODE has a unique equilibrium, then the sojourn time distribution computed via propagation of chaos using this equilibrium as initial condition is the limit of the stationary sojourn time distributions of the finite-particle models.

In Section 6 we turn our attention to models where the motion of typical customers between the queues can take place at different times. In the context of some simple examples, we discuss how one can immediately write down a branching scheme that gives the correct asymptotic sojourn time distribution in the transient regime. The idea behind the validity of the branching scheme is that, with probability asymptotically approaching one, every time a typical customer moves it sees a queue that has not interacted with any of the queues influenced by the queues it has visited so far. The examples chosen to illustrate the ideas in Section 6 are quite artificial; however, they are just representative examples of problems of this kind, and the same proof technique can be used to deal with similar models in other contexts.

Section 7 is devoted to a few summarizing remarks. Before proceeding let us collect some notation at this point. If E denotes a metric space, then $C(E)$ will denote the space of real valued bounded continuous functions on E , $B(E)$ will denote the Banach space of bounded Borel measurable functions on E , and $\mathcal{M}(E)$ will denote the space of probability measures on E . Further, as usual, for any complete separable metric space E , we will let $D_E[0, \infty)$ denote the space of E -valued functions on $[0, \infty)$, which are right continuous and have left limits. $D_E[0, \infty)$ endowed with the Skorohod topology is separable and complete (see Ethier and Kurtz [3, Chapter 3, Section 5]).

A sequence $\{X_n\}$ of E -valued random variables is said to converge weakly to the E -valued random variable X if

$$\lim_{n \rightarrow \infty} E[f(X_n)] = E[f(X)], \quad f \in C(E).$$

We write $X_n \xrightarrow{d} X$ or $X_n \Rightarrow X$. Weak convergence of $D_E[0, \infty)$ -valued random variables is analogously defined. Throughout this paper E will be a compact metric space.

2. A SIMPLE ILLUSTRATIVE EXAMPLE

Consider a collection of n buffers, each of which can hold at most one customer. There are two kinds of arrival processes: simple arrivals that bring one

customer to a buffer, and pair arrivals that bring one customer to each of a pair of buffers. There is one arrival process of simple arrivals at each buffer, and these are independent Poisson processes of rate λ . Pair arrivals are as independent Poisson processes of rate $\alpha/(n - 1)$ for each pair of buffers, so that the rate of pair arrivals involving a specific buffer is α , and these are independent of simple arrivals. A simple arrival is blocked and rejected if its corresponding buffer is already full. A pair arrival is blocked and rejected if any one of the buffers it involves is full. Buffers are held for independent exponentially distributed times of rate 1, after which they become free. (Thus, if a pair arrival is accepted, then the customers it brings in are no longer distinguishable from those that came in as simple arrivals.) This model is a special case of one that has been considered by Hunt [7], Whitt [13], and Ziedins and Kelly [14].

The preceding system has a simple Markovian description. As a state space, we may simply take the number of occupied buffers. The state transition diagram is drawn in Figure 1.

Suppose we start the system with a fraction $u^n(0)$ buffers having customers, and let $(u^n(t), t \in [0, \infty))$ denote the resulting process of fraction of occupied buffers in the system. Let $u(t) \in [0, 1]$ evolve according to the ODE

$$\dot{u}(t) = -u(t) + \lambda(1 - u(t)) + \alpha[1 - u(t)]^2. \tag{2.1}$$

Let \Rightarrow denote weak convergence. Then the following can be easily proved following Gibbens et al. [4]; see also Theorem 3.1.

LEMMA 2.1: *If $u^n(0) \Rightarrow u(0)$ as $n \rightarrow \infty$, then the process $(u^n(t), t \in [0, \infty))$ converges weakly to $(u(t), t \in [0, \infty))$ given by the solution to ODE (2.1) started at $u(0)$.*

Next we adopt a somewhat perverse point of view. We isolate the first p buffers and take as state space the set $\bigcup_{\vec{e} \in \{0,1\}^p} \{(e_1, \dots, e_p, k), 0 \leq k \leq n - p\}$, where (e_1, \dots, e_p, k) denotes the state where the first p buffers are in state e_1, \dots, e_p , respectively, and there are k customers in the remaining $n - p$ buffers. The state transition diagram is drawn in Figure 2 for $p = 1$.

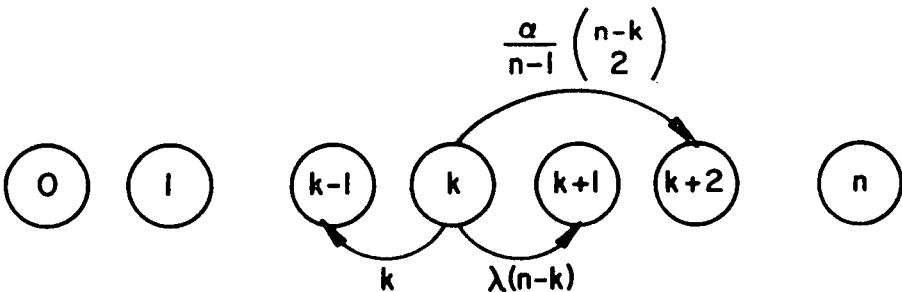


FIGURE 1. State transition diagram.

The concept of propagation of chaos is encapsulated in the following claim. It will be proved in a general framework in the next section.

THEOREM 2.2: Consider the preceding system and let $(x_1(t), \dots, x_p(t))$ denote the state of the first p buffers and $u^{p+1,n}(t)$ the fraction occupied among the buffers $p + 1 \leq j \leq n$ at time t . Suppose $(x_1(0), \dots, x_p(0), u^{p+1,n}(0))$ converges weakly to a product distribution $\mu^{(1)} \otimes \dots \otimes \mu^{(p)} \otimes \delta_{u(0)}$ in $E = \{0, 1\}^p \times [0, 1]$.

For $1 \leq j \leq p$, let $P^{\mu^{(j)}}$ be the probability measure on $D_{[0,1]}[0, \infty)$ corresponding to the time inhomogeneous Markov process with birth rate $\lambda + \alpha(1 - u(t))$, death rate 1, and initial distribution $\mu^{(j)}$, where $u(t)$ is given by ODE (2.1) with initial conditions $u(0)$. Then the process $(X_1(t), \dots, X_p(t), u^{p+1,n}(t))$ converges weakly to a product distribution $P^{\mu^{(1)}} \otimes P^{\mu^{(2)}} \otimes \dots \otimes P^{\mu^{(p)}} \otimes \delta_{u(t)}$ in $D_E[0, \infty)$.

Figure 3 gives a rate diagram of the evolution of buffers 1 through p . Note that they evolve independently according to time inhomogeneous Markov processes. One thinks about this as follows: The evolution of the fraction of occupied buffers among the buffers $p + 1 \leq j \leq n$ converges to ODE limit (2.1). Each of the buffers 1 through p evolves by means of three kinds of transitions: direct arrivals, departures, and pair arrivals involving one of the buffers $p + 1 \leq j \leq n$. Pair arrivals that involve two of the buffers $1 \leq j \leq p$ can be ignored in the limit. This also explains why, when the initial conditions of the buffers $1 \leq j \leq p$ are independent as in the preceding theorem, they remain independent as they evolve.

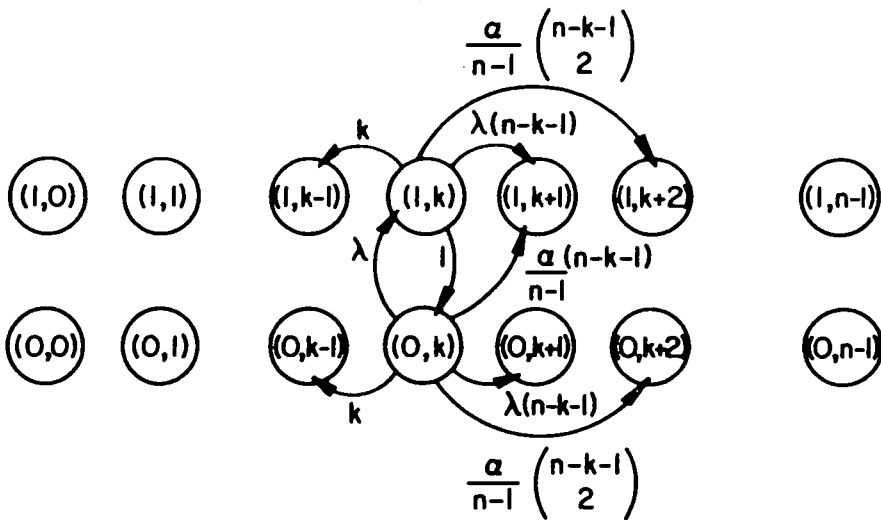


FIGURE 2. State transition diagram for $p = 1$.

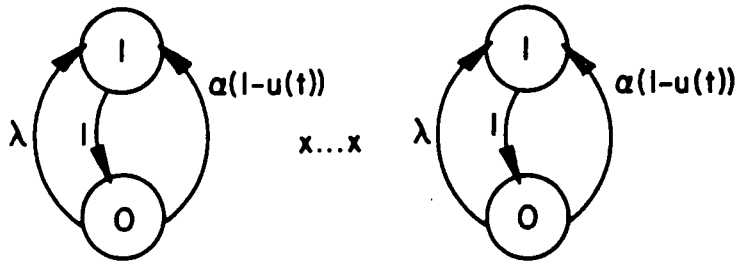


FIGURE 3. Evolution of buffers 1 through p .

Let $u^{(j)}(t)$ denote $P^{n(j)}(X_j(t) = 1)$. Note that $u^{(j)}(t)$ evolves according to the ODE

$$\dot{u}^{(j)} = -u^{(j)}(t) + \lambda[1 - u^{(j)}(t)] + \alpha[1 - u^{(j)}(t)][1 - u(t)]. \quad (2.2)$$

Under stationarity there must be a balance of the mean number of arrivals and the mean number of departures. Consider the equation

$$k = \lambda(n - k) + 2 \frac{\alpha}{n - 1} \frac{(n - k)(n - k - 1)}{2}.$$

Setting $k = \beta n$ and $n \rightarrow \infty$, this becomes

$$\beta = \lambda(1 - \beta) + \alpha(1 - \beta)^2. \quad (2.3)$$

Let β_0 be the unique solution of Eq. (2.3) in $(0,1)$ (see Fig. 4). We expect the stationary distribution of the number of occupied buffers to be sharply concentrated near $\beta_0 n$.

The intuition underlying Theorem 2.2 is particularly clear when $u(0) = \beta_0$. Then we can expect the fraction of occupied buffers to stay at β_0 . Thus, the rate of arrivals at any of the buffers $1 \leq j \leq p$ due to pair arrivals should be $\alpha(1 - \beta_0)$. This is precisely that given by ODE (2.2).

3. A SYSTEM WITH PAIRWISE INTERACTION

We shall introduce a Markovian system that is conceived to model the dynamics of a large system of randomly interacting particles and investigate a propagation of chaos result for it in the path space.

Consider a Markov process $X^n(t) = (X_1^n(t), X_2^n(t), \dots, X_n^n(t))$ on $S^n = S \times S \times \dots \times S$, where S is a finite set. $(X^n(t), t \in [0, \infty))$ is characterized by two sets of nonnegative constants $L = \{L^x(y) : x, y \in S, x \neq y\}$ and $K = \{K^{x,y}(x', y') : x, y, x', y' \in S, (x, y) \neq (x', y')\}$ and an infinitesimal generator given by

$$G_n = \sum_{k=1}^n L_k + \frac{1}{n-1} \sum_{1 \leq k < l \leq n} K_{k,l}$$

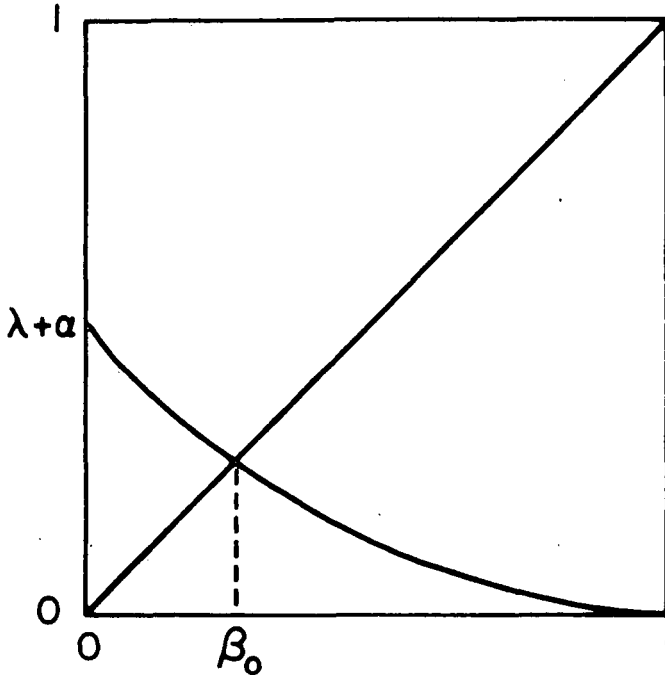


FIGURE 4. Solution to Eq. (2.3).

with

$$L_k \phi(\bar{x}) = \sum_{x'_k \in S} [\phi(\bar{x}'_k) - \phi(\bar{x})] L^{x_k}(x'_k)$$

and

$$K_{k,l} \phi(\bar{x}) = \sum_{x'_k \in S} \sum_{x'_l \in S} [\phi(\bar{x}'_{k,l}) - \phi(\bar{x})] K^{x_k, x_l}(x'_k, x'_l),$$

where $\bar{x} = (x_1, \dots, x_n) \in S^n$, ϕ is a real function on S^n , \bar{x}'_k (resp. $\bar{x}'_{k,l}$) is an element of S^n obtained from \bar{x} by replacing x_k (resp. x_k and x_l) with x'_k (resp. x'_k and x'_l), and the sum $\sum_{k < l}$ is taken over all pairs (k, l) such that $1 \leq k < l \leq n$. Also assume that $K^{x,y}(x', y') = K^{y,x}(y', x')$.

We think of S as the state space of an individual particle. Hence, $X_k^n(t)$ is the physical state of the k th particle in a system of n identical particles.

The process evolves as follows: Each particle k evolves autonomously through a Markovian motion according to L_k . Pairwise interaction between particles k and l is controlled by $K_{k,l}$: two particles in states x and y change simultaneously to states x' and y' , respectively, at rate $[K^{x,y}(x', y')]/(n - 1)$. The factor $n - 1$ is there so that the interaction rate per particle is constant as $n \rightarrow \infty$. Notice that the example discussed in Section 1 fits in this framework with $S = \{0, 1\}$, $L^0(1) = \lambda$, $L^1(0) = 1$, and $K^{0,0}(1, 1) = \alpha$.

Another way to think of the pairwise interaction is as follows: For each $y, x', y' \in S$, each particle in state x chooses another particle at random from the remaining $n - 1$ at rate $\frac{1}{2}K^{x,y}(x', y')$ and, if this particle is in state y , they change together to states x' and y' , respectively. If the chosen particle is not in state y , nothing happens.

Now let $u_i^n(t) = \frac{1}{n} \sum_{l=1}^n 1(X_l^n(t) = i)$. Then $u^n(t) = (u_1^n(t), \dots, u_{|S|}^n(t))$ is a Markov chain on the $|S|$ -dimensional simplex Δ given by

$$\Delta = \left\{ \bar{u} \in R^{|S|}, \sum_{i=1}^{|S|} u_i = 1 \right\}.$$

For $i \neq j, 1 \leq i, j \leq |S|$, let T_{ij} be an operator defined on $\bar{u} \in \Delta$ by

$$T_{ij}\bar{u} = \bar{u} + \frac{1}{n} (e_j - e_i),$$

where e_i is the unit vector in the i th direction. Then the infinitesimal generator A_n of the Markov process $u^n(t)$ on Δ is the operator given by

$$\begin{aligned} A_n\phi(\bar{u}) &= \sum_{\substack{i,j \in S \\ i \neq j}} [\phi(T_{ij}\bar{u}) - \phi(\bar{u})] L^i(j) u_i n \\ &+ \sum_{\substack{i,j,i',j' \in S \\ i \neq i', (i,i') \neq (j,j')}} [\phi(T_{ij}T_{i'j'}\bar{u}) - \phi(\bar{u})] \frac{K^{i,i'}(j,j')}{n-1} \frac{u_i n u_{i'} n}{2} \\ &+ \sum_{\substack{i,j,j' \in S \\ (i,i) \neq (j,j')}} [\phi(T_{ij}T_{ij'}\bar{u}) - \phi(\bar{u})] \frac{K^{i,i}(j,j')}{n-1} \frac{u_i n (u_i n - 1)}{2}, \end{aligned}$$

where ϕ is a continuous function on Δ .

Let $u(t) \in \Delta$ evolve according to the following equation started with $u(0)$:

$$\begin{aligned} \dot{u}_i(t) &= \sum_{\substack{j \in S \\ j \neq i}} L^j(i) u_j - \sum_{\substack{j \in S \\ j \neq i}} L^i(j) u_i \\ &+ \sum_{i' \in S} \sum_{j \in S} K_1^{i',j}(i) u_{i'} u_j - \sum_{i' \in S} \sum_{j \in S} K_1^{i',j}(i') u_i u_j \end{aligned} \tag{3.1}$$

for $i \in S$, where $K_1^{x,y}(x') = \sum_{y' \in S} K^{x,y}(x', y')$.

Then the idea of propagation of chaos is captured by the following theorem.

THEOREM 3.1: *For the preceding system of n interacting particles, let $(X_1(t), \dots, X_p(t))$ denote the state of the first p particles and $u_x^{p+1,n}(t)$ the fraction of particles $p + 1 \leq l \leq n$ that are in state x at time t . Let $u^{p+1,n}(t) = (u_x^{p+1,n}(t), x \in S)$. Suppose $(X_1(0), \dots, X_p(0), u^{p+1,n}(0))$ converges weakly to a product distribution $\mu^{(1)} \otimes \dots \otimes \mu^{(p)} \otimes \delta_{u(0)}$ in $E = S^p \times \Delta$. Let $u(t)$ solve ODE (3.1) starting at $u(0)$. For $1 \leq l \leq p$, let $P^{\mu^{(l)}}$ be the probability measure on $D_E[0, \infty)$ corresponding to the time inhomogeneous Markov chain $X(t)$ with state space*

S , with initial distribution $\mu^{(l)}$, and such that the rate of jumping from state s to state s' is

$$\lambda(u(t), s, s') = L^s(s') + \sum_{\substack{i, j \in S \\ (i, s) \neq (j, s')}} K^{i, s}(j, s') u_i(t). \tag{3.2}$$

Then the process $(X_1(t), \dots, X_p(t), u^{p+1, n}(t))$ converges weakly to a product distribution $P^{\mu^{(1)}} \otimes \dots \otimes P^{\mu^{(p)}} \otimes \delta_{u(t)}$ in $D_E[0, \infty)$.

Remarks:

1. Theorem 3.1 describes not only the limiting behavior of a single particle within a finite set of particles but also the limiting value at any given time of the fraction of particles in a given state.
2. The meaning of this result is that the evolution of p particles in a sea of a large number of particles can be described by replacing the interactions of the individual particles in the sea by the empirical averages of their effects and ignoring the mutual interactions of the p particles. This is possible because of the symmetry in the interactions. Note that one can easily write down the limiting evolution of $(X_1(t), \dots, X_p(t), u^{p+1, n}(t))$ for more general limiting initial distributions by thinking of them as mixtures of the initial distributions described in Theorem 3.1.

PROOF: We consider a system of n particles and look at the evolution of the first p particles. Consider the Markov process $Y^n(t) = (X_1(t), \dots, X_p(t), u^{p+1, n}(t))$ taking its values in $S^p \times \Delta$. Then $(Y^n(t), t \in [0, \infty))$ has a generator given by

$$\begin{aligned} & G_n \phi(x_1, \dots, x_p, \bar{u}) \\ &= \sum_{\substack{i, j \in S \\ i \neq j}} [\phi(\bar{x}, T_{ij} \bar{u}) - \phi(\bar{x}, \bar{u})] L^i(j) u_i(n-p) \\ &+ \sum_{\substack{i, j, i', j' \in S \\ (i, i') \neq (j, j'), i \neq i'}} [\phi(\bar{x}, T_{ij} T_{i'j'} \bar{u}) - \phi(\bar{x}, \bar{u})] \frac{K^{i, i'}(j, j')}{n-1} \frac{u_i u_{i'}(n-p)^2}{2} \\ &+ \sum_{\substack{i, j, j' \in S \\ (i, j) \neq (j, j')}} [\phi(\bar{x}, T_{ij} T_{ij'} \bar{u}) - \phi(\bar{x}, \bar{u})] \frac{K^{i, i}(j, j')}{n-1} \frac{u_i(n-p)(u_i(n-p)-1)}{2} \\ &+ \sum_{l=1}^p \sum_{x_j \in S} [\phi(\bar{x}'_l, \bar{u}) - \phi(\bar{x}, \bar{u})] L^{x_l}(x'_l) \\ &+ \sum_{l < k} \sum_{\substack{x'_l, x'_k \in S \\ (x'_l, x'_k) \neq (x'_l, x'_k)}} [\phi(\bar{x}'_{l, k}, \bar{u}) - \phi(\bar{x}, \bar{u})] \frac{K^{x_l, x_k}(x'_l, x'_k)}{n-1} \\ &+ \sum_{l=1}^p \sum_{\substack{i, j, x'_l \in S \\ (i, x'_l) \neq (j, x'_l)}} [\phi(\bar{x}'_l, T_{ij} \bar{u}) - \phi(\bar{x}, \bar{u})] \frac{K^{i, x'_l}(j, x'_l)}{n-1} u_i(n-p), \tag{3.3} \end{aligned}$$

where $\phi \in C(S^p \times \Delta)$.

Let $F(\bar{x}, \bar{u}) = (f_1(\bar{x}, \bar{u}), \dots, f_{|S|}(\bar{x}, \bar{u}))$, where $f_i(\bar{x}, \bar{u})$ is the right-hand side of ODE (3.1) for $u_i(t)$ (note that $F(\bar{x}, \bar{u})$ does not depend on \bar{x}). Let us define the following operator:

$$G\phi(\bar{x}, \bar{u}) = \sum_{l=1}^p \sum_{x_j \in S} [\phi(\bar{x}'_l, \bar{u}) - \phi(\bar{x}, \bar{u})] \lambda(\bar{u}, x_l, x'_l) + F(\bar{x}, \bar{u}) \cdot \nabla_{\bar{u}} \phi(\bar{x}, \bar{u}) \tag{3.4}$$

for all $\phi \in C_1(S^p \times \Delta)$.

Now $G = G_1 + G_2$, where

$$G_2\phi(\bar{x}, \bar{u}) = \sum_{l=1}^p \sum_{x_j \in S} [\phi(\bar{x}'_l, \bar{u}) - \phi(\bar{x}, \bar{u})] \lambda(\bar{u}, x_l, x'_l)$$

with $\mathcal{D}(G_2) = C(S^p \times \Delta)$,

$$G_1\phi(\bar{x}, \bar{u}) = \sum_{i \in S} f_i(\bar{x}, \bar{u}) \frac{\partial \phi}{\partial u_i} \quad \text{with } \mathcal{D}(G_1) = C_1(S^p \times \Delta).$$

G_2 is a bounded dissipative linear operator on $C(S^p \times \Delta)$ (it satisfies the positive maximum principle) and, hence, clearly generates a Feller semigroup $\{T_2(t)\}$ on $C(S^p \times \Delta)$ (see Lemma A.4).

Now let $z : [0, \infty) \times S^p \times \Delta \rightarrow S^p \times \Delta$ be the unique solution of the initial value problem:

$$\begin{aligned} \dot{z}(t, \bar{x}, \bar{u}) &= F(z(t, \bar{x}, \bar{u})), \\ z(0, \bar{x}, \bar{u}) &= (\bar{x}, \bar{u}). \end{aligned}$$

The existence and uniqueness of z follows because F is Lipschitz continuous on $S^p \times \Delta$. Therefore, the formula

$$T_1(t)f(\bar{x}, \bar{u}) = f(z(t, \bar{x}, \bar{u}))$$

defines a strongly continuous positive contraction semigroup $\{T_1(t)\}$ on $C(S^p \times \Delta)$ because $z(t + s, \bar{x}, \bar{u}) = z(t, z(s, \bar{x}, \bar{u}))$. Moreover, because F has continuous first partial derivatives, so does z [6, Theorem 3.1, p. 95]. Hence, using the chain rule we conclude that

$$T_1(t) : C_1(S^p \times \Delta) \rightarrow C_1(S^p \times \Delta). \tag{3.5}$$

The infinitesimal generator A of $\{T_1(t)\}$ is closed by Lemma A.1. $C_1(S^p \times \Delta)$ is a core for A by Lemma A.2. Hence, A , by Eq. (3.5), equals the closure of its restriction to $C_1(S^p \times \Delta)$; i.e., $A = \bar{G}_1$.

Now using Lemma A.5, we conclude that there exists a strongly continuous positive contraction semigroup $\{T(t)\}$ on $C(S^p \times \Delta)$ generated by $\bar{G}_1 + \bar{G}_2$. Because $(1, 0)$ belongs to the graph of G , $\{T(t)\}$ is conservative; hence, it is a Feller semigroup. Because \bar{G}_2 is bounded and $C_1(S^p \times \Delta)$ is a core for \bar{G}_1 , it follows that $C_1(S^p \times \Delta)$ is a core for $\bar{G}_1 + \bar{G}_2$.

Clearly e^{tG_n} is a Feller semigroup on $C(S^p \times \Delta)$. Suppose that for each $f \in C(S^p \times \Delta)$

$$\lim_{n \rightarrow \infty} e^{tG_n} f = T(t)f \quad \text{for } t \geq 0. \tag{3.6}$$

Then, if $Y^n(0)$ has limiting distribution $\mu^{(1)} \otimes \mu^{(2)} \dots \otimes \mu^{(p)} \otimes \delta_{u(0)} \in M(S^p \times \Delta)$, there exists a strong Markov process $Y(t)$ corresponding to $\{T(t)\}$ with initial distribution $\mu^{(1)} \otimes \mu^{(2)} \otimes \dots \otimes \mu^{(p)} \otimes \delta_{u(0)}$ and sample paths in $D_{S^p \times \Delta}[0, \infty)$ by Lemma A.7. Moreover, Y^n converges weakly to Y .

Note that under $T(t)$ the evolution on $S^p \times \Delta$ corresponds to the S^p part being p independent copies of a time inhomogeneous Markov process with initial distribution $\mu^{(i)}$, $i = 1, \dots, p$, and rates given by Eq. (3.2) and with the Δ part being a continuous deterministic motion started at $u(0)$ and satisfying ODE (3.1).

Hence, all we need to show is Eq. (3.6). By Lemma A.3 it is enough to show that for each f belonging to a core D of G , $G_n f \rightarrow Gf$. Hence, it suffices to show that $\lim_{n \rightarrow \infty} G_n \phi(\bar{x}, \bar{u}) = G\phi(\bar{x}, \bar{u})$ for $(\bar{x}, \bar{u}) \in S^p \times \Delta$, $\phi \in C_1(S^p \times \Delta)$. We consider the terms of $G_n \phi(\bar{x}, \bar{u})$ separately.

The first term

$$\begin{aligned} & \sum_{\substack{i, j \in S \\ i \neq j}} [\phi(\bar{x}, T_{ij}\bar{u}) - \phi(\bar{x}, \bar{u})] L^i(j)u_i(n-p) \\ &= \sum_{i, j \in S, i \neq j} \left[-\frac{1}{n} \frac{\partial \phi}{\partial u_i} + \frac{1}{n} \frac{\partial \phi}{\partial u_j} + o_n\left(\frac{1}{n}\right) \right] L^i(j)u_i(n-p) \\ &\rightarrow \sum_{\substack{i, j \in S \\ i \neq j}} \left(-\frac{\partial \phi}{\partial u_i} + \frac{\partial \phi}{\partial u_j} \right) L^i(j)u_i \quad \text{as } n \rightarrow \infty \\ &= \sum_{i \in S} \frac{\partial \phi}{\partial u_i} \sum_{\substack{j \in S \\ i \neq j}} [L^j(i)u_j - L^i(j)u_i]. \end{aligned}$$

The second term

$$\begin{aligned} & \sum_{\substack{i, j, i', j' \in S \\ i \neq i', (i, i') \neq (j, j')}} [\phi(\bar{x}, T_{ij}T_{i'j'}\bar{u}) - \phi(\bar{x}, \bar{u})] \frac{K^{i, i'}(j, j')}{n-1} \frac{u_i u_{i'}(n-p)^2}{2} \\ &= \sum_{\substack{i, j, i', j' \in S \\ i \neq i', (i, i') \neq (j, j')}} \left[\frac{1}{n} \left(\frac{\partial \phi}{\partial u_j} + \frac{\partial \phi}{\partial u_{j'}} - \frac{\partial \phi}{\partial u_i} - \frac{\partial \phi}{\partial u_{i'}} \right) + o_n\left(\frac{1}{n}\right) \right] \\ &\quad \times \frac{K^{i, i'}(j, j') u_i u_{i'}(n-p)^2}{2(n-1)} \\ &\rightarrow \sum_{\substack{i, j, i', j' \in S \\ i \neq i', (i, i') \neq (j, j')}} \left[\frac{\partial \phi}{\partial u_j} + \frac{\partial \phi}{\partial u_{j'}} - \frac{\partial \phi}{\partial u_i} - \frac{\partial \phi}{\partial u_{i'}} \right] K^{i, i'}(j, j') \frac{u_i u_{i'}}{2} \quad \text{as } n \rightarrow \infty. \end{aligned}$$

The third term in the same way converges to

$$\sum_{\substack{i,j,j' \in S \\ (i,i) \neq (j,j')}} \left[\frac{\partial \phi}{\partial u_j} + \frac{\partial \phi}{\partial u_{j'}} - 2 \frac{\partial \phi}{\partial u_i} \right] K^{i,i}(j,j') \frac{u_i^2}{2}.$$

Adding up the second and third terms, we get

$$\sum_{i \in S} \frac{\partial \phi}{\partial u_i} \left[\sum_{i' \in S} \sum_{j \in S} K^{i',j}(i) u_{i'} u_j - \sum_{i' \in S} \sum_{j \in S} K^{i',j}(i') u_{i'} u_j \right].$$

So clearly the first three terms converge to $F(\vec{x}, \vec{u}) \cdot \nabla_{\vec{u}} \phi(\vec{x}, \vec{u})$, which is the second term of the right-hand side of Eq. (3.4).

The fifth term clearly goes to 0. The last term

$$\begin{aligned} & \sum_{l=1}^p \sum_{i,j,x'_j \in S} [\phi(\vec{x}'_l, T_{ij} \vec{u}) - \phi(\vec{x}, \vec{u})] \frac{K^{i,x'_j}(j,x'_j)}{n-1} u_i (n-p) \\ &= \sum_{l=1}^p \sum_{i,j,x'_j \in S} \left[\phi(\vec{x}'_l, \vec{u}) - \phi(\vec{x}, \vec{u}) + \frac{1}{n} \frac{\partial \phi}{\partial u_j} - \frac{1}{n} \frac{\partial \phi}{\partial u_i} + o_n\left(\frac{1}{n}\right) \right] \\ & \quad \times \frac{K^{i,x'_j}(j,x'_j) u_i (n-p)}{n-1} \\ & \rightarrow \sum_{l=1}^p \sum_{\substack{i,j,x'_j \in S \\ (i,x'_j) \neq (j,x'_j)}} [\phi(\vec{x}'_l, \vec{u}) - \phi(\vec{x}, \vec{u})] K^{i,x'_j}(j,x'_j) u_i \quad \text{as } n \rightarrow \infty. \end{aligned}$$

By adding the fourth term, one gets

$$\sum_{l=1}^p \sum_{x'_j \in S} [\phi(\vec{x}'_l, \vec{u}) - \phi(\vec{x}, \vec{u})] \left[L^{x'_j}(x'_j) + \sum_{\substack{i,j \in S \\ (i,x'_j) \neq (j,x'_j)}} K^{i,x'_j}(j,x'_j) \right] u_i.$$

Clearly this is the first term of the right-hand side of Eq. (3.4), verifying Eq. (3.6). ■

4. APPLICATION TO DYNAMIC ROUTING MODELS

Our focus in this paper is on the use of propagation of chaos as a tool to compute sojourn times in large networks of interacting queues. In this section we illustrate this through some simple models for dynamic routing.

In recent years there has been considerable interest in the performance of dynamic routing strategies in circuit-switched networks. Dynamic routing schemes adaptively adjust routing patterns within the network, making better use of spare capacity and also providing extra flexibility and robustness to failures or overloads. Among the dynamic routing strategies, a natural one is random alternate routing, which works as follows: Every call that arrives to the network has a fixed first-choice route, a set of possible second-choice routes,

and possibly a third, fourth, and so forth. Whenever its first-choice route is free, the call will occupy it. Otherwise, an alternate route is chosen from the set of possible second-choice routes. If the route selected is busy, then the call may be allowed to select from a set of third-choice routes and, if necessary, a fourth and so forth. The set of alternate routes at any time can depend on the state of the network.

Here we consider a simple model introduced by Gibbens et al. [4]. There are n links, each link comprised of C circuits. At each link, calls arrive as a Poisson process of rate ν . If the link is not saturated, then the call occupies one circuit. If the link is saturated, then the call chooses two distinct links at random from the $n - 1$ remaining links. If neither one is saturated, then the call occupies one circuit from each of these two links. Otherwise, the call is lost. All circuit holding times are exponentially distributed with unit mean, independent of one another and of the arrival times. Furthermore, a call holding circuits from two links is assumed to release them independently.

Let $u_k^n(t)$, $0 \leq k \leq C$, be the fraction of the n links that have k occupied circuits at time t . Then $u^n(t) = (u_0^n(t), u_1^n(t), \dots, u_C^n(t))$ evolves on a C -dimensional simplex. Gibbens et al. [4] have shown that as $n \rightarrow \infty$, if the initial condition $u^n(0)$ converges weakly to a limit $u(0)$, then the process $(u^n(t), t \in [0, \infty))$ converges weakly to a deterministic process $(u(t), t \in [0, \infty))$, satisfying the following set of differential equations:

$$\begin{aligned} \dot{u}_0 &= u_1 - (\nu + 2\nu u_C(1 - u_C))u_0, \\ \dot{u}_k &= (k + 1)u_{k+1} + (\nu + 2\nu u_C(1 - u_C))u_{k-1} \\ &\quad - (k + \nu + 2\nu u_C(1 - u_C))u_k, \quad 0 < k < C, \\ \dot{u}_C &= -Cu_C + (\nu + 2\nu u_C(1 - u_C))u_{C-1}. \end{aligned} \tag{4.1}$$

Equations in ODE (4.1) describe the evolution of the empirical fraction of number of links having a given occupancy at a given time. A more detailed picture is given by a description of the evolution of a single link in the network. Such a description is possible on the basis of a somewhat more general version of Theorem 3.1. Let $S = \{0, 1, \dots, C\}$ and $M(S)$ the space of probability distributions on S . Let $(X^n(t), t \geq 0)$ denote the Markov chain on S^n corresponding to the preceding dynamic routing model with n links. Let $(X_1(t), \dots, X_p(t))$ denote the states of the first p links and $u_k^{p+1, n}(t)$, $0 \leq k \leq C$, the fraction of the links $p + 1 \leq j \leq n$ that have k occupied circuits. Let $u^{p+1, n}(t) = (u_0^{p+1, n}(t), \dots, u_C^{p+1, n}(t))$. Then we have Theorem 4.1.

THEOREM 4.1: *If $(X_1(0), \dots, X_p(0), u^{p+1, n}(0))$ converges weakly to the product $\mu^{(1)} \otimes \dots \otimes \mu^{(p)} \otimes \delta_{u(0)}$ in $E = S^p \times M(S)$, then the process $(X_1(t), \dots, X_p(t), u^{p+1, n}(t))$ converges weakly to $P^{\mu^{(1)}} \otimes \dots \otimes P^{\mu^{(p)}} \otimes \delta_{u(t)}$ in $D_E[0, \infty)$, where $u(t)$ is given by the solution to ODE (4.1) starting at $u(0)$ and $P^{\mu^{(l)}} \in D_S[0, \infty)$ ($1 \leq l \leq p$) is the time inhomogeneous birth and death process with*

initial distribution $\mu^{(1)}$, birth rates $b_j = \nu + 2\nu u_C(t)(1 - u_C(t))$, $0 \leq j \leq C - 1$, and death rates $d_j = j$, for $0 < j \leq C$.

PROOF: This follows directly from a version of Theorem 3.1 that allows for three-particle interaction. ■

Based on the preceding, we can give an expression for the asymptotic distribution of the sojourn time of a typical customer in the preceding dynamic routing model.

THEOREM 4.2: *Suppose the situation is as that in Theorem 4.1 with $p = 1$, and let $u_k^{(1)}(t) = P^{\mu^{(1)}}(X_1(t) = k)$.*

Consider a customer arriving at link 1 at time t_0 , and let T_n take values in $R_+ \cup \{\beta\}$ with

$$P(T_n \leq t) = P(\text{sojourn time of the job that corresponds to the customer} \leq t),$$

$$P(T_n = \beta) = P(\text{customer is blocked}).$$

Then $T_n \xrightarrow{d} T$ as $n \rightarrow \infty$ where

$$P(T \leq t) = [1 - u_C^{(1)}(t_0)][1 - e^{-t}] + u_C^{(1)}(t_0)[1 - u_C(t_0)]^2(1 - e^{-t})^2, \tag{4.2}$$

$$P(T = \beta) = u_C^{(1)}(t_0)[1 - (1 - u_C(t_0))^2].$$

Furthermore, $u_C^{(1)}(t)$ solves the equations

$$\begin{aligned} \dot{u}_0^{(1)} &= u_1^{(1)} - (\nu + 2\nu u_C(1 - u_C))u_0^{(1)}, \\ \dot{u}_k^{(1)} &= (k + 1)u_{k+1}^{(1)} + (\nu + 2\nu u_C(1 - u_C))u_{k-1}^{(1)} \\ &\quad - (k + \nu + 2\nu u_C(1 - u_C))u_k^{(1)}, \quad 0 < k < C, \\ \dot{u}_C^{(1)} &= -Cu_C^{(1)} + (\nu + 2\nu u_C(1 - u_C))u_{C-1}^{(1)}. \end{aligned} \tag{4.3}$$

PROOF: Let Y_1 be the state of link 1 at t_0^- , and Y_2, Y_3 the state of the two queues the customer occupies if queue 1 is full. Then for $k_2 \neq k_3$,

$$P((Y_1, Y_2, Y_3) = (k_1, k_2, k_3))$$

$$= \frac{1}{\binom{n-1}{2}} \sum_{2 \leq j < k \leq n} P(X_1^n(t_0) = k_1, X_j^n(t_0) = k_2, X_k^n(t_0) = k_3)$$

$$= \sum_{m+l \leq n-1} P\left(X_1(t_0) = k_1, u_{k_2}^{2,n}(t_0) = \frac{m}{n-1}, u_{k_3}^{2,n}(t_0) = \frac{l}{n-1}\right)$$

$$\times \frac{ml}{\binom{n-1}{2}}$$

$$\rightarrow u_{k_1}^{(1)}(t_0)u_{k_2}(t_0)u_{k_3}(t_0)$$

as $n \rightarrow \infty$ by Theorem 4.1. A similar calculation holds for $k_2 = k_3$.

Because $P(T_n < t) = (1 - e^{-t})P(Y_1 \neq C) + (1 - e^{-t})^2 \cdot (P(Y_1 = C, Y_2 < C, Y_3 < C))$ and $P(T_n = \beta) = P(Y_1 = C, \max(Y_2, Y_3) = C)$, ODE (4.2) follows. Equations in ODE (4.3) are a direct consequence of the evolution equation for the distribution of $X^{(1)}(t)$. ■

Remarks:

1. Note that one can write down the limiting sojourn time distribution for more general initial conditions by thinking of them as mixtures of those in Theorem 4.1.
2. Of course, sojourn times are of little interest in actual circuit-switched networks, because once a call establishes a connection, there is no queuing and the sojourn time is just the holding time of the call. (The form of the preceding sojourn time is an artifact of the model, coming from the assumption that a call using two links releases them independently.) In virtual circuit networks where the time for completion of the call may depend on the length of the circuit, sojourn times may be of interest. For example, one may think of the preceding model as representing the sojourn time of a call along a link by an exponentially distributed random variable of unit mean, and for virtual circuits consisting of two links, ignoring the fact that one of the links must be traversed before the other.
3. It is natural to want to compute the stationary sojourn time of a typical customer in a large system. That is, we consider the n link system in its stationary distribution, which is exchangeable by symmetry, and ask for the limiting distribution of the sojourn time of a typical customer as the total number of links becomes large. The description of sojourn times in this stationary regime is complicated by the fact that for large enough values of C equations in ODE (4.1) admit two stable equilibria for ν in a range (ν_m, ν_M) (see Gibbens et al. [4]). We expect that there will be a critical value $\nu_m < \nu_C < \nu_M$ such that for $\nu \in (\nu_m, \nu_C)$ (resp. $\nu \in (\nu_C, \nu_M)$) the lower (resp. upper) stable equilibrium point corresponds to the stationary regime. Identifying ν_C requires an analysis of the exit times from the basins of attraction of the stable equilibria—the stationary regime corresponds to the deeper basin.

Once the deeper basin is identified, the sojourn time of a typical customer in the stationary regime is described by equations in ODE (4.2) with the appropriate u_C . Of course, this limiting description is somewhat deceptive. The limiting sojourn time of a typical customer will correspond to one or the other equilibrium only because the large finite system predominantly spends time in the mode corresponding to that equilibrium. In fact, the large finite system will move between operating regimes corresponding to the different equilibria and a correct understanding of the system will require an analysis of the sojourn times in each regime. Some further insight into the possibility of multiple regimes

of operation in networks with dynamic routing is given by Anantharam [1]. When ODE (4.1) has a unique solution, such difficulties do not arise (see Section 5 for a discussion of this).

Similar expressions can be written for more realistic models that allow for the possibility of retries and trunk reservation. Here a call that is not successfully routed on its first-choice route chooses two links from the $n - 1$ remaining links. If both chosen links have less than $C - s$ busy circuits, then the call occupies that pair of links. If not, the call tries a pair from the $n - 3$ remaining links. The call is lost after it has tried r pairs. The parameter s is known as the trunk reservation parameter, and r is the number of retries. The model described at the beginning of this section corresponds to the case $s = 0$ and $r = 1$. One can state theorems similar to Theorems 4.1 and 4.2 in this context.

One can also consider the possibility of queuing for virtual circuits. For example, in the preceding examples consider the situation where each link can carry C calls and buffer $K - C$ additional calls. Then the sojourn time expressions will involve the distribution of sojourn times in an $M/M/C/K$ queue.

The examples in this section were chosen to illustrate the applicability of this method. Similar ideas can be useful in a wide variety of situations. From the preceding examples, it should be clear by now how to apply the method when all movement of customers between the queues takes place at a single time. The more interesting situations, where movements between queues take place at different times, will be discussed in Section 6.

5. STATIONARY SOJOURN TIME DISTRIBUTIONS

In the preceding models, the differential equations describing the evolution of the empirical distribution of particles may have more than one equilibrium solution. Hence, to describe the sojourn time of a typical customer in the stationary regime requires identifying which equilibrium of the ODE dominates. This requires an analysis of the exit times from the basins of attraction of the stable equilibria—the dominating stationary regime should correspond to the deeper basin. If, however, the limiting differential equations have a unique solution, determining the sojourn time distribution in stationarity is easier.

THEOREM 5.1: *Let π_n be the stationary distribution of the Markov process $X^n(t)$ described by rates L and K as in Section 3, which we assume is irreducible. Let $u_x^n(t)$ denote the fraction of particles that are in state x at time t and $u^n(t) = (u_x^n(t), x \in S)$. By Theorem 3.1 we know that if $u^n(0)$ converges weakly to $u(\cdot, 0)$, then $(u^n(t), t \in [0, \infty))$ converges weakly to $(u(t), t \in [0, \infty))$, which solves ODE (3.1). Suppose there is a unique equilibrium π to this ODE. Let $\pi_{n|p}$ denote the restriction of π_n to the first p coordinates. Then for any $p \geq 1$, $\pi_{n|p}$ converges weakly to the p -fold product $\pi \otimes \cdots \otimes \pi$ as $n \rightarrow \infty$.*

PROOF: We first consider $p = 1$. Let $\vec{\gamma} = (\gamma_k, 1 \leq k \leq |S|)$ be the function from S^n to $M(S)$ such that $\gamma_k(\vec{x}) = \frac{1}{n} \sum_{i=1}^n 1(x_i = k)$. Let $X^n(0) \stackrel{d}{=} \pi^n$; hence,

$X^n(t) \stackrel{d}{=} \pi^n$. Because $M(S)$ is compact, so is $M(M(S))$ in the weak topology. Therefore, $\bar{\gamma}(X^n(0))$ has a convergent subsequence $\bar{\gamma}(X^{n_k}(0))$ such that $\bar{\gamma}(X^{n_k}(0)) \xrightarrow{d} u(0)$ for some $u(0) \in M(S)$. It follows that $(X^{n_k}(t)) \xrightarrow{d} u(t)$, where $u(t)$ solves ODE (3.1). In stationarity, however, we have $(X^{n_k}(t)) \stackrel{d}{=} (X^{n_k}(0))$. Because $u(t)$ converges to π , we must have $u(0) \sim \delta_\pi$. Thus, every subsequential weak limit of $\bar{\gamma}(X^n(0))$ is π , so $\bar{\gamma}(X^n(0))$ converges weakly to π . But it is easy to see that $\bar{\gamma}(X^n(0))$ is just $\pi_{n|1}$. Hence, $\pi_{n|1}$ converges weakly to π .

For $p = 2$ we again start the chain $X^n(t)$ in π_n . Let $M_n(S)$ denote the subset of $M(S)$ consisting of empirical distributions based on n particles and $M_n(S \times S)$ similarly for empirical distributions of pairs. Let $F_n : M_n(S) \rightarrow M_n(S \times S)$ be the mapping that constructs pair measures. Let $F : M(S) \rightarrow M(S \times S)$ construct product measure and also extend F_n to $F_n : M(S) \rightarrow M(S \times S)$ by linear interpolation. Along a subsequence such that $\bar{\gamma}(X^{n_k}(0)) \xrightarrow{d} u(0)$, we have $\bar{\gamma}(X^{n_k}(t)) \xrightarrow{d} u(t)$. Hence, $F \circ \bar{\gamma}(X^{n_k}(t)) \xrightarrow{d} Fou(t)$. But $u(t) \rightarrow \pi$, so $Fou(t) \rightarrow \pi \times \pi$. We must therefore have $Fou(0) \sim \delta_{\pi \times \pi}$. By Theorem 4.1 in Billingsley [2, p. 25], we have $F_n \circ \bar{\gamma}(X^{n_k}(0)) \rightarrow \pi \times \pi$. Thus, $\pi_{n|2}$ converges weakly to $\pi \times \pi$.

The same argument can be repeated for $p > 2$. ■

6. MODELS WITH MOVEMENT AT SEVERAL TIMES

The technique described in this paper for computing the sojourn time distribution of typical customers is not limited to movement between queues occurring at a fixed time. For situations where the movement occurs with delays (e.g., models with impatience), a branching scheme can be used to write down the limiting distribution of the sojourn time of a typical customer. As a simple illustrative example, consider a system of n M/M/1/2 queues with impatience. The arrival rate of customers to each queue is ν , and the service rate at each queue is 1. An arriving customer who finds the queue full is rejected. A customer in the second position gets impatient at rate β , chooses a queue at random from the $(n - 1)$ remaining queues, and moves to it only if it is empty. The arrival, service, and impatience processes are all mutually independent.

Let $S = \{0, 1, 2\}$, and let $(X^n(t), t \geq 0)$ denote the Markov chain on S^n corresponding to the model described earlier. Let $u^n(t) = (u_0^n(t), u_1^n(t), u_2^n(t))$ denote the fractions of the n queues that have k customers at time t , $0 \leq k \leq 2$. Then one can show that as $n \rightarrow \infty$, if the initial condition $u^n(0)$ converges weakly to a limit $u(0)$, then the process $u^n(t)$ converges weakly to a deterministic process $u(t)$ satisfying the following:

$$\begin{aligned}
 \dot{u}_0 &= u_1 - \nu u_0 - \beta u_0 u_2, \\
 \dot{u}_1 &= u_2 + \nu u_0 - u_1 + 2\beta u_2 u_0 - \nu u_1, \\
 \dot{u}_2 &= \nu u_1 - u_2 - \beta u_2 u_0.
 \end{aligned}
 \tag{6.1}$$

It is also easy to see that these ODEs have a unique equilibrium (π_0, π_1, π_2) .

The following is an example of the kind of theorem one can prove.

THEOREM 6.1: *Suppose we start $(X^n(t), t \geq 0)$ with initial distribution μ_n , which is symmetric with respect to permutations, and let $\mu_{n|p}$ denote the restriction of μ_n to the first p coordinates. Suppose $\mu_{n|1} \rightarrow \mu$ so that $\mu_{n|p} \rightarrow \mu \otimes \dots \otimes \mu$. Let μ_t denote the solution of ODE (6.1) starting from μ . Then we have the following:*

- (i) *Consider a customer arriving to queue 1 at time 0. Let T_n be the sojourn time of this customer in the system conditional on it being accepted. Then T_n converges weakly to T where*

$$T \stackrel{d}{=} \frac{\mu(0)}{(\mu(0) + \mu(1))} E_1 + \frac{\mu(1)}{(\mu(0) + \mu(1))} (R + E_1),$$

where $P(R > t) = e^{-\int_0^t (1 + \beta \mu_s(0)) ds}$, E_a denotes an exponentially distributed random variable of mean a , and R is independent of E_1 .

- (ii) *Consider the system in stationarity, and let the sojourn time of a typical customer conditional on it being accepted be Z_n . Then Z_n converges weakly to Z where*

$$Z \stackrel{d}{=} E_1 + \frac{\pi_1}{\pi_0 + \pi_1} E_{1 + \pi_0 \beta}.$$

PROOF:

- (i) *Because a tagged customer arriving at queue 1 at time 0 sees the initial distribution, we have $T_n \stackrel{d}{=} \mu_{n|1}(0) / [\mu_{n|1}(0) + \mu_{n|1}(1)] E_1 + \mu_{n|1}(1) / [\mu_{n|1}(1) + \mu_{n|1}(0)] (\tau_n^1 + E_1)$, where $\tau_n^1 = \min(E_1, \tau_n)$, with τ_n being the time for the tagged customer to get impatient and successfully transfer to an empty queue. Let $(\sigma_1, \sigma_2, \dots)$ be the virtual impatience process (a Poisson process of rate β) at queue 1. Let Y_1, Y_2, \dots be the states of the queues the tagged customer tries to move to at times $\sigma_1, \sigma_2, \dots$. Let R_1, R_2, \dots taking values in $\{2, \dots, n\}$ be the indices of these queues. Then*

$$\begin{aligned} P(\tau_n^1 > t) &= P(\sigma_1 > t, E_1 > t) + P(\sigma_1 + \sigma_2 > t, Y_1 \neq 0, \sigma_1 \leq t, E_1 > t) \\ &+ \dots + P\left(\left(\sum_{i=1}^k \sigma_i\right) > t, Y_1 \neq 0 \dots Y_{k-1} \neq 0, \right. \\ &\quad \left. \times \left(\sum_{i=1}^{k-1} \sigma_i\right) \leq t, E_1 > t\right) + \dots. \end{aligned}$$

The k th term can be written as

$$L_k = \int_{\sum_{i=1}^{k-1} s_i \leq t} P\left(\sigma_k > t - \left(\sum_{i=1}^{k-1} s_i\right), E_1 > t, B, D\right),$$

where

$$B = \{Y_1 \neq 0, \dots, Y_{k-1} \neq 0\}$$

and

$$D = \{\sigma_1 \in (s_1, s_1 + ds_1) \cdots \sigma_{k-1} \in (s_{k-1}, s_{k-1} + ds_{k-1})\}.$$

Conditioning on the queues checked for potential transfer by the virtual impatience process at queue 1, the integrand can be written as

$$\begin{aligned} & \frac{1}{(n-1)^{k-1}} \sum_{2 \leq i_1, \dots, i_{k-1} \leq n} P\left(\sigma_k > t - \left(\sum_{i=1}^{k-1} s_i\right), \right. \\ & \qquad \qquad \qquad \left. E_1 > t, B, D \mid R_1 = i_1, \dots, R_{k-1} = i_{k-1}\right) \\ &= \frac{1}{(n-1)^{k-1}} \sum_{2 \leq i_1, \dots, i_{k-1} \leq n}^* P\left(\sigma_k > t - \left(\sum_{i=1}^{k-1} s_i\right), \right. \\ & \qquad \qquad \qquad \left. E_1 > t, B, D \mid R_1 = i_1, \dots, R_{k-1} = i_{k-1}\right) + o_n(1) \\ &= P\left(\sigma_k > t - \left(\sum_{i=1}^{k-1} s_i\right), E_1 > t, X_2^n(s_{1-}) \neq 0, \dots, X_k^n\left(\left(\sum_{i=1}^{k-1} s_i\right)\right) \right. \\ & \qquad \qquad \qquad \left. \neq 0, D \mid R_1 = 2, \dots, R_{k-1} = k\right) + o_n(1), \end{aligned}$$

where \sum^* denotes the sum over distinct i_1, \dots, i_{k-1} . Here, in writing the first equation, we observed that there are, at most, $\binom{k-1}{2} (n-1)^{k-2}$ choices of $2 \leq i_1, \dots, i_{k-1} \leq n$ that are not distinct, and in writing the second equation we used this and rewrote B given the conditioning.

It helps now to have a graphical representation of the process (see Fig. 5). Each queue has an associated horizontal time line. Queue i has three associated Poisson processes: an arrival process $A(i)$ of rate ν , a departure process $D(i)$ of rate 1, and an impatience process $I(i)$ of rate β . All Poisson processes in sight are independent. At the times of $A(i)$, we write an α on the time line of queue i , and at the times of $D(i)$ we write a δ . At the times of $I(i)$ we draw a vertical arrow from queue i to the queue that would be checked by the second customer at queue i , if any. This checked queue is determined by i.i.d. random variables $(R_n(i), n \geq 1)$, which are also independent of the Poisson processes. Clearly the process $(X^n(t), t \geq 0)$ can be constructed from any initial condition using this graphical representation, interpreting δ as virtual departure, α as arrival, and vertical arrows as virtual transfers due to impatience.

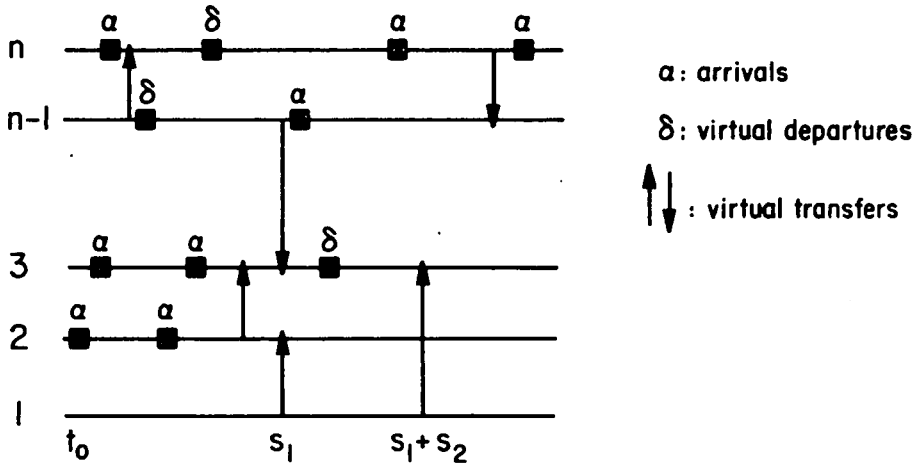


FIGURE 5. Representation of $X^n(t)$.

Consider the event $K = \{\text{tracing back from } X_2^n(s_{1-}), \dots, X_k^n((\sum_{i=0}^{k-1} s_i)_-)$ does not involve queue 1}. Here tracing backward means determining all possible initial conditions of all the queues that are consistent with any given set of values for $X_2^n(s_{1-}), \dots, X_k^n((\sum_{i=0}^{k-1} s_i)_-)$. This can be done by following those time lines that are impinged upon by or impinge upon the existing time lines as we work our way backward in time. We say that tracing backward does not involve queue 1 if the time line of queue 1 is never involved.

It is easy to see that $P(K | R_1 = 2, \dots, R_{k-1} = k) \rightarrow 1$ as $n \rightarrow \infty$. Hence, the integrand in the definition of L_k can be written as

$$P\left(\sigma_k > t - \left(\sum_{i=1}^{k-1} s_i\right), E_1 > t, X_2^n(s_{1-}) \neq 0, \dots, X_k^n\left(\left(\sum_{i=1}^{k-1} s_i\right)_-\right) \neq 0, D | K, R_1 = 2, \dots, R_{k-1} = k\right) + o_n(1),$$

but conditional on K , the evolution of queues $2, \dots, k$ is the same as in an $n - 1$ queue system. So the preceding integrand equals

$$P\left(\sigma_k > t - \left(\sum_{i=1}^{k-1} s_i\right), D\right) (1 - \mu_{s_1}(0)) \dots (1 - \mu_{s_1 + \dots + s_{k-1}}(0)) \exp(-t) + o_n(1).$$

Putting the pieces together, it is easy to see that τ_n^1 converges weakly to R where $P(R > s + ds | R > s) = (1 + \beta \mu_s(0)) ds + o(ds)$. Here the vanishing of the sum of the $o_n(1)$ terms follows from the Poisson tail

on the number of virtual attempts made by queue 1 before time t . From this the result follows.

- (ii) We can carry the previous argument here, and in this case the result follows using Theorem 5.1. ■

Remark: The theorem was stated as above to simplify the notation in the proof. A similar theorem can be easily proved for more general initial conditions such as those in Theorem 4.1 and then for completely general initial distributions by taking a suitable mixture (see Remark 1 following Theorem 4.2).

Using a similar proof technique, one can write down expressions for the asymptotic sojourn time distributions in networks with more complicated dynamics. For example, consider a network of $M/M/C/K$ queues with customers of two classes, c_1 and c_2 . At each queue customers of class c_1 arrive according to a Poisson process of rate ν , and they are rejected if the queue is full. A customer in a queue that is not in service becomes impatient at rate β , checks one of the other queues at random, and if there is a free server at the queue, it enters service as a customer of class c_2 . Customers of class c_1 require $\exp(1)$ service, and customers of class c_2 require $\exp(\alpha)$ service, where $\alpha < 1$. This is a crude model for dynamic routing with queuing for virtual circuits and impatience, the idea being that alternately routed calls take longer to get through the system. Blocking probabilities can be determined from the asymptotic ODE in a manner similar to earlier results. Asymptotic sojourn time distributions can be calculated by a proof technique identical to that earlier. Using these one can study the trade-offs involved in allowing queuing with impatience for virtual circuits.

Let us describe the limiting expression for the sojourn time distribution of a typical customer. The n queue system can be described by a Markov chain $(X^n(t), t \geq 0)$ taking values in S^n , where $S = \{(l_1, l_2), l_1 + l_2 \leq K, l_2 \leq C\}$, where a state (l_1, l_2) denotes that there are l_1 customers of class c_1 and l_2 customers of class c_2 in the link. Suppose the system is started with an initial distribution μ_n that is symmetric and satisfies $\lim_{n \rightarrow \infty} \mu_n|_m = \mu^{\otimes m}$ for each $m \geq 1$. Let μ_s denote the solution to the ODE starting with initial distribution μ . Then, conditional on the customer seeing a free server on arrival, its sojourn time distribution is, of course, $\exp(1)$. Conditional on the customer seeing k_1 customers of class c_1 and $k_2 = C - k_1$ customers of class c_2 in service and $0 \leq l < K - C$ customers (necessarily of class c_1) ahead of it waiting for service, the customer's limiting sojourn time distribution is described as follows: Run two clocks of time-varying rates $\beta \sum_{l_1+l_2 < C} \mu_s(l_1, l_2)$ and $k_1 + k_2\alpha + l\beta \sum_{l_1+l_2 < C} \mu_s(l_1, l_2)$, respectively. If the first clock runs out before the second, add on an $\exp(\alpha)$ distribution and this is the sojourn time. If the second clock runs out before the first at time τ , restart it with rate $k_1 + k_2\alpha + (l-1)\beta \sum_{l_1+l_2 < C} \mu_s(l_1, l_2)$ with probability $(k_1 + l\beta \sum_{l_1+l_2 < C} \mu_\tau(l_1, l_2)) (k_1 + k_2\alpha + l\beta \sum_{l_1+l_2 < C} \mu_\tau(l_1, l_2))^{-1}$ and rate $k_1 + 1 + (k_2 - 1)\alpha + (l-1)\beta \sum_{l_1+l_2 < C} \mu_s(l_1, l_2)$ with probability $(k_2\alpha) (k_1 + k_2\alpha + l\beta \sum_{l_1+l_2 < C} \mu_\tau(l_1, l_2))^{-1}$, and so on. If the customer

enters service via the second clock running out before the first when there are no waiting customers, then we add on an $\exp(1)$ time to determine the sojourn time distribution.

7. CONCLUDING REMARKS

Our purpose in this paper has been to discuss a technique suggested by some statistical mechanical ideas that can be used to compute sojourn time distributions of typical customers in various models of networks of queues with interaction, where there is suitable symmetry in the dynamics. The models discussed in this paper are only crude approximations of reality. Nevertheless, we feel that there is some purpose served in becoming aware of the possibility of using propagation of chaos to argue the independence of the behavior over time of any given collection of elements in a large network. We think of the approach here as a functional version of the Erlang fixed-point approximation, which bears a relation to the usual Erlang fixed-point approximation analogous to that which ODE limits bear to the usual law of large numbers. Hence, the approach should be particularly useful in situations where the Erlang fixed-point equations have proved to be adequate.

References

1. Anantharam, V. (1991). A mean field limit for a lattice caricature of dynamic routing in circuit switched networks. *Annals of Applied Probability* 1: 481–503.
2. Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.
3. Ethier, S.N. & Kurtz, T.G. (1986). *Markov processes: Characterization and convergence*. New York: Wiley.
4. Gibbens, R.J., Hunt, P.J., & Kelly, F.P. (1990). Bistability in communication networks. In G.R. Grimmett & D.J.A. Welsh (eds.), *Disorder in physical systems*. Oxford University Press, pp. 113–128.
5. Hajek, B. & Krishna, T. (1990). Bounds on the accuracy of reduced-load blocking formula in some simple circuit-switched networks. *Proceedings of the Bilkent University Conference on Communication, Control and Signal Processing*. Ankara, Turkey: Bilkent University, pp. 1633–1642.
6. Hartman, T. (1964). *Ordinary differential equations*. New York: Wiley.
7. Hunt, P.J. (1990). Limit theorems for stochastic networks. Ph.D. thesis, University of Cambridge.
8. Kac, M. (1956). Foundations of kinetic theory. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, pp. 154–161.
9. Kelly, F.P. (1986). Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability* 18: 473–505.
10. Kelly, F.P. (1988). Routing in circuit-switched networks: Optimization, shadow price and decentralization. *Advances in Applied Probability* 20: 112–144.
11. Sznitman, A.-S. (1989). Propagation of chaos. *Ecole d'Été de Probabilités de Saint-Flour*, Lecture Notes in Math. No. 1464, Berlin: Springer (preprint).
12. Uchiyama, K. (1986). Fluctuation in population dynamics. *Lecture Notes in Biomathematics* 70: 222–229.

13. Whitt, W. (1985). Blocking when service is required from several facilities simultaneously. *AT&T Technical Journal* 64: 1807–1856.
14. Ziedins, I. & Kelly, F.P. (1989). Limit theorems for loss networks with diverse routing. *Advances in Applied Probability* 21: 804–830.

APPENDIX

Weak Convergence of Markov Processes

In this appendix we recapitulate a few results from the weak convergence theory of Markov processes that are necessary for our proof of Theorem 3.1. Our main reference for these results is Ethier and Kurtz [3].

A semigroup $\{T(t)\}$ of bounded linear operators on a Banach space L (with norm $\|\cdot\|$) is said to be strongly continuous if $\lim_{t \rightarrow 0} T(t)f = f$ for every $f \in L$; it is said to be a contraction semigroup if $\|T(t)\| \leq 1$.

A (possibly unbounded) linear operator is a linear mapping whose domain $\mathfrak{D}(A)$ is a subspace of L and whose graph is given by $\mathfrak{G}(A) = \{(f, Af) : f \in \mathfrak{D}(A)\}$. A is said to be closed if $\mathfrak{G}(A)$ is a closed subspace of $L \times L$.

The generator of a semigroup $\{T(t)\}$ on L is the linear operator A defined by

$$Af = \lim_{h \rightarrow 0} \frac{T(h)f - f}{h}.$$

The domain of A , $\mathfrak{D}(A)$ is the collection of functions $f \in L$ for which the preceding limit exists.

A linear operator A on L is said to be closable if it has a closed linear extension. If A is closable, then the closure \bar{A} of A is the minimal closed linear extension of A ; more specifically, it is the closed linear operator B whose graph is the closure in $(L \times L)$ of the graph of A . If A is closed, a subspace D of $\mathfrak{D}(A)$ is said to be a core for A if the closure of the restriction of A to D is equal to A . We have the following lemma [3, Corollary 1.6, p. 10].

LEMMA A.1: *If A is the generator of a strongly continuous contraction semigroup $\{T(t)\}$ on L , then $\mathfrak{D}(A)$ is dense in L and A is closed.*

The next two lemmas are Proposition 3.3 and Theorem 6.1 in Ethier and Kurtz [3, pp. 17 and 28, respectively].

LEMMA A.2: *Let A be the generator of a strongly continuous contraction semigroup $\{T(t)\}$ on L . Let D_0 and D be dense subspaces of L with $D_0 \subset D \subset \mathfrak{D}(A)$. If $T(t) : D_0 \rightarrow D$ for all $t \geq 0$, then D is a core for A .*

LEMMA A.3: *For $n = 1, 2, \dots$, let $\{T_n(t)\}$ and $\{T(t)\}$ be a strongly continuous contraction semigroup on the Banach space L with generators A_n and A . Let D be a core for A . Then the following are equivalent:*

- (i) For each $f \in L$, $T_n(t)f \rightarrow T(t)f$ for all $t \geq 0$.
- (ii) For each $f \in D$, there exists $f_n \in \mathcal{D}(A_n)$ for each $n \geq 1$ such that $f_n \rightarrow f$ and $A_n f_n \rightarrow Af$.

A linear operator A is said to be dissipative if $\|\lambda f - Af\| \geq \lambda \|f\|$ for every $f \in \mathcal{D}(A)$ and $\lambda > 0$. An operator A satisfies the positive maximum principle if whenever $f \in \mathcal{D}(A)$, $x_0 \in E$, and $\sup_{x \in E} f(x) = f(x_0) \geq 0$, we have $Af(x_0) \leq 0$. An operator on $C(E)$ satisfying the positive maximum principle is dissipative [3, Lemma 2.1, p. 165].

The following lemma follows from Ethier and Kurtz [3, Theorem 2.6, p. 13].

LEMMA A.4: A bounded operator B that is dissipative generates a strongly continuous contraction semigroup.

The following perturbation result follows from Ethier and Kurtz [3, Theorem 7.1, p. 37].

LEMMA A.5: Let A be a linear operator such that \bar{A} generates a strongly continuous contraction semigroup $\{T(t)\}$ on L , and let \bar{B} be a bounded linear operator on L such that $\{e^{t\bar{B}}\}$ is a contraction semigroup. Then $\overline{A + B}$ generates a strongly continuous contraction semigroup $\{S(t)\}$ on L . Moreover, $\overline{A + B} = \bar{A} + \bar{B}$.

A semigroup $\{T(t)\}$ on L is positive if $T(t)f \geq 0$ for all $f \geq 0$. It is conservative if $T(t)1 = 1$ where 1 is the identity. If E is a complete separable metric space, any strongly continuous contraction semigroup on $C(E)$ that is positive and conservative is called a Feller semigroup.

Let $\{T(t)\}$ be a semigroup on a closed subspace $L \subset B(E)$. We say that an E -valued Markov process X corresponds to $\{T(t)\}$ if

$$E[f(X(t + s)) | X(u), u \leq s] = T(t)f(X(s))$$

for all $s, t \geq 0$ and $f \in L$. Then we have the following lemma [3, Proposition 1.6, p. 161].

LEMMA A.6: Let E be a complete separable metric space. Let X be an E -valued Markov process with initial distribution ν corresponding to a semigroup $\{T(t)\}$ on $C(E)$. Then $\{T(t)\}$ and ν determine the finite dimensional distributions of X .

Finally we have the following [3, Theorem 2.5, p. 167].

LEMMA A.7: Let E be a compact metric space. For $n = 1, 2, \dots$, let $\{T_n(t)\}$ be a Feller semigroup on $C(E)$, and suppose X_n is a Markov process corresponding to $\{T_n(t)\}$ with sample paths in $D_E[0, \infty)$. Suppose that $\{T(t)\}$ is a Feller semigroup on $C(E)$ and that for each $f \in C(E)$:

$$\lim_{n \rightarrow \infty} T_n(t)f = T(t)f \quad \text{for } t \geq 0.$$

If $X_n(0)$ has limiting distribution $\nu \in M(E)$, then there is a Markov process X corresponding to $\{T(t)\}$ with initial distribution ν and sample paths in $D_E[0, \infty)$ and $X_n \Rightarrow X$.