A new method for analysing the equilibrium and time-dependent behaviour of Markovian models *

P.K. Pollett
M.R. Thompson
Department of Mathematics, The University of Queensland, Brisbane, Australia.

Abstract

Many large-scale stochastic systems, such as telecommunications networks, can be modelled using a continuous-time Markov chain. However, it is frequently the case that a satisfactory analysis of their time-dependent, or even equilibrium behaviour, is impossible. In this paper we propose a new method of analyzing Markovian models, whereby the existing transition structure is replaced by a more amenable one. Using rates of transition given by the equilibrium *expected rates* of the corresponding transitions of the original chain, we are able to approximate its behaviour. We present two formulations of the idea of expected rates. The first provides a method for analysing time-dependent behaviour, while the second provides a highly accurate means of analysing equilibrium behaviour. We shall illustrate our approach with reference to a variety of models, giving particular attention to queueing and loss networks.

1 Introduction

In order to understand the rationale of expected rates, consider any large-scale stochastic system whose natural state description is Markovian, yet its equilibrium or time-dependent behaviour is difficult to analyze. The system in question might be a communications network, whose state records the numbers of calls on the various routes through the network (each call using resources at several communications links). The idea is to find an alternative state description, together with an approximating transition structure, which can be analyzed more simply. An alternative description for the communications network might focus on the links, rather than the routes, say recording the resource usage on those links. Since each link will usually service several different routes, this description is unlikely to be Markovian. However, we might usefully approximate the behaviour of the network by considering the links in isolation, and model the resource usage on any given link by a Markov chain whose rates of transition are given by the equilibrium expected rates of the corresponding transitions of the original chain.

We begin by proposing a formulation of the idea of expected rates, which is similar to one suggested to us by Peter Taylor as a "generalized reduced load concept". We illustrate how it applies to the study of time-dependent behaviour in Markovian queueing networks. We shall see that the resulting approximations may be accurate for only a limited range of parameter values. We will then describe a variant of the basic idea which offers much greater promise, for it encapsulates several recent methods for analyzing loss networks that are known to be highly accurate in a wide variety of circumstances. Our method is appropriate for estimating equilibrium quantities in Markovian models for which there is no appropriate product-form equilibrium distribution.

2 Expected rates

Let $(X(t),\ t \geq 0)$ be a continuous-time Markov chain over a denumerable state space S with transition rates $Q = (q(x,y),\ x,y \in S)$, where for simplicity $q(x,x) = 0$, and set $q(x) = \sum_{y \in S} q(x,y)$.

*This work was funded by the Australian Research Council (Grant No. A00104575).

Suppose that S is irreducible (and hence $q(x) > 0$ for all $x \in S$) and positive recurrent, and let $\pi = (\pi(x),\ x \in S)$ be the unique equilibrium distribution of the chain. Thus,

$$\sum_{x \in S} \pi(x) q(x,y) = \pi(y) q(y), \qquad y \in S.$$

Now let $(X_n,\ n = 0, 1, \dots)$ be the jump chain, that is, the discrete-time Markov chain over S, with $X_0 = X(0)$, that records the sequence of states visited. Note that S is also irreducible for the jump chain because its transition probabilities $P = (p(x,y),\ x, y \in S)$ are given by $p(x,y) = q(x,y)/q(x)$. We shall assume that $\sum_{x \in S} \pi(x) q(x) < \infty$, so that the jump chain admits an equilibrium distribution $m = (m(x),\ x \in S)$ given by $m(x) = \pi(x) q(x) / \sum_{y \in S} \pi(y) q(y)$, $x \in S$; see Exercise 1.1.5 of Kelly [9]. Notice that m coincides with π *only* when $q(x)$ is the same for all x, and so only in this exceptional case can the two chains be stationary together.

Now identify a set of transitions $A \subseteq \tilde{S}$, where $\tilde{S} = S \times S$, and define $r(A)$ by

$$r(A) = \mathbb{E}_m \left(q(X_n, X_{n+1}) | (X_n, X_{n+1}) \in A \right), \tag{1}$$

where $\mathbb{E}_m(\cdot)$ denotes expectation with respect to the distribution m. Thus, $r(A)$ is the equilibrium expected rate of transition, given that the transition is in A. Notice that $r(A)$ does not depend on n because, under m, (X_n, X_{n+1}) forms a stationary sequence. Indeed, this sequence is a Markov chain with transition probabilities $p((u,x);(x,y)) = p(x,y)$ and equilibrium distribution $m(x,y) = m(x)p(x,y)$; see Proposition 2.1 of Kelly and Pollett [10]. We can evaluate $r(A)$ as follows:

$$\begin{aligned} r(A) &= \sum_{(x,y) \in A} q(x,y) \frac{\Pr(X_n = x, X_{n+1} = y)}{\sum_{(u,v) \in A} \Pr(X_n = u, X_{n+1} = v)} \\ &= \sum_{(x,y) \in A} q(x) p(x,y) \frac{m(x)p(x,y)}{\sum_{(u,v) \in A} m(u)p(u,v)} \\ &= \frac{\sum_{(x,y) \in A} q(x) m(x) p(x,y)^2}{\sum_{(x,y) \in A} m(x) p(x,y)} = \frac{\sum_{(x,y) \in A} \pi(x) q(x,y)^2}{\sum_{(x,y) \in A} \pi(x) q(x,y)} . \end{aligned} \tag{2}$$

Remark It is natural for the expectation in (1) to be taken with respect to the equilibrium distribution of the jump chain. However, if it had been taken with respect to π, thus giving a quantity $r_n(A)$ which would generally depend on n, we would have $r_n(A) \to r(A)$ as $n \to \infty$, at least formally, since because the jump chain is aperiodic, $\Pr(X_n = x) \to m(x)$.

Example 1 To illustrate how expected rates can be evaluated, consider the $M/M/1$ queue. It has Poisson arrivals at rate $\alpha > 0$, independent exponentially distributed service times with unit mean, and a single server operating at rate $\phi > 0$, serving customers one at a time in the order in which they arrive. With the state $X(t)$ representing the number of customers in the system at time t, we have $S = \{0, 1, \dots\}$, $q(x, x+1) = \alpha$ and $q(x, x-1) = \phi$ for $x \geq 1$, with all other transition rates equal to 0. Define transitions $A = \{(x, x+1),\ x = 0, 1, \dots\}$ and $D = \{(x, x-1),\ x = 1, 2, \dots\}$, corresponding to arrivals and departures, respectively. Assuming that the traffic intensity $\rho = \alpha/\phi$ is strictly less than 1, an equilibrium distribution exists for both $X(t)$ and its jump chain, and, $\pi(x) = (1-\rho)\rho^x$. Using (2) we find, perhaps not unexpectedly, that $r(A) = \alpha$ (notice that, more generally, if $q(x,y) = \alpha$ for $x, y \in A$, so that transitions in A form a Poisson process with rate α, then $r(A) = \alpha$). We also find that $r(D) = \phi$, but note that the equilibrium expected departure rate is $\mathbb{E}_\pi(\phi 1_{\{X(t)>0\}}) = \phi \Pr(X(t) > 0) = \alpha$, which is different from $r(D)$. We see that the expected rates approximation for this simple queueing system is the *same* as the original system.

3 Markovian queueing networks

In this section we shall study a network of queues, with the queues labelled $1, \dots, J$. If customers can enter or leave the network, it is said to be *open*. In this case customers arrive at queue i from outside the network as a Poisson stream with rate ν_i (if $\nu_i = 0$ there is no exogenous arrival

process at that queue). Otherwise, a fixed number N of customers circulate, and the network is said to be *closed*. After completing service at queue i, a customer either leaves the network, with probability λ_{i0}, or proceeds to another queue j, with probability λ_{ij} (in the closed case we take $\lambda_{i0} = 0$). For simplicity, we shall assume that $\lambda_{ii} = 0$. Clearly $\sum_j \lambda_{ij} = 1$. We shall assume that these parameters are chosen so that a customer can reach any queue from anywhere in the network. In the open case we shall also assume that a customer can reach any queue from outside the network and eventually leave the network starting from anywhere. In the closed case these conditions ensure that the routing matrix (λ_{ij}) is irreducible and, hence, that there is a unique collection $(\alpha_1, \alpha_2, \ldots, \alpha_J)$ of strictly positive numbers which satisfy the *traffic equations* $\alpha_j = \sum_i \alpha_i \lambda_{ij}$, $j = 1, 2, \ldots, J$. Here we may assume without loss of generality that $\sum_j \alpha_j = 1$. In the open case these conditions ensure that there is a unique positive solution $(\alpha_1, \alpha_2, \ldots, \alpha_J)$ to the equations $\alpha_j = \nu_j + \sum_i \alpha_i \lambda_{ij}$, $j = 1, 2, \ldots, J$. In this case α_j is the arrival rate at queue j, while in the closed case α_j is *proportional to* the arrival rate at queue j. Service times of customers at the various queues in the network are assumed to be independent exponentially distributed random variables with unit mean, and independent of the arrival and routing processes. When there are n customers at a given queue j, a service effort of $\phi_j(n)$ is offered. We shall assume that $\phi_j(0) = 0$ and $\phi_j(n) > 0$ whenever $n \geq 1$. For example, when $\phi_j(n) = \phi_j n$, every customer at queue j gets the *same* service effort ϕ_j (the infinite-server queue), while if $\phi_j(n) = \phi_j \min\{n, s_j\}$, for $n \geq 1$, then at most s_j customers receive service, each at the same rate ϕ_j (the s_j-server queue). In this latter case $\rho_j = \alpha_j/(\phi_j s_j)$ is called the *traffic intensity* at queue j.

We have described the basic migration process of Whittle [21] (see also Whittle [22]), a special case of which was considered first by Jackson [8]; for further details see Chapter 2 of Kelly [9]. The equilibrium behaviour of these networks is well understood, and summarized in Theorems 2.3 and 2.4 of Kelly [9]. The network can be described by a continuous-time Markov chain with state $\boldsymbol{n} = (n_1, n_2, \ldots, n_J)$, where n_j is the number of customers at queue j (including those in service). In the open case $S = Z_+^J$ and the transition rates are given by

$$\begin{aligned} q(\boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_j) &= \nu_j && \text{(external arrival at queue } j\text{)} \\ q(\boldsymbol{n}, \boldsymbol{n} - \boldsymbol{e}_i) &= \lambda_{i0}\phi_i(n_i) && \text{(external departure at queue } i\text{)} \\ q(\boldsymbol{n}, \boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j) &= \lambda_{ij}\phi_i(n_i) && \text{(movement from queue } i \text{ to queue } j\text{)}\,, \end{aligned}$$

where $\boldsymbol{e}_j$ is the unit vector in Z_+^J with a 1 as its j-th entry. An equilibrium distribution exists if $b_j^{-1} := 1 + \sum_{n=1}^{\infty} (\alpha_j^n / \prod_{r=1}^n \phi_j(r)) < \infty$ for all j, in which case

$$\pi(\boldsymbol{n}) = \prod_{j=1}^{J} \pi_j(n_j), \quad \text{where} \quad \pi_j(n) = b_j \, \frac{\alpha_j^n}{\prod_{r=1}^n \phi_j(r)}. \tag{3}$$

Thus, in equilibrium, $n_1, n_2, \ldots, n_J$ are *independent* and each queue j behaves *as if* it were isolated with Poisson input at rate α_j.

In the closed case S $(= S_N)$ is the finite subset of Z_+^J with $\sum_j n_j = N$, where recall that N is the total number of customers in the network. The transition rates are now simply

$$q(\boldsymbol{n}, \boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j) = \lambda_{ij}\phi_i(n_i) \qquad \text{(movement from queue } i \text{ to queue } j\text{)}\,.$$

An equilibrium distribution always exists and is given by

$$\pi(\boldsymbol{n}) \; (= \pi_N(\boldsymbol{n})) \; = B_N \prod_{j=1}^{J} \frac{\alpha_j^{n_j}}{\prod_{r=1}^{n_j} \phi_j(r)}\,,$$

where B_N is a normalizing constant chosen so that π sums to 1 over S_N.

There are very few explicit results concerning the time-dependent behaviour of these networks, and a product form such as (3) is exhibited rather rarely by the transient distribution. Indeed, Boucherie and Taylor [4] have shown that the networks with all queues being $\cdot/M/\infty$ are the *only* ones with a transient product-form distribution (among a much larger class of Markovian networks than the ones considered here). We propose the following approximation using expected rates.

Define

$$\begin{aligned} A_k(m) &= \{(\boldsymbol{m},\boldsymbol{n}) \in \tilde{S} : m_k = m, n_k = m+1\}, \qquad m \geq 0, \\ D_k(m) &= \{(\boldsymbol{m},\boldsymbol{n}) \in \tilde{S} : m_k = m, n_k = m-1\}, \qquad m \geq 1, \end{aligned} \tag{4}$$

where recall that $\tilde{S} = S \times S$. These represent, respectively, an arrival and a departure transition at queue k when there are m individuals at that queue, and so $a_k(m) = r(A_k(m))$ and $d_k(m) = r(D_k(m))$ will give the expected (state-dependent) arrival and departure rates for queue k under the equilibrium distribution of the jump chain. We propose to approximate the behaviour of the network by a system of isolated queues, with each queue k modelled as a birth-death process with birth rates $q_k(m, m+1) = a_k(m)$, $m \geq 0$, and death rates $q_k(m, m-1) = d_k(m)$, $m \geq 1$.

On summing $m(\boldsymbol{m})q(\boldsymbol{m},\boldsymbol{n})$ and $m(\boldsymbol{m})q(\boldsymbol{m},\boldsymbol{n})^2$ over $(\boldsymbol{m},\boldsymbol{n})$ in $A_k(m)$ and in $D_k(m)$, we find that the expected rates can be expressed in terms of π. In the open case $a_k(m) = a_k$ is the same for all m and given by

$$a_k = \frac{1}{\alpha_k}\left(\nu_k^2 + \sum_j \alpha_j \lambda_{jk}^2 \sum_{n=0}^{\infty} \pi_j(n)\phi_j(n+1)\right),$$

while $d_k(m) = d_k\phi_k(m)$, where $d_k = \lambda_{k0}^2 + \sum_j \lambda_{kj}^2$. In the closed case

$$a_k(m) = \frac{\sum_j \alpha_j \lambda_{jk}^2 \mathbb{E}_{\pi_{N-1}}\left(\phi_j(n_j+1)1_{\{n_k=m\}}\right)}{\alpha_k \mathrm{Pr}_{\pi_{N-1}}(n_k = m)},$$

where $\mathrm{Pr}_{\pi_{N-1}}(n_k = m)$ is the equilibrium probability that there are m customers at queue k in a network with $N-1$ customers circulating, while $d_k(m) = d_k\phi_k(m)$, where $d_k = \sum_j \lambda_{kj}^2$. For both the open and closed cases $d_k(m)$ is given explicitly. In contrast, the expected arrival rates can be evaluated explicitly only in special cases. For the open network, if $\phi_j(n) = \phi_j$ for $n \geq 1$ ($\phi_j(0) = 0$), then $\sum_n \pi_j(n)\phi_j(n+1) = \phi_j$, while if $\phi_j(n) = \phi_j n$, then $\sum_n \pi_j(n)\phi_j(n+1) = \alpha_j + \phi_j$. Thus if $\phi_j(n) = \phi_j$ for every j, $a_k = (\nu_k^2 + \sum_j \phi_j \alpha_j \lambda_{jk}^2)/\alpha_k$, while if $\phi_j(n) = \phi_j n$ for every j, then $a_k = (\nu_k^2 + \sum_j \alpha_j(\alpha_j + \phi_j)\lambda_{jk}^2)/\alpha_k$. For the closed network, if $\phi_j(n) = \phi_j$ for every j, then $a_k(m) = (\sum_j \phi_j \alpha_j \lambda_{jk}^2)/\alpha_k$, while if $\phi_j(n) = \phi_j n$ for every j, then

$$a_k(m) = \frac{1}{\alpha_k}\left(\sum_j \phi_j \alpha_j \lambda_{jk}^2 + \frac{\phi_k C(N-1-m)}{\phi_k - C\alpha_k} \sum_j \alpha_j^2 \lambda_{jk}^2\right),$$

where $C^{-1} = \sum_j (\alpha_j/\phi_j)$.

Example 2 In order to assess the accuracy of our method, we shall examine a network for which there are explicit results for describing time-dependent behaviour. The network we shall consider has $\phi_j(n) = \phi_j n$ for every j, so that each queue has infinitely many servers, with the servers at queue j operating at rate ϕ_j. Explicit results exist, provided we assume that the network is completely empty at time 0. Kingman [14] showed that if $n_j(0) = 0$ for all j, then, for every $t > 0$, $n_1(t), n_2(t), \ldots, n_J(t)$ are *independent* Poisson random variables with $n_j(t)$ having mean $\mu\beta_j(t)$, where $\mu = \sum_j \nu_j$ is the total exogenous arrival rate, and $\beta_j(t) = \int_0^t p_j(t)dt$, where $p_j(t)$ is the probability that an individual, entering the network at time 0, is in queue j after time t. For details, see Theorem 4.2 of Kelly [9]. (Note that, by Fubini's theorem, $\beta_j(t)$ is the expected total time the single individual spends in queue j up to time t.) Kingman's result holds in much greater generality than might be indicated by the present context. For the particular Markovian network in question, $p_j(t)$ can be evaluated further.

If an individual arrives at the network at time $t = 0$, then he will enter queue j with probability $p_j(0) = \nu_j/\mu$. Define $p_0(t) = 1 - \sum_j p_j(t)$ to be the probability that the individual has left the network by time t, and note that $p_0(0) = 0$. The movement of the individual through the network can be thought of as a random walk in continuous time on the set of indices $\{0, 1, 2, \ldots, J\}$, recording his present location, with 0 (an absorbing state) indicating that he has left the network. Note that, under the conditions we have imposed, $\{1, 2, \ldots, J\}$ is an irreducible class for the random

walk. Since service times are exponentially distributed with mean 1 and the service rate at queue j is ϕ_j, the rate at which the individual moves from queue i to queue j is $r_{ij} = \phi_i \lambda_{ij}$, and, from queue i to the outside, the rate is $r_{i0} = \phi_i \lambda_{i0}$. Therefore $p_j(t)$, $j = 0, 1, \ldots, J$, satisfies a set of forward equations $p_j{}'(t) = \sum_{i=0}^{J} p_i(t) r_{ij}$, $j = 0, 1, \ldots, J$, where, for $j = 1, 2, \ldots, J$, $r_{jj} = -\phi_j$ and $r_{0j} = r_{00} = 0$. These integrate to give

$$\mu p_j(t) = \nu_j e^{-\phi_j t} + \sum_{i=1}^{J} \int_0^t \mu p_i(u) e^{-\phi_j (t-u)} du \, \phi_i \lambda_{ij}, \qquad j = 1, \ldots, J,$$

remembering that $\lambda_{jj} = 0$, and $p_0(t) = \sum_{i=1}^{J} \int_0^t p_i(u) du \, \phi_i \lambda_{i0}$. Recall that $n_j(t)$, $j = 1, 2, \ldots, J$, are independent Poisson random variables with $\mathbb{E}(n_j(t)) = \int_0^t \mu p_j(s) ds$.

The expected rates approximation has each queue isolated, with queue k being an infinite-server queue having arrival rate $a_k = (\nu_k^2 + \sum_j \alpha_j(\alpha_j + \phi_j)\lambda_{jk}^2)/\alpha_k$, and each of its servers operating at rate $b_k = d_k \phi_k$, where $d_k = \lambda_{k0}^2 + \sum_j \lambda_{kj}^2$. Thus $n_k(t)$ is a Poisson random variable with mean $(a_k/b_k)(1 - \exp(-b_k t))$. The expected rates approximation is given explicitly, but how accurate is it?

We shall specialize to the case of a symmetric network, for which the mean of $n_j(t)$ can be evaluated explicitly. Suppose that $\nu_j = \nu$ and $\lambda_{j0} = \lambda_0$ for each j, and that $\lambda_{ij} = (1 - \lambda_0)/(J - 1)$ for $j \neq i$, so that traffic equations have the solution $\alpha_j = \nu/\lambda_0$ (which does not depend on J). Suppose that $\phi_j = \phi$ is also the same for all j. Then, $p_0(t) = 1 - e^{-\phi\lambda_0 t}$ and, for $j = 1, 2, \ldots, J$, $p_j(t) = (1/J) e^{-\phi\lambda_0 t}$, so that $n_j(t)$ is a Poisson random variable with

$$\mathbb{E}(n_j(t)) = \frac{\nu}{\phi\lambda_0} \left(1 - \exp(-\phi\lambda_0 t)\right).$$

This should be compared with the expected rates approximation, which gives

$$\mathbb{E}(n_j(t)) \simeq \left(\frac{\nu}{\phi\lambda_0} + \frac{(1-\lambda_0)^2}{(J-1)\lambda_0^2 + (1-\lambda_0)^2} \right) \left(1 - \exp\left(-\phi \left(\lambda_0^2 + \frac{(1-\lambda_0)^2}{J-1} \right) t \right) \right).$$

We would expect this approximation to be accurate only in cases there the network is large, or when λ_0 is close to 1. This is illustrated in Figure 1, where the relative error in approximating $\mathbb{E}(n_j(10))$ is plotted for a network with $\nu = 1.0$ and $\phi = 2.0$. The top pane has $J = 10$ and λ_0 varying from 0.1 to 1.0, while the bottom pane has $\lambda_0 = 0.9$ and J varying from 2 to 100.

4 An alternative approach

One of the drawbacks of defining expected rates in terms of *transitions* (ordered pairs of states) is that the expectation is evaluated with respect to the equilibrium distribution of the jump chain. We now describe an alternative approach, based on the equilibrium distribution π. Returning to the notation of Section 2, define $A(x) = \{y \in S : (x, y) \in A\}$, $x \in S$, for any particular set $A \subseteq \tilde{S}$ of transitions, let $\underline{A} = \{x \in S : (x, y) \in A \text{ for some } y \in S\}$, and, in place of (1), let

$$r(A) = \mathbb{E}_\pi \Big(q\big(X(t), A(X(t))\big) \,\Big|\, X(t) \in \underline{A} \Big) = \sum_{x \in \underline{A}} q(x, A(x)) \, \frac{\pi(x)}{\pi(\underline{A})},$$

where, for $B \subseteq S$, $q(x, B) = \sum_{y \in B} q(x, y)$ and $\pi(B) = \sum_{y \in B} \pi(y)$. Note that, in contrast to (1), the expectation here is taken with respect to π.

For the Markovian networks studied in the previous section, this approach does not yield anything new. For example, using the transitions defined by (4), and the notation $a_k(m) = r(A_k(m))$ and $d_k(m) = r(D_k(m))$, we find that

$$\begin{aligned} a_k(m) &= \sum_{\boldsymbol{n} \in S : n_k = m} \left(\nu_k + \sum_i \phi_i(n_i) \lambda_{ik} \right) \frac{\pi(\boldsymbol{n})}{\pi_k(m)} \\ &= \nu_k + \sum_{i \neq k} \left(\sum_{\boldsymbol{n} \in S : n_k = m} \phi_i(n_i) \prod_{l \neq k} \pi_l(n_l) \right) \lambda_{ik} = \nu_k + \sum_{i \neq k} \alpha_i \lambda_{ik} = \alpha_k \,, \end{aligned}$$

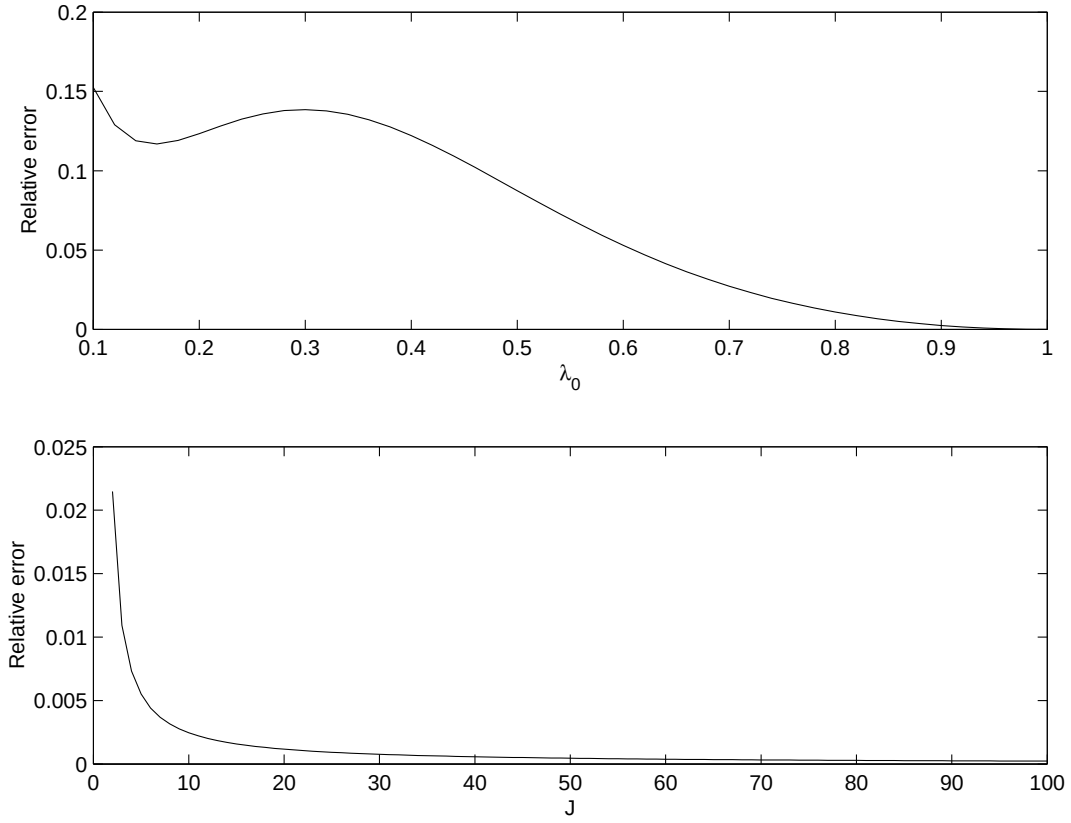

Figure 1: Accuracy of the expected rates approximation for a network of infinite-server queues.

and $d_k(m) = \phi_k(m)$. Hence, a proposal to approximate the behaviour of the network by a system of isolated queues, with each queue j modelled as a birth-death process with birth rates α_j and death rates $\phi_j(n)$, would result in a system whose equilibrium behaviour is the *same* as the original model; only in exceptional circumstances (Example 2) would the transient behaviour be the same. This is a natural approach to analyzing queueing networks; see Kühn [15] for an extension to networks with general service time distributions, which involves the additional matching of higher order moments of the arrival and service processes.

Using a slight variation of this new rationale, we can apply the technique to cases where there is no appropriate product form. The idea is to *impose* a product form π for the equilibrium distribution (or some set of marginal distributions), to evaluate the expected rates using *this* distribution, and then to use these rates to update π. By doing this repeatedly, we would hope to find the product form which best approximates the behaviour of the original model (or some particular quantity of interest, such as a performance measure). We shall illustrate this by looking at an important class of models called *loss networks*.

The basic model describes a circuit-switching network with fixed routing, such as a telephone network, but it also arises in the study of local area networks, multi-processing architectures, database management systems, mobile/cellular radio and broadband packet networks (see Kelly [13] for an excellent review).

The network is composed of communications links, and any route in the network can be expressed as a subset of $\{1, 2, \ldots, J\}$, where J is the number of links. Let $\mathcal{R}$ be the set of all routes. Calls using route r are offered at rate ν_r as a Poisson stream, and use $\lambda_{jr} (\geq 0)$ circuits from link j, the total number of circuits on link j being C_j. We assume that $\mathcal{R}$ indexes independent Poisson processes. Calls requesting route r are blocked and lost if, on *any* link j, there are fewer than λ_{jr} free circuits. Otherwise, the call is connected and simultaneously holds λ_{jr} circuits on each link j for the duration of the call. For simplicity, we shall take $\lambda_{jr} \in \{0, 1\}$. Call durations are independent and identically distributed exponential random variables with unit mean, and are

independent of the arrival processes.

Let $\boldsymbol{n} = (n_r,\ r \in \mathcal{R})$, where n_r is the number of calls in progress using route r, let $\boldsymbol{C} = (C_j,\ j = 1, \ldots, J)$, and let $\boldsymbol{\Lambda} = (\lambda_{jr},\ r \in \mathcal{R},\ j = 1, \ldots, J)$. Then, $(\boldsymbol{n}(t),\ t \geq 0)$ is a continuous-time Markov chain taking values in $S = S(\boldsymbol{C}) = \left\{\boldsymbol{n} \in \mathrm{Z}_+^{\mathcal{R}} \ :\ \boldsymbol{\Lambda n} \leq \boldsymbol{C}\right\}$, with transition rates given by

$$\begin{aligned} q(\boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_r) &= \nu_r, && \text{if } \boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_r \in S, && \text{(call connected on route } r\text{)} \\ q(\boldsymbol{n}, \boldsymbol{n} - \boldsymbol{e}_r) &= n_r, && \text{if } \boldsymbol{n}, \boldsymbol{n} - \boldsymbol{e}_r \in S, && \text{(call cleared on route } r\text{)} \end{aligned}$$

and equal to 0 otherwise; here $\boldsymbol{e}_r$ is the unit vector indicating just one call in progress on route r. It can be shown (see for example Kelly [13]) that the equilibrium distribution is given by $\pi(\boldsymbol{n}) = B \prod_{r \in \mathcal{R}} (\nu_r^{n_r}/n_r!)$, where $B = B(\boldsymbol{C})$ is a normalizing constant chosen so that π sums to 1 over $S(\boldsymbol{C})$. Most of the usual measures of performance of the network can be evaluated in terms of π. For example, the equilibrium probability that a route-r call is blocked is given by $1 - B(\boldsymbol{C})/B(\boldsymbol{C} - \boldsymbol{\Lambda} \boldsymbol{e}_r)$. However, although one has an explicit expression for the blocking probability in terms of B, the latter cannot (usually) be computed in polynomial time (see for example Kelly [12]). Thus, for networks with even moderate capacity, one is forced to use approximation methods.

Define, for each $k \in \{1, 2, \ldots, J\}$, the following sets of transitions:

$$\begin{aligned} A_k(u) &= \{(\boldsymbol{m}, \boldsymbol{n}) \in \tilde{S} : (\boldsymbol{\Lambda m})_k = u, (\boldsymbol{\Lambda n})_k = u + 1\}, \quad u = 0, 1, \ldots, C_k - 1, \\ D_k(u) &= \{(\boldsymbol{m}, \boldsymbol{n}) \in \tilde{S} : (\boldsymbol{\Lambda m})_k = u, (\boldsymbol{\Lambda n})_k = u - 1\}, \quad u = 1, 2, \ldots, C_k. \end{aligned}$$

These comprise all transitions corresponding to an increase, respectively decrease, in the usage on link k when there are u circuits in use on that link. Now, for each $\boldsymbol{m} \in S$ such that $(\boldsymbol{\Lambda m})_k = u$, define $A_k(\boldsymbol{m}, u) = \{\boldsymbol{n} \in S : (\boldsymbol{m}, \boldsymbol{n}) \in A_k(u)\}$, and $D_k(\boldsymbol{m}, u)$ similarly in terms of $D_k(u)$. Then,

$$q(\boldsymbol{m}, A_k(\boldsymbol{m}, u)) = \sum_{r \in \mathcal{R}: k \in r} \nu_r \prod_{i \in r - \{k\}} 1_{\{u_i < C_i\}} 1_{\{u_k = u\}} = \sum_{r \in \mathcal{R}} \lambda_{kr} \nu_r \prod_{i \in r - \{k\}} 1_{\{u_i < C_i\}} 1_{\{u_k = u\}}$$

and $q(\boldsymbol{m}, D_k(\boldsymbol{m}, u)) = \sum_{r \in \mathcal{R}: k \in r} m_r 1_{\{u_k = u\}} = \sum_{r \in \mathcal{R}} \lambda_{kr} m_r 1_{\{u_k = u\}} = u_k 1_{\{u_k = u\}}$, where here we have used the notation $u_i = (\boldsymbol{\Lambda m})_i$ for the number of circuits in use on link i when the state is $\boldsymbol{m}$. Since $q(\boldsymbol{m}, D_k(\boldsymbol{m}, u)) = u_k 1_{\{u_k = u\}}$, we shall always have $d_k(u) := r(D_k(u)) = u$. In order to evaluate $a_k(u) := r(A_k(u))$, we shall impose a product form distribution for $\boldsymbol{u} = (u_1, u_2, \ldots, u_J)$:

$$\pi(\boldsymbol{u}) = \prod_{j=1}^{J} \pi_j(u_j), \quad \text{where} \quad \pi_j(u) = \frac{a_j^u}{u!} \left(\sum_{v=0}^{C_j} \frac{a_j^v}{v!} \right)^{-1}, \quad u = 0, 1, \ldots, C_j. \tag{5}$$

This *would* be the equilibrium distribution for $\boldsymbol{u}$ were the individual links isolated from one another, with calls offered to link j as a Poisson stream with rate a_j, and blocked if they arrive to find C_j circuits in use. (However, this will not be true, except in trivial cases; indeed, under the inherited transition structure $\boldsymbol{u} \to \boldsymbol{u} \pm \boldsymbol{\Lambda} \boldsymbol{e}_r$, $r \in \mathcal{R}$, the process $\boldsymbol{u}(t)$ taking values in $\{\boldsymbol{u} \in \mathrm{Z}_+^J \ : \boldsymbol{u} \leq \boldsymbol{C}\}$ will not generally be Markovian.) Using (5), we find that

$$\begin{aligned} a_k(u) &= \sum_{r \in \mathcal{R}} \lambda_{kr} \nu_r \mathrm{Pr}_\pi \Big(u_i < C_i,\ \forall i \in r - \{k\} \ \Big|\ u_k = u \Big), \\ &= \sum_{r \in \mathcal{R}} \lambda_{kr} \nu_r \prod_{i \in r - \{k\}} (1 - L_i), \end{aligned} \tag{6}$$

where $L_i = \pi_i(C_i)$ is the probability that a call is blocked on link i, and is given by $L_i = E(a_i, C_i)$, where $E(a, C) = ba^C/C!$ with $b^{-1} = \sum_{v=0}^{C} (a^v/v!)$ (*Erlang's formula*). Notice that $a_k(u)$ does not depend on u, and so both of our expected rates are consistent with (5). The idea now is to update (5) replacing a_k by (6) in the hope that (5) better approximates the marginal distribution of $\boldsymbol{u}$ for the original model. If the sequence of iterates for a_k converges for each k, then the corresponding limiting values, $L_1, L_2, \ldots, L_J$, for the link blocking probabilities will satisfy

$$L_j = E\left(\sum_{r \in \mathcal{R}} \lambda_{jr} \nu_r \prod_{i \in r - \{j\}} (1 - L_i),\ C_j \right).$$

These equations *do* have a fixed point, called the *Erlang fixed point* (EFP); this follows from the *Brouwer fixed point theorem*, for they define a continuous mapping from a compact convex set $[0,1]^J$ into itself. The uniqueness of the Erlang fixed point, as well as the required convergence, was established by Kelly [11].

This approximation for the blocking probabilities, which is widely known as the *Erlang fixed point approximation*, is one of a wider class of *reduced load* approximations, named as such, because in using (5) one is effectively thinning the offered traffic at link j by an amount determined by the level of blocking at other links. There are several limiting regimes under which the EFP approximation is asymptotically *exact*. The first is one in which the topology of the network is held fixed, while capacities and arrival rates at the links become large (Kelly [11]); this has become known as the *Kelly limiting regime*, or (somewhat misleadingly) as the *heavy traffic limit*. Under the second limiting regime, called *diverse routing*, the number of links, and the number of routes which use those links, become large, while the capacities are held fixed and the arrival rates on multi-link routes become small (see for example Hunt [7], Whitt [20], and Ziedins and Kelly [23]).

In order to illustrate the accuracy of the method, we shall consider briefly a network with K links forming a loop and with each link having the same capacity C. Suppose that there are two types of traffic: one-link routes (type-1 traffic) and two-link routes comprising pairs of adjacent links (type-2 traffic). Type-t traffic is offered at rate ν_t on each type-t route. Because of the

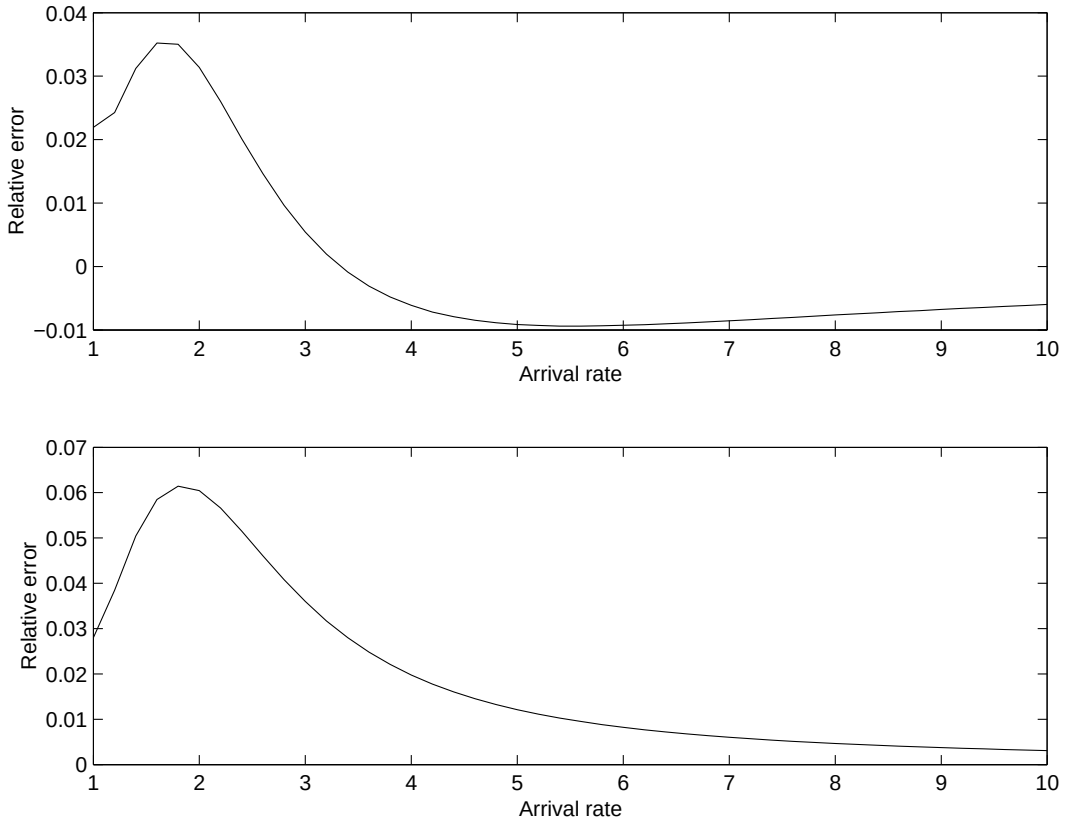

Figure 2: Accuracy of the EFP approximation for a ring network with two types of traffic.

anticipated correlation between the occupancy at adjacent links, we would not expect the EFP approximation to perform particularly well. However, as we shall see, the approximation is very accurate. If L_t is the EFP approximation for the loss probability of type-t calls, then $L_1 = B$ and $L_2 = 1 - (1-B)^2$, where the B is the unique solution to $B = E(\nu_1 + 2\nu_2(1-B), C)$. This is illustrated in Figure 2 for a network with $C = 10$, $K = 10$ and $\nu_1 = \nu_2 =$Arrival rate. The top (respectively bottom) pane shows the relative error in the EFP approximation for type-1 (respectively type-2) calls.

The EFP approximation can be improved in a number of ways. For example, if we were to

suppose that *pairs* of links or, more generally, *subnetworks* of links behave independently, then we could evaluate expected rates for these subnetworks, based on a product-form distribution π similar to (5) with $\boldsymbol{u}$ partitioned appropriately, and thus produce an iterative scheme for determining π. Several other reduced-load methods can be viewed in this way, for example the methods of Bebbington *et al.* [1, 2, 3], Ciardo and Trivedi [5], Coyle *et al.* [6], Pallant [16], Thompson [17, 18], and Thompson and Pollett [19].

5 Concluding remarks

We have presented two approaches to the basic idea of using expected rates in approximating the behaviour of complex Markovian systems. The first allows one to estimate time-dependent behaviour, and is useful in analysing queueing networks. The second approach, which is useful in estimating equilibrium behaviour, encapsulates several approximations for loss networks which are known to be asymptotically exact. This latter approach offers great promise and warrants further investigation. We are presently looking at some general formulations, as well as methods for specific models, including loss networks with admission controls and queueing networks with blocking.

Acknowledgements

We would like to thank Peter Taylor for suggesting the idea of expected rates as a "generalized reduced load concept", this being very close to our first formulation. We are grateful to the referees for valuable comments and suggestions.

References

[1] Bebbington, M.S., Pollett, P.K. and Ziedins, I. (1998) Two-link approximation schemes for linear loss networks without controls, *J. Korean Math. Soc.* **35**, 539–557.

[2] Bebbington, M.S., Pollett, P.K. and Ziedins, I. (2001) Product form approximations for highly linear loss networks with trunk reservation, *Math. Comput. Modelling* (to appear).

[3] Bebbington, M.S., Pollett, P.K. and Ziedins, I. (2002) Two-link approximation schemes for loss networks with linear structure and trunk reservation, *Telecommun. Sys.* **19**, 187–207.

[4] Boucherie, R. and Taylor, P. (1993) Transient product form distributions in queueing networks, *Discrete Event Dynamical Sys.* **3**, 375–396.

[5] Ciardo, G. and Trivedi, K.S. (1991) A decomposition approach for stochastic petri net models, *Proc. 4th Int. Workshop on Petri Nets and Performance Models*, IEEE Comp. Soc. Press, pp. 74–85.

[6] Coyle, A.J., Henderson, W. and Taylor, P.G. (1995) Decomposition methods for loss Networks with circuit reservation, In (Ed. W. Henderson) *Proc. 7th Austral. Teletraffic Res. Seminar*, Teletraffic Research Centre, University of Adelaide, Adelaide, pp. 229–241.

[7] Hunt, P.J. (1995) Loss networks under diverse routing: the symmetric star network, *Adv. Appl. Prob.* **25**, 255–272.

[8] Jackson, J.R. (1963) Jobshop-like queueing systems, *Mgmt. Sci.* **10**, 131–142.

[9] Kelly, F.P. (1979) *Reversibility and Stochastic Networks*, Wiley, Chichester.

[10] Kelly, F.P. and Pollett, P.K. (1983) Sojourn times in closed queueing networks, *Adv. Appl. Prob.* **15**, 638–656.

[11] Kelly, F.P. (1986b) Blocking probabilities in large circuit-switched networks, *Adv. Appl. Prob.* **18**, 473–505.

[12] Kelly, F.P. (1986a) Blocking and routing in circuit-switched networks, In O.J. Boxma, J.W. Cohen and H.C. Tijms (eds), *Teletraffic Analysis and Computer Performance Evaluation*, Elsevier Science (North-Holland), pp. 37–45.

[13] Kelly, F.P. (1991) Loss networks, *Ann. Appl. Prob.* **1**, 319–378.

[14] Kingman, J.F.C. (1969) Markov population processes, *J. Appl. Prob.* **6**, 1–18.

[15] Kühn, P.J. (1979) Approximate analysis of general queueing networks by decomposition. *IEEE Trans. Commun.* **27**, 113–126.

[16] Pallant, D. (1992) A reduced load approximation for cellular mobile networks including handovers, *Austral. Telecom. Res.* **26**, 21–30.

[17] Thompson, M.R. (2000) Analysis of a ring-based loss network, *Int. Trans. Operat. Res.* **7**, 419–429.

[18] Thompson, M.R. (2001) *Fixed Point Methods for Loss Networks*, PhD Thesis, The University of Queensland.

[19] Thompson, M.R. and Pollett, P.K. (2001) *A reduced load approximation accounting for link interactions in a loss network*, Submitted for publication.

[20] Whitt, W. (1985) Blocking when service is required from several facilities simultaneously, *AT&T Tech. J.* **64**, 1807–1856.

[21] Whittle, P. (1967) Nonlinear migration processes, *Bull. Inst. Int. Statist.* **42**, 642–647.

[22] Whittle, P. (1968) Equilibrium distributions for an open migration process, *J. Appl. Probab.* **5**, 567–571.

[23] Ziedins, I.B. and Kelly, F.P. (1989) Limit theorems for loss networks with diverse routing, *Adv. Appl. Prob.* **21**, 804–830.