Networks of queues

1 Time reversal of Markov chains

In studying the $M/M/1$ queue, we made use of reversibility in order to compute its invariant distribution. While this could have been done, in principle, by solving the global balance equations, reversibility made it much easier as we only needed to solve the local balance equations. This becomes even more important when trying to obtain the invariant distribution of Markov chains whose state space is multi-dimensional, as is the case for networks of queues. (For an $M/M/1$ queue, the state space is $\mathbb{Z}_+$, the set of non-negative integers, whereas for a network of d such queues, it would be $\mathbb{Z}_+^d$.) In general, there is no good way of obtaining the invariant distribution unless the Markov chain is reversible. Unfortunately, not all Markov chains of interest are reversible! However, there is a class of queueing networks called Jackson networks which, while they aren't reversible, can be analysed using techniques related to time reversal. So we'll start with a study of those ideas, leading up to the development of Kelly's lemma.

Let $(X_t, t \in \mathbb{R})$ be a continuous time Markov chain on some countable state space S, with transition rate matrix Q. Recall that any probability vector π on the state space S which solves the global balanace equations $\pi Q = \mathbf{0}$ (or, writing it out componentwise, $\sum_{j \neq k} \pi_j q_{jk} = -\pi_k q_{kk} = \pi_k \sum_{j \neq k} q_{kj}$ for all $k \in S$) is an invariant distribution of this Markov chain. If an invariant distribution exists, then it is unique if the Markov chain is irreducible (i.e., it is possible to move between any two states, though not necessarily in one step).

We say that the Markov chain X_t is stationary if X_t has the same distribution as X_{t+u} for any u and, similarly, all finite-dimensional distributions are left unchanged by a time shift: the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is the same as the joint distribution of $(X_{t_1+u}, X_{t_2+u}, \ldots, X_{t_k+u})$ for all u, and all

k and $t_1, t_2, \ldots, t_k$. This implies that for any fixed t, X_t has distribution π, and for any fixed $t_1 \leq t_2 \leq t_k$, the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is given by

$$\begin{aligned}&P(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_k} = x_k)\\&= \pi_{x_1} e^{Q(t_2-t_1)}(x_1, x_2) \cdots e^{Q(t_k-t_{k-1})}(x_{k-1}, x_k),\end{aligned}$$

where $e^{Qt}(i,j)$ denotes the ij^{th} element of the matrix e^{Qt}.

Warning: Stationarity only implies that the marginal distribution at any *fixed time* is the invariant distribution π. It doesn't imply that joint distributions are products of the invariant distribution; $P(X_{t_1} = x_1, X_{t_2} = x_2) \neq \pi(x_1)\pi(x_2)$. It also doesn't say anything about *random times*. For example, the distribution at the first time after 0 that the Markov chain visits x (a random time) is obviously unit mass concentrated at x rather than the invariant distribution π.

Suppose that the Markov chain $(X_t, t \in \mathbb{R})$ is irreducible with unique invariant distribution π, and that it is stationarity. Let $Y_t = X_{s-t}$ denote its time reversal around some fixed time s. Then $(Y_t, t \in \mathbb{R})$ is also a stationary continuous-time Markov chain, with the same invariant distribution π but with transition rate matrix $\tilde{Q}$ given by

$$\tilde{q}_{jk} = \frac{\pi_k q_{kj}}{\pi_j}, \quad j, k \in S.$$

The intuition is that, in the long term, the rate at which one observes jumps from j to k in forward time is $\pi_j q_{jk}$. This is because the Markov chain spends a fraction π_j of its time in state j and, when in state j, jumps to state k occur at rate q_{jk}. So this should also be the rate at which we observe jumps from k to j for the reversed process. But for the reversed process, by the same reasoning, this rate is $\tilde{\pi}_k \tilde{q}_{kj}$, where $\tilde{\pi}$ is the invariant distribution of the reversed process. Hence,

$$\pi_j q_{jk} = \tilde{\pi}_k \tilde{q}_{kj}.$$

But the invariant distribution is the same as the long-term fraction of time spent in each state (by the ergodic theorem for Markov chains), and hence is the same whether we are watching the process in forward or reverse time. So $\tilde{\pi}_k = \pi_k$. Substituting this above, we get $\tilde{q}_{kj} = \pi_j q_{jk}/\pi_k$, as claimed. We have only made this argument intuitively so far (we took for granted that the time reversal was Markovian, for instance). See the proof of Kelly's lemma below for full details.

We found reversibility very useful in computing invariant distributions because it meant that we could solve local or detailed balance equations rather than global balance equations, and usually the local balance equations are simpler. There is a similar way in which time reversal can help; if we can guess the invariant distribution, it gives an easy way of verifying if the guess is correct.

Lemma 1 (Kelly's lemma) *Let $\{X_t, t \in \mathbb{R}\}$ be a stationary CTMP with transition rates q_{ij}. If we can find a set of non-negative numbers $\tilde{q}_{ij}$, $i, j \in S$, and non-negative numbers π_j, $j \in S$ summing to one, such that*

$$\begin{aligned} \sum_{k \neq j} \tilde{q}_{jk} &= \sum_{k \neq j} q_{jk} \quad \forall\, j \in S, & (1) \\ \pi_j q_{jk} &= \pi_k \tilde{q}_{kj} \quad \forall\, j, k \in S, & (2) \end{aligned}$$

then the $\tilde{q}_{jk}$ are the transition rates for the reversed process and π is an invariant distribution for both forward and reversed processes.

Proof. If π and $\tilde{Q}$ satisfy the assumptions of the lemma, then for each $k \in S$, we have

$$\begin{aligned} \sum_{j \neq k} \pi_j q_{jk} &= \sum_{j \neq k} \pi_k \tilde{q}_{kj} \quad \text{by (2)} \\ &= \pi_k \sum_{j \neq k} q_{kj} \quad \text{by (1)} \\ &= -\pi_k q_{kk}, \end{aligned}$$

where the last equality holds because the row sums of the Q matrix are zero by definition. So π satisfies the global balance equations $\pi Q = 0$. Since π_j sum to 1 by the assumptions of the lemma, π is an invariant distribution.

It remains to show that the reversed process is Markovian with transition rates $\tilde{q}_{jk}$. Fix $s \in \mathbb{R}$ and define $Y_t = X_{s-t}$ for all $t \in \mathbb{R}$. Now, for $\{Y_t, t \in \mathbb{R}\}$ to be a Markov process, we require that for all $\tau \in \mathbb{R}$, $\{Y_t, t < \tau\}$ is conditionally independent of $\{Y_t, t > \tau\}$ given Y_τ. But this is the same as requiring that $\{X_t, t > s-\tau\}$ be conditionally independent of $\{X_t, t < s-\tau\}$ given $X_{s-\tau}$, which holds by the Markov property of the $\{X_t\}$ process. Hence, $\{Y_t\}$ is a Markov process.

We now compute its transition rates. We have

$$\begin{aligned} P(Y_{t+h}=k|Y_t=j) &= P(X_{s-t-h}=k|X_{s-t}=j) \\ &= \frac{P(X_{s-t-h}=k, X_{s-t}=j)}{P(X_{s-t}=j)} \\ &= \frac{P(X_{s-t-h}=k)P(X_{s-t}=j|P(X_{s-t-h}=k)}{P(X_{s-t}=j)} \\ &= \frac{\pi_k(q_{kj}h+o(h))}{\pi_j}. \end{aligned}$$

Taking limits as h tends to zero, we see that the transition rates of the reversed process are $q'_{kj}=\pi_k q_{kj}/\pi_j$, which is the same as $\tilde{q}_{kj}$ given in the statement of the lemma. Finally, we need to check that π also satisfies the global balance equations for the transition rates $\tilde{Q}$. This is exactly analogous to the corresponding proof of $\pi Q = 0$ and is omitted. This completes the proof of the lemma. □

2 Jackson networks

We shall now turn our attention to studying a special class of networks of queues, called Jackson networks. These are very popular in practical applications of queueing theory, in part because it is possible to explicitly calculate their invariant distribution, and it has a particularly simple form.

A network consisting of a finite number, J, of queues is called a Jackson network if:

1. The arrival processes into different queues are independent Poisson processes. Let η_j denote the external arrival rate into queue j.

2. Each queue has a single server operating a FCFS policy, and the service times of customers at each queue are iid exponential random variables, with distribution $Exp(\mu_j)$ at the j^{th}. Moreover, the service times at different queues are mutually independent.

3. Routing is Markovian: a customer finishing service at queue j is routed to queue k with some fixed probability r_{jk}, and leaves the system with probability $r_{j0} = 1 - \sum_{k=1}^{J} r_{jk}$. Routing decisions are mutually independent, and independent of the past history of customers (for

example, their arrival and service times, and their route through the network up to that time).

4. The network is open, i.e., with probability one, every customer can only visit finitely many queues before leaving the network.

The routing matrix R has non-negative entries, and all its row sums must be less than or equal to 1, as $\sum_{j=1}^{J} r_{ij}$ is the probability that a customer leaving queue i is routed to some other queue, rather than leaving the network. Such a matrix is called substochastic. (Recall that it is called stochastic if the entries are non-negative and all the row sums are equal to 1.) Note that it is possible for r_{jj} to be positive, i.e., a customer leaving a queue may return to it immediately. More generally, a customer may also return after visiting some other queues. But we cannot have $r_{jj} = 1$, nor can it be possible for a customer to revisit any queue with probability 1 because that would require the customer to circulate around the network forever, which contradicts the requirement for the network to be open. In particular, at least some of the row sums of the R matrix have to be strictly smaller than 1. Thus, the picture of a Jackson network is that a customer enters some queue, visits a deterministic or random subset of queues (possibly more than once), and eventually leaves the network.

2.1 Traffic equations

Consider a single queue in a Jackson network in isolation. It gets external arrivals according to a Poisson process. In addition, customers also enter the queue after having been served at some other queue and being rerouted. Let us denote by λ_j the total arrival rate into queue j, of both new and rerouted customers. How can we compute λ_j?

Suppose each queue in the network is stable, i.e., the queue length doesn't grow unboundedly over time. Then, it must be the case that, in the long run, every customer entering the queue must eventually leave the queue (otherwise the queue would build up). Hence, the long-run departure rate from queue j must be the same as the long-run arrival rate, which we denoted λ_j. Now, the arrivals into queue j are made up of the external arrivals, as well as departures from other queues that get routed to it. Hence, we must have

$$\lambda_j = \eta_j + \sum_{i=1}^{J} \lambda_i r_{ij}, \quad j = 1, \ldots, J, \tag{3}$$

since a fraction r_{ij} of all departures from queue i are routed to queue j. These equations, one for each $j = 1, \ldots, J$, are called the traffic equations. We can write them in matrix form as

$$\lambda = \eta + \lambda R, \tag{4}$$

where λ and η are row vectors, and R is the $J \times J$ routing matrix withe entries r_{ij}. The traffic equations have a unique solution, given by

$$\lambda = \eta(I - R)^{-1} = \eta\big(I + R + R^2 + R^3 + \ldots\big). \tag{5}$$

The reason that the matrix $I - R$ is invertible, and that the series $I + R + R^2 + \ldots$ converges, has to do with the assumption of the Jackson network being open. The intuition is as follows. The ij^{th} element of the matrix R^n is the probability that a customer leaving queue i ends up in queue j after n routing steps (r_{ij} is the probability of going from queue i to queue j in one step); hence $\sum_{n=1}^{\infty}(R^n)_{ij}$ denotes the expected total number of visits to queue j made by a customer leaving queue i before it leaves the system. The assumption that the queueing network is open is equivalent to the assumption that this expected total number of visits is finite for every pair of queues; a customer can't circulate around for ever.

Examples

1. Consider a network of 3 queues with external arrival rate $\eta_1 = 2$ into the first queue, no external arrivals into the other queues, and with routing matrix
$$R = \begin{pmatrix} 0.3 & 0.5 & 0 \\ 0 & 0.2 & 0.8 \\ 0.3 & 0 & 0.7 \end{pmatrix}$$
The traffic equations can be written as
$$\begin{aligned} \lambda_1 &= \eta_1 + 0.3\lambda_1 + 0.3\lambda_3, \\ \lambda_2 &= 0.5\lambda_1 + 0.2\lambda_2, \\ \lambda_3 &= 0.8\lambda_2 + 0.7\lambda_3. \end{aligned}$$
Solving these equations, we obtain $\lambda_1 = 10$, $\lambda_2 = \frac{25}{4}$ and $\lambda_3 = \frac{50}{3}$.

2. Consider the same network above but with routing matrix
$$R = \begin{pmatrix} 0.3 & 0.5 & 0 \\ 0 & 0.2 & 0.8 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

This network is not open since customers who enter queues 2 or 3 can't leave this subset of queues. Hence, the total number of customers in these two queue will tend to infinity over time, though some customers who enter the system will leave, since $r_{10} = 1 - 0.8 = 0.2 > 0$.

3. Consider the same network above but with routing matrix

$$R = \begin{pmatrix} 0.3 & 0 & 0 \\ 0 & 0.2 & 0.8 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

The network is again not open. In this case though, all customers who enter queue 1 eventually leave the network (after a geometric number of returns to queue 1), and there are no arrivals into queues 2 or 3. If there are any customers in this subset of queues initially, they will continue to circulate forever, but these queues won't grow over time.

2.2 Invariant distribution of Jackson networks

Consider a Jackson network with J nodes, and let $X(t) = (X_1(t), \ldots, X_J(t))$ denote the vector of queue lengths at time t. Then, $(X(t), t \in \mathbb{R})$ is a continuous time Markov process (CTMP). To see this, note that the conditional distribution of $X(t+s)$ given $X(u)$ for all $u \leq t$ depend only on $X(t)$ and on the arrivals, services and routing decisions between times t and $t+s$. By the assumptions defining a Jackson network, these are independent of $X(u)$, $u < t$ (and also of $X(t)$).

Let us now compute the transition rates of this CTMP. Suppose the Markov process is in state $\mathbf{n} \in \mathbb{Z}_+^J$. What are the possible transitions from this state? The process could go to state $n + e_i$ if there is an external arrival into queue i. (Here, e_i denotes the i^{th} unit vector, namely the vector $\begin{pmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}$ where there is a 1 in the i^{th} position and zeros elsewhere.) If $n_i \geq 1$, it could also go to state $n - e_i$ if there is a service at queue i, and the customer departs the system upon service, or to state $n - e_i + e_j$ if there is a service at queue i and the customer is thereupon routed to queue j. We can write down the rates for all these transitions in terms of the external arrival rates, the service rates and the routing probabilities.

We have

$$\begin{aligned} q(\mathbf{n}, \mathbf{n}+e_i) &= \eta_i, & (6)\\ q(\mathbf{n}, \mathbf{n}-e_i) &= \mu_i r_{i0} 1(n_i \geq 1), & (7)\\ q(\mathbf{n}, \mathbf{n}-e_i+e_j) &= \mu_i r_{ij} 1(n_i \geq 1). & (8)\end{aligned}$$

We want to compute the invariant distribution of this CTMP. The direct approach to doing so would be to solve the global balance equations, but it isn't obvious how to solve them (and since they are a system of infinitely many simultaneous equations, they can't be solved numerically). So we shall approach it indirectly, using Kelly's lemma. In order to use Kelly's lemma, we need to guess at both the invariant distribution and the transition rates of the time reversed Markov chain, and then verify that the guess is correct.

We start with the time reversal. Consider the Jackson network described above. What is the rate at which customers enter queue i from outside in the reversed process? In the long run, this has to be the same as the rate at which customers from queue i depart the system in the forward process. Since customers enter queue i at total rate λ_i (here λ is the solution of the traffic equations and λ_i includes both external arrivals and rerouted customers), then this must also be the total rate at which customers leave queue i, assuming that the queue is stable, i.e., that $\lambda_i < \mu_i$. Of all customers leaving queue i, a fraction r_{i0} depart the system. Hence, the long run rate at which customers from queue i depart the system in the forward process is $\lambda_i r_{i0}$. This leads us to guess that this is the rate of external arrivals into queue i for the reversed process which we denote $\tilde{\eta}_i$, i.e., that

$$\tilde{\eta}_i = \lambda_i r_{i0}, \quad i = 1, \ldots, J. \tag{9}$$

Note that it is far from obvious that the process of departures from some queue to outside the network (and hence the arrival process in reverse time) is Poisson. At this point, we are simply guessing that it is so and writing down our guess for the corresponding transition rates.

Likewise, if we assume that the time reversal is also a Jackson network, we can guess its various other parameters. The service rate at queue i, which is the reciprocal of the mean time between when a customer starts being served and finishes being served, is the same in forward and reverse time (start and finish reverse roles, which leaves the time between start and finish unchanged). Likewise, the total rate at which customers enter, or depart, a queue is the same in forward and reverse time provided all queues are stable, so that all arrivals depart. Using tildes to denote time-reversed parameters, we can thus write

$$\tilde{\lambda}_i = \lambda_i, \ \tilde{\mu}_i = \mu_i, \quad i = 1, \ldots, J. \tag{10}$$

Next, the rate at which customers move from queue i to queue j in forward time is $\lambda_i r_{ij}$, since they leave queue i at rate λ_i and a fraction r_{ij} of them move to queue j. This has to be the same as the rate at which customers move from queue j to queue i in reversed time which, by the same reasoning, is $\tilde{\lambda}_j \tilde{r}_{ji}$. From this, we obtain the routing parameters of the time-reversed Jackson network as

$$\tilde{r}_{ji} = \frac{\lambda_i r_{ij}}{\tilde{\lambda}_j} = \frac{\lambda_i r_{ij}}{\lambda_j}. \tag{11}$$

We can check that the matrix $\tilde{R}$ is substochastic. Clearly, $\tilde{r}_{ji}$ given by (11) is non-negative. Moreover,

$$\sum_{i=1}^{J} \tilde{r}_{ji} = \frac{1}{\lambda_j} \sum_{i=1}^{J} \lambda_i r_{ij} = \frac{\sum_{i=1}^{J} \lambda_i r_{ij}}{\eta_j + \sum_{i=1}^{J} \lambda_i r_{ij}} \leq 1,$$

where we have used the traffic equation (3) for λ_j in order to get the second equality above. You might also want to check for yourselves that the vector λ satisfies the traffic equations with the time-reversed parameters $\tilde{\eta}$ and $\mathbf{R}$.

Having guessed at the time-reversed parameters, we guess that the time reversed process behaves as a CTMP with transition rates given by equations analogous to (6,7,8), but with η_i and r_{ij} replaced by $\tilde{\eta}_i$ and $\tilde{r}_{ij}$ respectively.

Next, we need to guess the invariant distribution for this Markov process. The total arrival rate into queue i is λ_i, the service rate is μ_i and service times are exponential, and there is a single server. Therefore, if the arrival process into queue i were Poisson, the invariant distribution of this queue would be geometric, $\pi^i(n) = (1 - \rho_i)\rho_i^n$, $n = 0, 1, 2, \ldots$, where $\rho_i = \lambda_i / \mu_i$. Unfortunately, the total arrival process is not Poisson in general (though the external arrival process is). Intuitively, the reason is that, if it is possible for a departure from queue i to eventually revisit queue i with positive probability, then if there have been a lot of arrivals to queue i in the past, this will increase the conditional rate of future arrivals as some of these will return to queue i later. Thus, the future of the arrival process is not independent of its past, which means that it cannot be Poisson.

Nevertheless, let's be wildly optimistic and assume that the invariant queue length distribution at each queue is geometric, as above. In fact, let's go even further and assume that the joint distribution is product form, i.e., that

$$\pi(n_1, n_2, \ldots, n_J) = \prod_{i=1}^{J} \pi^i(n_i) = \prod_{i=1}^{J} (1-\rho_i)\rho_i^{n_i}. \tag{12}$$

This would be the case if the queues were independent, but in fact they are not independent. So there really is no good justification for our guess but let's try it out anyway.

Now that we have a guess for the invariant distribution π and the reversed transition rate matrix $\tilde{Q}$, we are in a position to check the assumptions of Kelly's lemma. To check the first assumption, we need to show that

$$\sum_{\mathbf{m}\neq\mathbf{n}} \tilde{q}(\mathbf{n},\mathbf{m}) = \sum_{\mathbf{m}\neq\mathbf{n}} q(\mathbf{n},\mathbf{m}). \tag{13}$$

Now, all the possible transitions out of state $\mathbf{n}$ and their rates are specified by equations (6,7,8). Hence,

$$\begin{aligned} \sum_{\mathbf{m}\neq\mathbf{n}} \tilde{q}(\mathbf{n},\mathbf{m}) &= \sum_{i=1}^{J} \Big(\tilde{\eta}_i + \tilde{\mu}_i 1(n_i > 0) \Big(\tilde{r}_{i0} + \sum_{j=1}^{J} \tilde{r}_{ij} \Big) \Big) \\ &= \sum_{i=1}^{J} (\lambda_i r_{i0} + \mu_i 1(n_i > 0)), \end{aligned}$$

where we have used (9), (10), and the fact that $\tilde{r}_{i0} = 1 - \sum_{j=1}^{J} \tilde{r}_{ij}$ by definition, to obtain the last equality. Likewise,

$$\sum_{\mathbf{m}\neq\mathbf{n}} q(\mathbf{n},\mathbf{m}) = \sum_{i=1}^{J} (\eta_i + \mu_i 1(n_i > 0)).$$

Hence, in order to verify (13), we need to verify that

$$\sum_{i=1}^{J} \lambda_i r_{i0} = \sum_{i=1}^{J} \eta_i. \tag{14}$$

But the LHS of the above expression is the total rate at which customers leave the system, while the RHS is the total rate at which customers enter the system. Since the network is open, these two quantities have to be the same, provided all queues are stable, i.e., $\lambda_i < \mu_i$ for all $i = 1, \ldots, J$. If this condition is not satisfied, then the system does not have an invariant distribution since at least some of the queues grow to infinity. (If this intuitive argument is not to your satisfaction, then you can prove (14) more formally. Let $\mathbf{1}$ denote the column vector of all ones, of length J. Then, $\sum_{i=1}^{J} \eta_i = \eta\mathbf{1}$, while

$$\sum_{i=1}^{J} \lambda_i r_{i0} = \sum_{i=1}^{J} \lambda_i (1 - \sum_{j=1}^{J} r_{ij}) = \lambda(I - R)\mathbf{1},$$

which is equal to $\eta\mathbf{1}$ by (4).)

Next, we need to check the second assumption of Kelly's lemma, i.e., for all possible states $\mathbf{m}$ and $\mathbf{n}$, we need to show that $\pi(\mathbf{n})\tilde{q}(\mathbf{n}, \mathbf{m}) = \pi(\mathbf{m})q(\mathbf{m}, \mathbf{n})$. If both transition probabilities $\tilde{q}(\mathbf{n}, \mathbf{m})$ and $q(\mathbf{m}, \mathbf{n})$ are zero, then this equation holds with 0=0. So we only need to check this condition when at least one of the corresponding transition probabilities is non-zero. Fix i between 1 and J, and consider $\mathbf{m} = \mathbf{n} + e_i$. Then,

$$\tilde{q}(\mathbf{n}, \mathbf{m}) = \tilde{\eta}_i = \lambda_i r_{i0},$$

by (9), whereas

$$q(\mathbf{m}, \mathbf{n}) = \mu_i r_{i0} 1(m_i > 0) = \mu_i r_{i0} 1(n_i + 1 > 0) = \mu_i r_{i0}.$$

Hence, the condition, $\pi(\mathbf{n})\tilde{q}(\mathbf{n}, \mathbf{m}) = \pi(\mathbf{m})q(\mathbf{m}, \mathbf{n})$, that we need to check can be written as

$$\prod_{j=1}^{J} (1 - \rho_j)\rho_j^{n_j} \lambda_i r_{i0} = \prod_{j \neq i} (1 - \rho_j)\rho_j^{n_j} (1 - \rho_i)\rho_i^{n_i+1} \mu_i r_{i0}.$$

If we cancel out all common terms, the above reads $\lambda_i = \rho_i \mu_i$, which holds by the definition of ρ_i. Thus, we have verified the second assumption of Kelly's lemma for $\mathbf{m} = \mathbf{n} + e_i$, for any i between 1 and J. We similarly need to verify it for $\mathbf{m} = \mathbf{n} - e_i$ and $\mathbf{m} = \mathbf{n} - e_i + e_j$. As the steps are very similar, the details are omitted, though you might want to check it for yourself.

Since the assumptions of Kelly's lemma are verified, its conclusions hold, i.e., our guesses for the invariant distribution π and the reversed parameters $\tilde{Q}$ are correct. Let us summarise this important result as a theorem.

Theorem 1 *Consider a network consisting of J single server nodes, with external arrivals into node i according to a Poisson process of rate η_i, iid* *exponential service times at node i with mean $1/\mu_i$, and routing matrix R specifying the probabilities r_{ij} that a customer leaving node i is routed to node j. Suppose that the routing decisions at each node are iid, and that the routing matrix is such that the network is open. Suppose too that the arrival and service processes and routing decisions at different nodes are mutually independent. Finally, assume that the network is stable, i.e., that the solutions λ_j, $j = 1, \dots, J$, of the traffic equations satisfy $\lambda_j < \mu_j$ for all $j = 1, \dots, J$. Then, this Jackson network has unique invariant distribution π given by*

$$\pi(\mathbf{n}) = \prod_{j=1}^{J} (1 - \rho_j)\rho_j^{n_j},$$

where $\rho_j = \lambda_j/\mu_j$.

3 Generalisations

A natural question to ask at this point is whether there are other queueing network models for which we can compute invariant distributions. There are a small number of other such examples, some of which we now describe. We start with single queues and then go on to networks.

M/G/∞ queue Consider a queue with Poisson arrivals at rate λ, iid service times with an arbitrary distribution (G stands for general), and infinitely many servers (so that all customers are served in parallel). Denote the mean service time by $E[S]$ and define $\rho = \lambda E[S]$. If the service time had the *Exp*(μ) distribution, then the mean service time would be $1/\mu$, so the definition of ρ is consistent with its definition in the $M/M/\infty$ queue. The invariant queue length π for the $M/G/\infty$ queueing model is the same as for the corresponding $M/M/\infty$ model, namely the Poisson(ρ) distribution:

$$\pi(n) = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 0, 1, 2, \dots$$

We say that the invariant distribution in the $M/G/\infty$ queue has the *insensitivity* property; it only depends on the mean service time and is not sensitive to the actual distribution of the service time.

M/G/1-PS queue This is a queue with Poisson arrivals, general service time distributions and a single server which uses a processor-sharing service discipline. This means that it allocates an equal portion of its service capacity to each customer in the queue. Say for example that the queue is initially empty, a customer enters at time $T_1 = 0$ with a service time requirement of $S_1 = 10$ units, the next customer arrives at time $T_2 = 3$ with a service requirement of $S_2 = 5$ units, and the third customer arrives at time $T_3 = 20$. The first customer gets the full attention of the server from time 0 till time 3, so will have a residual service requirement of 10-3=7 units at the arrival time of the second customer. From time T_2 onwards, both customers in system each get only half the attention of the server. Their residual service requirements at time $T_2 = 3$ are 7 and 5 units respectively. It will thus take 10 time units for the service of the second customer to finish, and this customer will depart at time 3+10=13. During these 10 time units, the first customer will also receive 5 units of service, so its residual service requirement at time 13 will be 7-5=2 units. After the departure of the second customer, the first customer will get the undivided attention of the server, hence it will take 2 time units to finish being served. Thus, the first customer will depart at time 15, before the third customer has entered.

The above example describes how a processor-sharing queue works. Customers arrive at times T_1, T_2, $T_3, \ldots$ bringing with them random service requirements S_1, S_2, $S_3, \ldots$, which are iid with a general distribution. Let λ denote the rate of the Poisson arrival process. Then $T_2 - T_1$, $T_3 - T_2, \ldots$ are iid $Exp(\lambda)$ random variables. There is infinite waiting room at the queue, and a single server which works at unit rate. The server divides its effort equally among all customers who are in the queue. Let $E[S]$ the mean service time, and define $\rho = \lambda E[S]$. The invariant distribution π of this queue is geometric, given by

$$\pi(n) = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots$$

In other words, the invariant distribution is the same as for an $M/M/1$ queue with the same mean service time requirement. The invariant distribution is also *insensitive* in that it only depends on the mean service time and not on the service time distribution.

M/G/1-LIFO or M/G/1-LCFS queue This is a queue with Poisson arrivals, whose rate we denote by λ, and iid service times with general distribution, whose mean we denote by $E[S]$. We define $\rho = \lambda E[S]$. There is a single server and infinite waiting room. The service discipline is Last-In-First-Out, also called Last-Come-First-Served. To be more precise, the service discipline is LIFO/LCFS with *pre-emptive resume.* Pre-emptive means that, when a new job/customer comes in, the server drops the job it is working on, and moves to the new job. Resume means that any work already done on that job is not wasted; when the server is finished with the new jobs, it can resume work on the old job from where it left off. The results stated below only apply to this model. Non-preemptive, and pre-emptive restart (where work done on dropped jobs is wasted and has to be restarted from scratch) disciplines have quite different behaviour. The invariant distribution, π, for the $M/G/1$-LIFO is geometric,

$$\pi(n) = (1-\rho)\rho^n, \quad n = 0, 1, 2, \dots,$$

which is the same as for an $M/M/1$ queue with the same mean service time requirement. The invariant distribution is again *insensitive* in that it only depends on the mean service time and not on the service time distribution.

The FIFO service policy that we started studying in the context of $M/M/1$ queues seems the most natural, and is also the most familiar from daily life. So it might appear that PS and LIFO policies are mathematical abstractions of little practical use. In fact, they are quite widely used in practice, especially in computer systems. A computer may be working on several different tasks or "threads" in parallel. It will typically do so by using a 'round-robin' service discipline, doing a little bit of work on each task before moving on to the next. If the amount of work done on each task before moving to the next is small compared to the size of the task, then the processor-sharing model is a good approximation to the round-robin model, and much more tractable mathematically. The TCP protocol, which regulates data transmission over the Internet, shares the link bandwidth approximately equally among all transmissions sharing a link. Processor sharing models are widely used to describe its behaviour. The LIFO policy is also used to schedule tasks in computer systems, where tasks are placed in a 'stack' from bottom to top by one process and retrieved from top to bottom by another.

What is the reason for all these different disciplines? One reason to prefer processor-sharing over FIFO in some situations is that customers derive some benefit from partially completed tasks, and so it may be better to partially complete multiple jobs in parallel rather than perform them sequentially, starting one only after the previous one has finished. Another reason is that, if jobs require highly variable service times, a single large job can hold up many smaller jobs in FIFO but this doesn't happen with PS. It also happens with LIFO but has a smaller impact. Whereas in FIFO, the longer a job the more jobs it holds up (as it holds up all jobs which arrived while it was being worked on), in LIFO it only holds up the jobs whose service it interrupted. Thus, in LIFO, both short and long jobs hold up only the same number of jobs on average.

Having considered these different queueing models in isolation, we now consider networks made up of such queues. It turns out that, if any combination of these queues, namely $M/M/1$, $M/G/\infty$ (including $M/M/\infty$ as a special case), $M/G/1$-PS, or $M/G/1$-LIFO, are connected together in a network with Markovian routing, and if the routing matrix satisfies the requirements for the network to be open, then the invariant distribution is product form,

$$\pi(\mathbf{n}) = \prod_{i=1}^{J} \pi^i(n_i),$$

where each π^i is either geometric or Poisson, depending on whether it is a single-server or infinite-server queue respectively.

Unfortunately, there are only a few models of queueing networks for which it is possible to explicitly calculate invariant distributions, of which the above are the most common examples. In practical applications, one may typically try to use one of these models even if it doesn't exactly fit the situation. Another alternative is to start with a more accurate model that can't be solved explicitly, but use simulation to study its behaviour. Simulation is perhaps the only 'general-purpose' modelling technique, and is useful for generating quantitative predictions, but is not very good at yielding qualitative insights. This course aimed to introduce you to some of the most commonly used models and analytical techniques in queueing theory, and to equip you to apply them to a range of real-life problems.