THESIS FOR THE DEGREE OF LICENTIATE OF TECHNOLOGY

On Perfect Simulation of Certain Queueing Networks

Dan Mattsson

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY and GÖTEBORG UNIVERSITY
Göteborg, Sweden 2000

On Perfect Simulation of Certain Queueing Networks
Dan Mattsson

ISSN 0347–2809/NO 2000:19
Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31–772 1000

Chalmers University of Technology
Göteborg, Sweden 2000

Abstract

In this paper two models of queueing networks with finite waiting spaces are presented. They are similar to Jackson networks in that they have Poisson arrivals and exponential service time and obey Markovian routing. The two models are compared and it is shown how the coupling from the past algorithm can be applied to the network models to obtain samples from their stationary distributions. The paper contains several simulation examples.

Keywords: Markov Chain Monte Carlo, Perfect simulation, Propp-Wilson algorithm, Queueing theory, Jackson networks, Birth and Death processes.

MSC 2000 subject classifications: 60K20, 60K25, 68M20, 90B22, 65C05, 60E15, 60J80

Contents

Introduction 1

1 Comments on perfect simulation

2 Comments on queueing theory

1 Basic theory 5

1.1 Markov theory

1.2 Coupling theory

2 Queueing systems, Jackson networks 9

2.1 Open Jackson networks

2.2 Stationary distribution

2.3 Domination results

2.4 Comments on the models

3 Poisson calculus 21

3.1 Poisson representation

3.2 The two-dimensional Poisson embedding

3.3 Coupling of two birth and death processes

4 Coupling from the past 31

4.1 Fundamentals

4.2 Size of time-steps

4.3 Other algorithms

5 Simulations 39

5.1 Performance

5.2 Example of networks

5.3 Comments on the simulations

Conclusions 55

References 57

Preface

Acknowledgement

No man is an island. . .

and thus I am grateful for the good nature and grace of many people. Those that have made the work joyful, without whose help this report would never have been finished, and those without whose help it would have been achieved substantially faster. I am obliged to you all.

One man especially deserves mentioning: my advisor Torgny Lindvall, who was always ready to provide support, encouragement and optimism.

Dan Mattsson
Göteborg, April 2000

Introduction

This paper is devoted to the study of simulation of queueing networks, and showing that samples from the stationary distribution of the number of customers in the network can be obtained in a reasonable amount of time. The queueing systems considered are of Jackson type, i.e., they have Poisson arrivals and exponential service time and obey Markovian routing. What distinguishes them from usual Jackson networks is that each queue has a finite waiting space.

Simulation of such systems will hopefully yield some useful information about bottlenecks, i.e., nodes in the network where congestion is likely to occur, and predict the behaviour of the queueing system when some fundamental parameters change, for instance, increased waiting spaces or faster service rates.

This goal is achieved in several steps, as follows. Chapter 1 contains some basic theory presented for completeness but can be skipped by most readers. In Chapter 2 we present the class of queueing networks to be studied, and we show how to model them as general birth and death processes. The state space for such a process has a natural partial ordering, and we will later use this ordering in constructing an algorithm for obtaining samples from the stationary distribution of customers in the network. The algorithm is the coupling from the past algorithm treated in detail in Chapter 4. Chapter 3 contains the explicit construction of the general birth and death process used to model the queueing network. The essence of Chapter 4 is to describe and to adapt the coupling from the past algorithm to our queueing networks. The algorithm enables us to obtain samples from the stationary distribution. In Chapter 5 the simulations are presented, and the chapter begins with an analysis of the performance of the simulation algorithm for large networks possessing special structure. After that follows an analysis of certain interesting networks. The remaining chapter contains a discussion on limitations of the queueing network model and possible feasible extensions of the model for future research.

1 COMMENTS ON PERFECT SIMULATION

Perfect simulation is a term for algorithms that generate samples from an unknown stationary distribution of a Markov chain. In recent years the most celebrated algorithm has been the coupling from the past algorithm (CFTP), but other algorithms exist. Earlier perfect simulation was referred to as *exact simulation*.

In general we speak of MCMC* methods when we take samples from a Markov Chain $(X_t)_{t\geq 0}$ at some time points $t_1, t_2, \ldots$ If the Markov chain is ergodic and admits the stationary distribution π, we can obtain samples from a distribution close to π by observing the chain $(X_t)_{t\geq 0}$ at some distant time $t = \tau$. This strategy, although intuitively clear and often easily implemented, suffers from some major drawbacks. First, the generated samples are usually biased and *not* from the stationary distribution π. Secondly, in order to control bias, an appropriate time τ has to be determined, which may be analytically and empirically difficult.

The coupling from the past algorithm, suggested by Propp and Wilson[10], resolves this problem by first telling us when to stop, and secondly, the obtained sample is an unbiased sample from π. For an overview of different algorithms and a survey of applications where perfect simulations have fruitfully been used, see Wilson[14] and Dimakos[5].

2 COMMENTS ON QUEUEING THEORY

Classical queueing theory dates back to Erlang (1878–1929) who investigated the loss probability in a telephone switching system with a finite number of available channels. Requests, such as telephone calls or arriving customers, are accepted only if sufficient resources are available to deal with them. The resources are modelled as queues with finite waiting spaces and each queue is served by a number of servers emptying the queue and thus freeing resources. Requests are visualised as customers arriving in a queue and they are accepted if the waiting space has vacant positions. For a brief overview of queueing networks, consult Søren Asmussen's contribution to [2]. A detailed treatment of the topic can be found in Wolff[16].

The objective is to compute quantities like the probability that an arriving customer is rejected, estimated queue lengths, service times, average waiting

*Markov Chain Monte Carlo

times, etc. For a service provider it is often interesting to know what the mean queue lengths are as a time average, so that system capacity is not wasted, while customers using the resources have other measures of quality of service.

Current research is focused on queueing networks with finite buffers with different types of workload and blocking schemes. The particular kind of network studied depends on the application. For a classification of the available literature and models of queueing networks in manufacturing, see Rao et al.[11]. Blocking occur when a station has not enough free resources to handle a request, an arriving customer, and several blocking mechanisms have been suggested to model different realistic behaviours. The blocking schemes ranges from expunging blocked customers from the system, to letting them wait at different stations or placing them in a waiting orbit.

Closed form expressions for the stationary distribution of the customers in such systems are only available in a limited number of models, and huge effort has been put into approximative algorithms for calculating it and estimate the errors in the calculations.

Chapter 1

Basic theory

The aim of this chapter is to establish the notations and conventions used throughout the paper, as well as to formally define and present basic theorems and definitions.

The statements of Section 1.1 are given for completeness only. They can be found in any textbook on Markov theory and may be skipped by most readers. For proofs of theorems and such, we refer to Norris[9]. The last section contains important results from coupling theory which are crucial in the analysis in subsequent chapters. For a careful treatment of coupling theory and stochastic domination we refer to Lindvall[7].

We shall always let n denote a discrete index, $n \in \mathbb{Z}_+$ or $n \in \mathbb{Z}$, while t is a continuous one, $t \in \mathbb{R}_+$ or $t \in \mathbb{R}$. Consequently, $(X_n)_{n\geq 0}$ will denote a Markov chain in discrete time and $(X_t)_{t\geq 0}$ one in continuous time. Further, d will always denote a dimension, $d \in \mathbb{N}$, and $\boldsymbol{b}$ a d-dimensional vector of bounds,

$$\boldsymbol{b} = (b_1, \dots, b_d) \in \mathbb{N}^d.$$

We write $\mathbb{Z}_+^{\boldsymbol{b}}$ for the d-dimensional (finite) space

$$\{0, \dots, b_1\} \times \cdots \times \{0, \dots, b_d\}.$$

The k^{th} unit vector in $\mathbb{Z}_+^d$, or $\mathbb{Z}_+^{\boldsymbol{b}}$ and $\mathbb{R}^d$ for that matter, is always denoted by e_k, $k = 1, \dots, d$.

1.1 MARKOV THEORY

Definition 1.1 For a state space E, a *intensity matrix*, $Q = [q_{ij}]_{i,j\in E}$, is a matrix with elements satisfying

- $0 \leq -q_{ii} < \infty$ for all $i \in E$.
- $q_{ij} \geq 0$ for all $i, j \in E$ such that $i \neq j$.
- $\sum_{j\in E} q_{ij} = 0$ for all $i \in E$.

We will use the word 'rate' interchangeably with 'intensity' in the remaining of this paper.

Definition 1.2 Associated with the intensity matrix is a semigroup of *transition matrices*,

$$P_t = [p_{ij}(t)]_{i,j\in E}\,, \quad t \geq 0,$$

where P_t is the unique non-negative solution to the *backward equation*

$$\frac{d}{dt}P_t = QP_t, \quad P_0 = I,$$

and I is the identity matrix $I = [\mathbf{1}_{\{i=j\}}]_{i,j\in E}$.

A direct consequence of the semigroup property is that $P_{s+t} = P_sP_t$, for all $s, t \geq 0$.

Let μ be a probability measure on E.

Definition 1.3 We say that $(X_t)_{t\geq 0}$ is a Markov chain on E with initial distribution μ and intensity matrix Q if $X_0 \overset{\mathscr{D}}{=} \mu$ and for all $n \in \mathbb{N}$, $0 \leq t_1 \leq \cdots \leq t_n, i_1, \ldots, i_n \in E$

$$\mathsf{P}\left(X_{t_{n+1}} = i_{n+1} \,|\, X_{t_1} = i_1, \ldots, X_{t_n} = i_n\right) \;=\; p_{i_n,i_{n+1}}(t_{n+1} - t_n).$$

In Chapter 3 we shall construct the Markov chain $(X_t)_{t\geq 0}$ from a specified intensity matrix Q. For general state spaces we must make further assumptions about Q, e.g. that it is non-explosive, but for finite state spaces these conditions are all satisfied. See Brémaud[4] for a thorough treatment of the subject.

Definition 1.4 A state $j \in E$ is *accessible* from $i \in E$ if there exists a $t > 0$ such that $p_{ij}(t) > 0$. If every state j is accessible from every other state i, then the chain, and its intensity matrix, are said to be *irreducible*.

Suppose that we start the chain at time 0 in state $\boldsymbol{i}$, i.e., the initial distribution is $\mu = \delta_{\boldsymbol{i}}$ for some $\boldsymbol{i} \in E$.

Definition 1.5 The return time to state $\boldsymbol{i}$ is defined by

$$R_{\boldsymbol{i}} = \inf\{t > 0 : X_t = \boldsymbol{i},\ \exists s < t : X_s \neq \boldsymbol{i}\}.$$

We say that a state $\boldsymbol{i}$ is *recurrent* if

$$\mathsf{P}_{\boldsymbol{i}}(R_{\boldsymbol{i}} < \infty) = 1.$$

and *transient* otherwise. If $\boldsymbol{i}$ is recurrent and $\mathsf{E}_{\boldsymbol{i}}[R_{\boldsymbol{i}}] < \infty$ we say that $\boldsymbol{i}$ is *positive recurrent.*

Theorem 1.1 *Let $(X_t)_{t\geq 0}$ be an irreducible Markov chain on E with intensity matrix Q. Then every state in E is either (positive) recurrent or transient, and we say that the chain is (positive) recurrent or transient, accordingly.*

Since we shall restrict our attention to Markov chains with finite state spaces, typically $E = \mathbb{Z}_+^{\boldsymbol{b}}$, we note the following theorem:

Theorem 1.2 *If E is finite and $(X_t)_{t\geq 0}$ is irreducible, then $(X_t)_{t\geq 0}$ is positive recurrent.*

Definition 1.6 An irreducible, positive recurrent Markov chain is called *ergodic.*

Theorem 1.3 *Let $(X_t)_{t\geq 0}$ be an ergodic Markov chain on E with initial distribution π and intensity matrix Q. Then the distribution of X_t does not depend on $t > 0$ iff π is a solution to*

$$\pi(\boldsymbol{i}) \sum_{\substack{\boldsymbol{j}\in E \\ \boldsymbol{j}\neq\boldsymbol{i}}} q_{\boldsymbol{ij}} = \sum_{\substack{\boldsymbol{j}\in E \\ \boldsymbol{j}\neq\boldsymbol{i}}} \pi(\boldsymbol{j}) q_{\boldsymbol{ji}}, \quad \boldsymbol{i} \in E, \tag{1.1}$$

or, more compactly,

$$\pi Q = 0.$$

Equation (1.1) is sometimes referred to as the *full (or global) balance equation.*

Definition 1.7 A probability measure π that is a solution to (1.1) is called the *stationary distribution* of the Markov chain $(X_t)_{t\geq 0}$ with intensity matrix Q.

The next theorem states that the limiting distribution of an ergodic Markov chain is the stationary distribution.

Theorem 1.4 *Let $(X_t)_{t\geq 0}$ be an irreducible, recurrent Markov chain on E with intensity matrix Q and stationary distribution π. Then for the induced semigroups $(P_t)_{t\geq 0}$ we have*

$$\lim_{t\to\infty} p_{\boldsymbol{ij}}(t) = \lim_{t\to\infty} \mathsf{P}_{\boldsymbol{i}}(X_t = \boldsymbol{j}) = \pi(\boldsymbol{j}), \quad \boldsymbol{j} \in E,$$

regardless of the initial state $\boldsymbol{i} \in E$.

We note that this implies that $\pi(\boldsymbol{j}) = \lim_{t\to\infty} \mathsf{P}_{\mu}(X_t = \boldsymbol{j})$ for $\boldsymbol{j} \in E$, regardless of initial distribution μ. The 'only if' part of Theorem 1.3 means that we have no hope of obtaining a sample from the stationary distribution by simply observing $(X_t)_{t\geq 0}$ at a predefined time t, unless $(X_t)_{t\geq 0}$ is stationary from the start. The coupling from the past algorithm in Chapter 4 will enable us to obtain such samples by a clever choice of starting times of $(X_t)_{t\geq 0}$.

1.2 COUPLING THEORY

Definition 1.8 On $E = \mathbb{Z}_+^d$ we use the standard partial ordering $\preceq$, defined by

$$\boldsymbol{i} \preceq \boldsymbol{j} \text{ if } i_k \leq j_k, \text{ for all } k = 1, \ldots, d. \tag{1.2}$$

We will use the convention of writing $\boldsymbol{i} \prec \boldsymbol{j}$ if $\boldsymbol{i} \preceq \boldsymbol{j}$ and $\boldsymbol{i} \neq \boldsymbol{j}$.

We note that this partial ordering causes the state space to have unique minimal element $\mathbf{0} = (0, \ldots, 0)$. Note that if $E = \mathbb{Z}_+^{\boldsymbol{b}}$ we also have

$$\mathbf{0} \prec \boldsymbol{i} \prec \boldsymbol{b}, \quad \forall \boldsymbol{i} \in E \backslash \{\mathbf{0}, \boldsymbol{b}\}.$$

Definition 1.9 Let X and X' be two random variables on E. We say that X is *stochastically dominated* by X', written $X \overset{\mathscr{D}}{\preceq} X'$, if

$$\mathsf{P}(X \preceq \boldsymbol{i}) \geq \mathsf{P}(X' \preceq \boldsymbol{i}), \text{ for all } \boldsymbol{i} \in E.$$

The following is Strassen's celebrated theorem.

Theorem 1.5 *For two random variables on E, X and X', then $X \overset{\mathscr{D}}{\preceq} X'$ if and only if there exists a coupling $(\widetilde{X}, \widetilde{X}')$ of X and X' such that*

$$\widetilde{X} \preceq \widetilde{X}' \text{ a.s., where } \widetilde{X} \overset{\mathscr{D}}{=} X, \widetilde{X}' \overset{\mathscr{D}}{=} X'.$$

Chapter 2

Queueing systems, Jackson networks

For a queueing network with $d \geq 1$ stations, let $X_k(t)$ be the number of customers in queue $k = 1, \ldots, d$ at time t, and let

$$X_t = (X_1(t), \ldots, X_d(t))$$

denote the state of the queueing network at that time. One distinguishes between open and closed networks. In a *closed* queueing network the total number of customers in the system is constant, i.e., the state space E_c of X_t is

$$E_c = \left\{ (i_1, \ldots, i_d) \in \mathbb{Z}_+^d \colon \sum_{k=1}^{d} i_k = \text{some constant} \right\}.$$

We shall only consider networks where customers may enter and eventually leave the system. Such a network we call *open*, and we model $(X_t)_{t \geq 0}$ as a continuous time Markov chain with the special structure defined below. It is intuitively clear what we mean by open, but a formal definition is found in (2.3).

2.1 OPEN JACKSON NETWORKS

Let us first define the general birth and death process we shall use as a model of our queueing networks.

Definition 2.1 Let $(X_t)_{t\geq 0}$ be a *general birth and death process* on E, where

$$E = \mathbb{Z}_+^d$$

or some finite subset thereof. This means that $(X_t)_{t\geq 0}$ is a Markov chain on E and its intensity matrix $Q = [q_{\boldsymbol{ij}}]_{\boldsymbol{i},\boldsymbol{j}\in E}$ has non-zero transition intensities $q_{\boldsymbol{ij}}$, $\boldsymbol{i} \neq \boldsymbol{j}$, only for transitions of the form

1. $\boldsymbol{i}$ to $\boldsymbol{i} + e_k$, which is interpreted as an external arrival at queue k. These transitions occur with intensity $q_{\boldsymbol{i},\boldsymbol{i}+e_k} = \beta_k(\boldsymbol{i})$.
2. $\boldsymbol{i}$ to $\boldsymbol{i} - e_k$, a departure from queue k that leaves the network, which occur with intensity $\delta_k(\boldsymbol{i})$. That is, we have $q_{\boldsymbol{i}+e_k,\boldsymbol{i}} = \delta_k(\boldsymbol{i} + e_k)$. (2.1)
3. $\boldsymbol{i}$ to $\boldsymbol{i} + (e_m - e_k)$, a transfer of a customer from queue k to m, which occur with intensity $\gamma_{km}(\boldsymbol{i})$. That is, the elements of the intensity matrix are $q_{\boldsymbol{i}+e_k,\boldsymbol{i}+e_m} = \gamma_{km}(\boldsymbol{i} + e_k)$,

where $k, m \in \{1, \ldots, d\}$.

In our networks the arrival process at station k is a Poisson process with intensity β_k, and the service times at station k are independent and exponentially distributed. The service times at different stations are independent of each other and of the arrival processes.

Definition 2.2 A *Jackson network* is a queueing network with the structure above and is modelled by the general birth and death process $(X_t)_{t\geq 0}$ on $\mathbb{Z}_+^d$ with transitions as in (2.1), where

1. $\beta_k(\boldsymbol{i}) = \beta_k$,
2. $\delta_k(\boldsymbol{i}) = \delta_k(i_k) = \mu_k(i_k)p_k$, and (2.2)
3. $\gamma_{km}(\boldsymbol{i}) = \gamma_{km}(i_k) = \mu_k(i_k)p_{km}$,

when $k, m \in \{1, \ldots, d\}$ and p_{km} is the probability that a departure from station k goes to station m. We will use the convention to denote $\beta_k(\boldsymbol{i})$ by β_k when the former does not depend on $\boldsymbol{i}$, i.e., is constant. Consequently, $\gamma_{km}(i_k)$ emphasize that $\gamma_{km}(\boldsymbol{i})$ depends only on $\boldsymbol{i}$ through i_k.

The probability that a departure from station k leaves the system is

$$p_k = 1 - \sum_{m=1}^{d} p_{km},$$

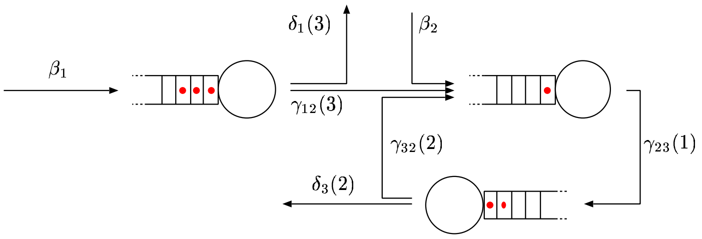

Figure 2.1. *A simple 3-station queueing network with state-dependent service rates,* $\mu_k(\boldsymbol{i})$. *In this model* $\beta_3 = 0$ *since we do not have external arrivals to station 3, and likewise* $\delta_2 = 0$ *and* $\gamma_{13} = \gamma_{21} = \gamma_{31} = 0$.

We shall also assume that the service rate at station k can be written

$$\mu_k(n) = \sum_{i=1}^{n} \nu_k(i) \quad \text{where } \nu_k(n) \geq 0 \text{ for } n \geq 1.$$

The definition of a Jackson network is rather general in the sense that it allows the service rate at station k, μ_k, to depend on the number of customers, i_k, present. This covers the situation when a queue is attended by more than one server. An example of a Jackson network is shown in Figure 2.1.

Definition 2.3 An *open Jackson network* is a Jackson network where the routing probabilities p_{km}, where $k, m \in \{1, \ldots, d\}$, satisfy:

- For all $m \in \{1, \ldots, d\}$ there exist $k, k_1, \ldots, k_l \in \{1, \ldots, d\}$ such that
$$\beta_k \cdot p_{kk_1} \cdot p_{k_1 k_2} \cdots p_{k_l m} > 0,$$
which means that every queue m is (indirectly) supplied with customers from the outside.
- For all $k \in \{1, \ldots, d\}$ there exist $m, k_1, \ldots, k_l \in \{1, \ldots, d\}$ such that (2.3)
$$p_{kk_1} \cdot p_{k_1 k_2} \cdots p_{k_l m} \cdot \delta_m(e_m) > 0,$$
which means that every queue k has an (indirect) way of expunging customers from the network.

In the following we shall assume that every Jackson network is open and thus omit the word ‘open’ from the definitions.

A common type of Jackson network occurs when we have single-server queues.

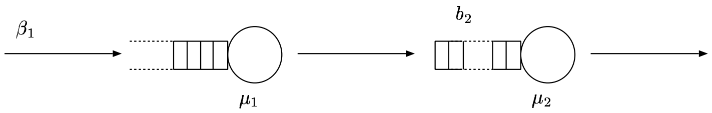

Figure 2.2. *This simple network (tandem queue) illustrates how blocking of customers can occur when a station has a finite waiting space.*

Definition 2.4 A *Jackson network with single-server queues* is modelled by the general birth and death process $(X_t)_{t\geq 0}$ on $\mathbb{Z}_+^d$ with transition intensities

$$\begin{aligned} &1.\ \beta_k(\boldsymbol{i}) = \beta_k \\ &2.\ \delta_k(\boldsymbol{i}) = \delta_k \mathbf{1}_{\{i_k>0\}} \\ &3.\ \gamma_{km}(\boldsymbol{i}) = \gamma_{km} \mathbf{1}_{\{i_k>0\}}, \end{aligned} \tag{2.4}$$

where $k, m \in \{1, \ldots, d\}$.

We shall focus on queueing networks with a structure similar to the Jackson network, so we refer to networks defined by Definition 2.2 as *standard Jackson networks* whenever distinction is necessary. The networks we shall consider consist of queues with finite buffers, and we first give an informal description of such a system.

Let us first consider a network of two queues in series. The first queue has Poisson arrivals at intensity β_1 and a server which works with rate μ_1. The second queue has a finite waiting space that can accommodate at most b_2 customers, and its server works with rate μ_2. See Figure 2.2. We impose a blocking mechanism for the first server, such that a customer finishing service at the first station, is looped back to the first server and receives a new independent service time whenever the secondary queue is full. In queueing theory context this is commonly referred to as *repetitive blocking*.

To generalise further, consider a network of d queues where queue k has a finite waiting space of size b_k. A customer that finishes service at station k choses to transfer to station m with probability p_{km} and leaves the network with probability p_k. Suppose that the chosen queue m is full, i.e., $i_m = b_m$, then the customer remains at station k an exponential amount of time before a new routing attempt is made. Note that it is possible for a retrying customer to chose another station than the original one. We say that we have a *repetitive blocking mechanism* with *random destination*, as opposed to a 'fixed destination' where the destination is determined after the first service time and cannot be altered later.

We are now ready to formally define our first model of such a network.

Definition 2.5 Let $\boldsymbol{b} \in \mathbb{N}^d$ be a vector of bounds and denote the state space $\mathbb{Z}_+^{\boldsymbol{b}}$ by $E_{\boldsymbol{b}}$. A *bounded Jackson network of Type 1* is modelled by the general birth and death process $(X_t)_{t \geq 0}$ on $E_{\boldsymbol{b}}$ with transition intensities

$$\begin{aligned} &1.\ \beta_k(\boldsymbol{i}) = \beta_k \mathbf{1}_{\{i_k < b_k\}} \\ &2.\ \delta_k(\boldsymbol{i}) = \mu_k(i_k) p_k \\ &3.\ \gamma_{km}(\boldsymbol{i}) = \mu_k(i_k) p_{km} \mathbf{1}_{\{i_m < b_m\}}, \end{aligned} \tag{2.5}$$

where $k, m \in \{1, \ldots, d\}$.

Compare with (2.2).

A slightly different network will be obtained if we let those customers who depart from station k, and are heading towards station m, *leave the system* when they find queue m full, i.e., $i_m = b_m$. This is the essence behind:

Definition 2.6 Let $\boldsymbol{b} \in \mathbb{N}^d$ be a vector of bounds and denote the state space $\mathbb{Z}_+^{\boldsymbol{b}}$ by $E_{\boldsymbol{b}}$. A *bounded Jackson network of Type 2* is modelled by the general birth and death process $(X_t)_{t \geq 0}$ on $E_{\boldsymbol{b}}$ with transition intensities

$$\begin{aligned} &1.\ \beta_k(\boldsymbol{i}) = \beta_k \mathbf{1}_{\{i_k < b_k\}} \\ &2.\ \delta_k(\boldsymbol{i}) = \mu_k(i_k) \left(p_k + \textstyle\sum_{m=1}^{d} p_{km} \mathbf{1}_{\{i_m = b_m\}} \right) \\ &3.\ \gamma_{km}(\boldsymbol{i}) = \mu_k(i_k) p_{km} \mathbf{1}_{\{i_m < b_m\}}, \end{aligned} \tag{2.6}$$

where $k, m \in \{1, \ldots, d\}$.

The difference between a Type 1 and a Type 2 bounded Jackson network may seem subtle regarding the definitions, but the difference in behaviour is substantial. For instance, the number of customers in a Type 1 network dominates the corresponding number in a network of Type 2, in a sense that soon will be made precise (glance ahead at Figure 2.4).

2.2 STATIONARY DISTRIBUTION

For these models of queueing networks our next step is to show how one can obtain the stationary distribution. First we note that we are only considering open Jackson networks, and we quote the following theorem from Walrand[13]:

Theorem 2.1 *A Jackson network is irreducible if and only if it is open.*

This combined with Theorem 1.2 tells us that a bounded Jackson network is ergodic. For a standard Jackson network a necessary and sufficient condition for positive recurrence (and hence ergodicity) is given in Theorem 2.2.

For a standard Jackson network the stationary distribution π of $(X_t)_{t\geq 0}$ can be calculated directly. Let the intensities be as in (2.2), and let λ_k be the net flow of customers entering queue k. If every customer eventually is served, then stationarity should mean that the average flow of customers finishing service at station k must equal λ_k. This heuristic argument tells us that under stationarity, we should have

$$\lambda_k \;=\; \beta_k + \sum_{m=1}^{d} \lambda_m p_{mk}, \quad k \in 1, \dots, d. \tag{2.7}$$

This system of equations is sometimes called the *detailed balance equations*, and when the chain $(X_t)_{t\geq 0}$ is irreducible, it has a unique nonnegative solution $(\lambda_1, \dots, \lambda_d)$.

Define normalising factors by

$$\kappa_k \;=\; 1 + \sum_{i_k=1}^{\infty} \prod_{i=1}^{i_k} \frac{\lambda_k}{\mu_k(i)}, \quad k = 1, \dots, d.$$

Theorem 2.2 *An open standard Jackson network $(X_t)_{t\geq 0}$ is positive recurrent (and hence ergodic) iff $\kappa_k < \infty$ for $k = 1, \dots, d$.*

The stationary distribution π of $(X_t)_{t\geq 0}$ is given by

$$\pi(\boldsymbol{i}) = \prod_{k=1}^{d} \pi_k(i_k), \text{ where } \pi_k(i_k) = \frac{1}{\kappa_k} \prod_{i=1}^{i_k} \frac{\lambda_k}{\mu_k(i)}. \tag{2.8}$$

A brute-force proof of this statement is to check that this particular expression for the stationary distribution is a solution to the full balance equation (1.1).

Definition 2.7 For a single-server Jackson network, $\mu_k(i_k) = \mu_k \mathbf{1}_{\{i_k>0\}}$, we define the *traffic intensity* ρ by

$$\rho_k = \frac{\lambda_k}{\mu_k}, \quad k = 1, \dots, d, \quad \rho = (\rho_1, \dots, \rho_d).$$

If (X_t) is ergodic, i.e., if $\rho_k < 1$ for $k = 1, \dots, d$, then the stationary distribution π is given by

$$\pi(\boldsymbol{i}) = \prod_{k=1}^{d} \pi_k(i_k) = \prod_{k=1}^{d} \rho_k^{i_k}(1 - \rho_k).$$

Unfortunately, for bounded Jackson networks the explicit solution for the stationary distribution is hard to find since the balance equation (1.1) is hard to solve.

Definition 2.8 When $\pi(\boldsymbol{i}) = \pi_1(i_1) \cdots \pi_d(i_d)$, the stationary distribution is said to be on a *product form*. A network of queues with such a distribution is said to be a *product-form network*.

A remarkable feature of such networks is that under stationarity at any given time the number of customers at a station is independent of the rest of the network. This is not the case for bounded Jackson networks. A network with finite waiting spaces at the stations has product form, if for instance customers that find a station full, immediately skip that station and continue as if they had received service there. Such a network is sometimes said to have a jump-over blocking scheme. According to [3] a formal proof of the product form of π can be found in Aarssen and van Dijk[1].

For the bounded Jackson networks neither the departure process of served customers nor the departure process of rejected customers is Poisson, but we note the following.

If the net-arrival process at station k in a bounded Jackson network of Type 2 is Poissonian, i.e., there is no inflow of customers from any* part of the network, then $\lambda_k = \beta_k$ and we can calculate the marginal distribution $\pi_k^{(2)}$ of a bounded Jackson network of Type 2 as in (2.8) but with

$$\kappa_k = 1 + \sum_{i_k=1}^{b_k} \prod_{i=1}^{i_k} \frac{\lambda_k}{\mu_k(i)}, \quad \pi_k^{(2)}(i_k) = \frac{1}{\kappa_k} \prod_{i=1}^{i_k} \frac{\lambda_k}{\mu_k(i)}, \quad i_k = 0, \ldots, b_k. \tag{2.9}$$

We will use this distribution for comparison with the estimated distributions obtained by simulation.

Before we leave the discussion on the stationary distribution, we note that in a stationary standard Jackson network the fraction of time during which queue k contains i_k customers is $\pi_k(i_k)$. The fraction of *arrivals* to queue k who see i_k customers there is also $\pi_k(i_k)$. For Poisson arrivals this feature is known as the PASTA†-property of a Poisson process, but it is still valid for a standard Jackson network in equilibrium where the net-arrival processes are *not* Poisson. This result is called the Arrival theorem for Jackson networks. Boucherie and van Dijk[3] contains a survey of when the arrival theorem is valid for queueing networks with blocking.

*Not even from k itself.

†Poisson Arrivals See Time Averages; see Wolff[16].

2.3 DOMINATION RESULTS

To be able to construct the algorithm for obtaining samples from the stationary distribution π we need some domination results. Impose on $E_{\boldsymbol{b}}$ the partial ordering $\preceq$ of (1.2). The element $\mathbf{0}$ represents an empty queueing network, and $\boldsymbol{b}$ is the state when every queueing position is occupied and every server is busy. Certainly, every other configuration of customers must necessarily lie between these two states.

Lemma 2.3 *Let $(X_t)_{t\geq 0}$ be a bounded Jackson network on $E_{\boldsymbol{b}}$ with intensities as in* (2.5) *or* (2.6). *Then for $\boldsymbol{i}, \boldsymbol{j} \in E_{\boldsymbol{b}}$, such that $\boldsymbol{i} \preceq \boldsymbol{j}$, we have*

$$\begin{aligned} \beta_k(\boldsymbol{i}) &\geq \beta_k(\boldsymbol{j}), \\ \delta_k(\boldsymbol{i}) &\leq \delta_k(\boldsymbol{j}), \text{ and} \\ \gamma_{km}(\boldsymbol{i})\mathbf{1}_{\{j_m<b_m\}} &\leq \gamma_{km}(\boldsymbol{j}). \end{aligned}$$

Proof: Direct identification yields

$$\begin{aligned} \beta_k(\boldsymbol{i}) &= \beta_k\mathbf{1}_{\{i_k<b_m\}} \geq \beta_k\mathbf{1}_{\{j_k<b_m\}} = \beta_k(\boldsymbol{j}), \\ \delta_k^{(1)}(\boldsymbol{i}) &= \mu_k(i_k)p_k \leq \mu_k(j_k)p_k = \delta_k^{(1)}(\boldsymbol{j}), \\ \delta_k^{(2)}(\boldsymbol{i}) &= \mu_k(i_k)\Big(p_k + \textstyle\sum_{m=1}^d p_{km}\mathbf{1}_{\{i_m=b_m\}}\Big) \\ &\leq \mu_k(j_k)\Big(p_k + \textstyle\sum_{m=1}^d p_{km}\mathbf{1}_{\{j_m=b_m\}}\Big) = \delta_k^{(2)}(\boldsymbol{j}), \text{ and} \\ \gamma_{km}(\boldsymbol{j}) &= \mu_k(j_k)p_{km}\mathbf{1}_{\{j_m<b_m\}}\mathbf{1}_{\{i_m<b_m\}} \geq \mu_k(i_k)p_{km}\mathbf{1}_{\{j_m<b_m\}}\mathbf{1}_{\{i_m<b_m\}} \\ &= \gamma_{km}(\boldsymbol{i})\mathbf{1}_{\{j_m<b_m\}}, \end{aligned}$$

where the superscript on the departure rate signifies the type of the bounded Jackson network. Note that $\delta_k^{(1)}(\boldsymbol{i}) \leq \delta_k^{(2)}(\boldsymbol{i})$ for all $\boldsymbol{i} \in E_{\boldsymbol{b}}$. □

Theorem 2.4 *If $(X_t)_{t\geq 0}$ is a bounded Jackson network with $X_0 = \mathbf{0}$, then X_t is stochastically increasing in t, i.e.,*

$$X_t \overset{\mathscr{D}}{\preceq} X_{t+s} \textit{ for all } t, s \geq 0.$$

Proof: Let $(X_t)_{t\geq 0}$ and $(X_t')_{t\geq 0}$ be two independent copies of a bounded Jackson network. We construct a coupling $(\widetilde{X}_t, \widetilde{X}_t')$ of (X_t) and (X_t) such that if $\widetilde{X}_0 \preceq \widetilde{X}_0'$, then

$$\widetilde{X}_t \preceq \widetilde{X}_t', \text{ for all } t \geq 0.$$

This coupling is obtained by letting the bivariate Markov chain $(\widetilde{X}_t, \widetilde{X}_t')$

have, when $(\widetilde{X}_t, \widetilde{X}'_t) = (\boldsymbol{i}, \boldsymbol{j})$, transition intensities as follows

$$\begin{array}{lll}
\text{from} & \text{to} & \text{with intensity} \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}+e_k, \boldsymbol{j}+e_k) & \beta_k(\boldsymbol{j}) \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}+e_k, \boldsymbol{j}) & \beta_k(\boldsymbol{i}) - \beta_k(\boldsymbol{j}) \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}-e_k, \boldsymbol{j}-e_k) & \delta_k(\boldsymbol{i}) \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}, \boldsymbol{j}-e_k) & \delta_k(\boldsymbol{j}) - \delta_k(\boldsymbol{i}) \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}+e_m-e_k, \boldsymbol{j}+e_m-e_k) & \gamma_{km}(\boldsymbol{i})\mathbf{1}_{\{j_m<b_m\}} \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}+e_m-e_k, \boldsymbol{j}) & \gamma_{km}(\boldsymbol{i})\mathbf{1}_{\{j_m=b_m\}} \\
(\boldsymbol{i},\boldsymbol{j}) & (\boldsymbol{i}, \boldsymbol{j}+e_m-e_k) & \gamma_{km}(\boldsymbol{j}) - \gamma_{km}(\boldsymbol{i})\mathbf{1}_{\{j_m<b_m\}}.
\end{array} \tag{2.10}$$

It is clear that $\widetilde{X}_t \overset{\mathscr{D}}{=} X_t$ and $\widetilde{X}'_t \overset{\mathscr{D}}{=} X'_t$, and $\widetilde{X}_t \preceq \widetilde{X}'_t$ for all $t \geq 0$ if $\widetilde{X}_0 \preceq \widetilde{X}'_0$. Now, fix $s > 0$ and let $\widetilde{X}_0 = \mathbf{0}$ and $\widetilde{X}'_0 = \widetilde{X}_s \overset{\mathscr{D}}{=} X_s$. This makes $\widetilde{X}_t \preceq \widetilde{X}_{t+s}$ and Theorem 1.5 gives the statement. □

An explicit construction of the coupling (2.10) is the subject of Chapter 3, and when using it we will start the bivariate chain $(\widetilde{X}_t, \widetilde{X}'_t)_{t\geq 0}$ in $(\widetilde{X}_0, \widetilde{X}'_0) = (\mathbf{0}, \boldsymbol{b})$.

Theorem 2.5 *Let $(X_t^{(1)})_{t\geq 0}$ be a bounded Jackson network of Type 1 and $(X_t^{(2)})_{t\geq 0}$ one of Type 2, where both have the same arrival intensities β_k, service rates $\mu_k(i_k)$ and routing probabilities p_{km}. Then*

$$X_t^{(2)} \overset{\mathscr{D}}{\preceq} X_t^{(1)} \textit{ for all } t \geq 0.$$

Proof: From the definitions we have that $\beta^{(1)}(\boldsymbol{i}) = \beta^{(2)}(\boldsymbol{i})$, and $\gamma_{km}^{(1)}(\boldsymbol{i}) = \gamma_{km}^{(2)}(\boldsymbol{i})$, $\forall \boldsymbol{i} \in E_{\boldsymbol{b}}$, $k \in \{1, \ldots, d\}$. We noted in the proof of Lemma 2.3 that $\delta_k^{(1)}(\boldsymbol{i}) \leq \delta_k^{(2)}(\boldsymbol{i})$ for all $\boldsymbol{i} \in E_{\boldsymbol{b}}$ and $k = 1, \ldots, d$, and in the same manner as the proof of Theorem 2.4 we can construct a coupling $(\widetilde{X}_t, \widetilde{X}'_t)$ of $(X_t^{(1)})$ and $(X_t^{(2)})$, such that

$$\widetilde{X}_t \overset{\mathscr{D}}{=} X_t^{(2)} \text{ and } \widetilde{X}'_t \overset{\mathscr{D}}{=} X_t^{(1)} \text{ (note the order)}, \quad t \geq 0,$$

and if $\widetilde{X}_0 \preceq \widetilde{X}'_0$, then

$$X_t^{(2)} \overset{\mathscr{D}}{=} \widetilde{X}_t \preceq \widetilde{X}'_t \overset{\mathscr{D}}{=} X_t^{(1)}, \text{ for all } t \geq 0.$$

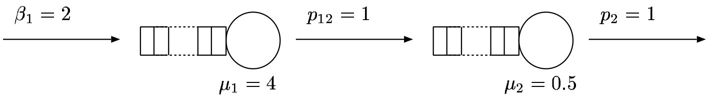

Figure 2.3. *This simple network (tandem queue) illustrates the difference between a Type 1 and a Type 2 bounded Jackson network. The secondary service rate is very small, so in a Type 1 Jackson network congestion will occur at the first station. The buffer sizes are given by* $\boldsymbol{b} = (10, 10)$.

Take for instance

from	to	with intensity
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i} + e_k, \boldsymbol{j} + e_k)$	$\beta_k^{(1)}(\boldsymbol{j})$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i} + e_k, \boldsymbol{j})$	$\beta_k^{(2)}(\boldsymbol{i}) - \beta_k^{(1)}(\boldsymbol{j})$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i} - e_k, \boldsymbol{j} - e_k)$	$\delta_k^{(1)}(\boldsymbol{i})$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i}, \boldsymbol{j} - e_k)$	$\delta_k^{(1)}(\boldsymbol{j}) - \delta_k^{(1)}(\boldsymbol{i})$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i} - e_k, \boldsymbol{j})$	$\delta_k^{(2)}(\boldsymbol{i}) - \delta_k^{(1)}(\boldsymbol{i})$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i} + e_m - e_k, \boldsymbol{j} + e_m - e_k)$	$\gamma_{km}^{(2)}(\boldsymbol{i})\mathbf{1}_{\{j_m < b_m\}}$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i} + e_m - e_k, \boldsymbol{j})$	$\gamma_{km}^{(2)}(\boldsymbol{i})\mathbf{1}_{\{j_m = b_m\}}$
$(\boldsymbol{i}, \boldsymbol{j})$	$(\boldsymbol{i}, \boldsymbol{j} + e_m - e_k)$	$\gamma_{km}^{(1)}(\boldsymbol{j}) - \gamma_{km}^{(2)}(\boldsymbol{i})\mathbf{1}_{\{j_m < b_m\}}$

□

The theorem states that a Jackson network of Type 1 tends to have more customers *at every station*. Consider for instance the network in Figure 2.3, and note that this is one of the situations where we actually can calculate the stationary distribution of the number of customers at the first station, π_1, for the bounded Jackson network of Type 2. Since the first station is a single-server queue, the traffic intensity is $\rho = 1/2$ and $\pi_1^{(2)}$, according to (2.9), is

$$\pi_1^{(2)}(\boldsymbol{i}) = \frac{1-\rho}{1-\rho^{b_1+1}}\rho^{i_k} = \frac{1}{1-2^{-11}}\left(\frac{1}{2}\right)^{i_k+1} \qquad i_k = 0, \dots, 10.$$

Figure 2.4 shows the estimated stationary distributions $\widehat{\pi}_1^{(1)}$ and $\widehat{\pi}_1^{(2)}$ for the network in Figure 2.3.

2.4 COMMENTS ON THE MODELS

We saw that the number of customers in a bounded Jackson network of Type 1 dominates the corresponding number in a network of Type 2. This can be generalised a little bit further by defining a bounded Jackson

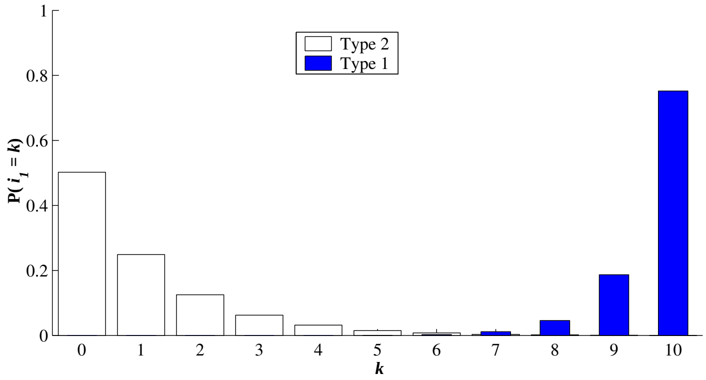

Figure 2.4. *The estimated distributions $\widehat{\pi}_1^{(1)}$ and $\widehat{\pi}_1^{(2)}$ of the bounded Jackson network in Figure 2.3 obtained from 50,000 samples. The dark bars correspond to estimates of $\pi_1^{(1)}(k)$, $k = 1, \dots, 10$, which dominate the estimates of $\pi_1^{(2)}(k)$ (white bars).*

network where *some* stations may dispose of the customers they cannot transfer. Let $X^{(n)}$ be a bounded Jackson network of type n, $n = 1, 2$, and X a general birth and death process on $E_{\boldsymbol{b}}$ where

$$\beta_k(\boldsymbol{i}) = \beta_k^{(1)}(\boldsymbol{i}), \gamma_{km}(\boldsymbol{i}) = \gamma_{km}^{(1)}(\boldsymbol{i}), \quad \forall k, m \in \{1, \dots, d\},$$

and

$$\delta_k(\boldsymbol{i}) = \delta_k^{(1)}(\boldsymbol{i}) \text{ for } k \in S, \ \delta_k(\boldsymbol{i}) = \delta_k^{(2)}(\boldsymbol{i}) \text{ for } k \notin S,$$

where S is some subset of $\{1, \dots, d\}$. Then X is a network of a mixture of Type 1 and Type 2 stations. In the same way as we proved Theorem 2.5 one can show that

$$X_t^{(2)} \overset{\mathcal{D}}{\preceq} X_t \overset{\mathcal{D}}{\preceq} X_t^{(1)} \text{ for all } t \geq 0$$

if $X_0^{(2)} \overset{\mathcal{D}}{\preceq} X_0 \overset{\mathcal{D}}{\preceq} X_0^{(1)}$.

A remark concerning a bounded Jackson network of Type 2 is that by definition a customer decides where to go next *before* departing from the present station. Suppose that $p_{kk} > 0$ for some $k \in \{1, \dots, d\}$. When queue k is full, no customers will recycle from k to k even if their own departure would leave room for their admittance. If this is not desirable, then point 2 in Definition 2.6 can be replaced by

$$\delta_k(\boldsymbol{i}) = \mu_k(i_k)\Big(p_k + \sum_{\substack{m=1 \\ m \neq k}}^{d} p_{km} \mathbf{1}_{\{i_m = b_m\}}\Big),$$

and the theorems in this paper are still valid.

Although we have not especially treated the multi-server case we would like to emphasise the versatility of the model. Recall that $\mu_k(\boldsymbol{i}) = \sum_{n=1}^{i_k} \nu_k(n)$. If we take for instance

$$\nu_k(1) > 0, \quad \nu_k(\lfloor 2b/3 \rfloor) > 0,$$

and all other $\nu_k(n)$ are zero, we have a situation where station k works at rate $\nu_k(1)$ until the buffer is two-thirds full. At that point a spare server is activated, increasing the service speed to $\nu_k(1) + \nu_k(\lfloor 2b/3 \rfloor)$.

2.4.a MARKOVIAN ROUTING

The way customers flow in a network is regulated by a routing mechanism, a protocol regulating which paths the customers should use to travel from one station to another. With this in mind, Markovian routing is very crude. For instance, as a model of a computer network it may not be adequate if a message (customer) originating from, say, a computer (station) should not be allowed to cycle infinitely through the system.

A way to solve this problem is to assign classes to customers, where each class represents a different routing scheme, i.e., letting the transition probability p_{km} depend on the current class of the customer. To generalise further, it is possible to allow customers to change classes during their passage through the networks, i.e., the transition probabilities take the form

$$p_{km}^{jl},$$

the probability that a j-type customer when finishing service at station k goes to station m and gets assigned the class l.

Chapter 3

Poisson calculus

This chapter is devoted to the construction of a d-dimensional birth and death process $(X_t)_{t\geq 0}$ on $E_{\boldsymbol{b}}$, describing a bounded Jackson network treated in Chapter 2. This will eventually lead us to the construction of the coupling used in (2.10).

3.1 POISSON REPRESENTATION

Take $d \in \mathbb{N}$, $\boldsymbol{b} \in \mathbb{N}^d$ and let $E_{\boldsymbol{b}} = \mathbb{Z}_+^{\boldsymbol{b}}$. Let $N_{\boldsymbol{ij}}$, $\boldsymbol{i}, \boldsymbol{j} \in E_{\boldsymbol{b}}$, be independent, homogeneous Poisson processes on $\mathbb{R}$, with intensities $q_{\boldsymbol{ij}}$, respectively. That is:

1. For all $n \geq 1$ and $s_1 < t_1 < \cdots < s_n < t_n \in \mathbb{R}_+$, the random variables $N_{\boldsymbol{ij}}(s_k, t_k]$, $k = 1, \ldots, n$, are independent.
2. For $s < t$, $N_{\boldsymbol{ij}}(s, t]$ is Poisson distributed with expected value $(t - s)q_{\boldsymbol{ij}}$.

This is rather over-parameterised since most of the intensities $q_{\boldsymbol{ij}}$ will be zero, but it is a convenient notation. The intensities $q_{\boldsymbol{ij}}$, for $\boldsymbol{i}, \boldsymbol{j} \in E_{\boldsymbol{b}}$, are those of a general birth and death process on $E_{\boldsymbol{b}}$ found in (2.1). Define the shifted processes $\theta_s N_{\boldsymbol{ij}}$ by

$$\theta_s N_{\boldsymbol{ij}}(0, t] = N_{\boldsymbol{ij}}(s, s + t], \quad s, t \geq 0.$$

We will now construct a continuous time Markov chain $(X_t)_{t\geq 0}$ from transitions of a discrete time chain $(X_n)_{n\geq 0}$ and a sequence of time-points $(T_n)_{n\geq 0}$, which in turn are generated by the processes $N_{\boldsymbol{ij}}$, $\boldsymbol{i}, \boldsymbol{j} \in E_{\boldsymbol{b}}$.

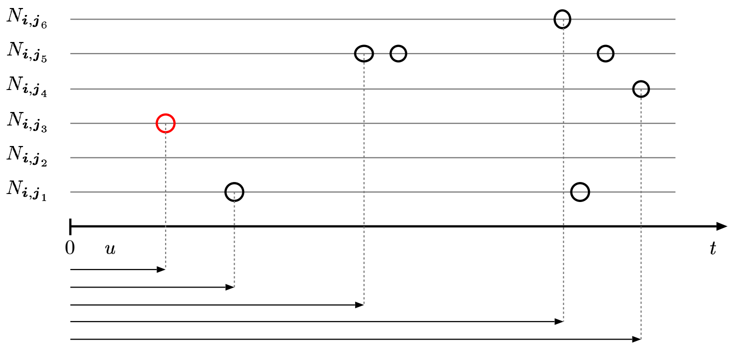
$N_{\boldsymbol{i},\boldsymbol{j}_6}$
$N_{\boldsymbol{i},\boldsymbol{j}_5}$
$N_{\boldsymbol{i},\boldsymbol{j}_4}$
$N_{\boldsymbol{i},\boldsymbol{j}_3}$
$N_{\boldsymbol{i},\boldsymbol{j}_2}$
$N_{\boldsymbol{i},\boldsymbol{j}_1}$
0
u
t

Figure 3.1. *When $X_0 = \boldsymbol{i}$, the pair X_1, Z_1 is determined by the first arrival in one of the Poisson processes $N_{\boldsymbol{ij}}$, $\boldsymbol{j} \in E_{\boldsymbol{b}}$. The figure shows six of those processes, $N_{\boldsymbol{i},\boldsymbol{j}_1}, \ldots, N_{\boldsymbol{i},\boldsymbol{j}_6}$, and the first arrival was in $N_{\boldsymbol{i},\boldsymbol{j}_3}$ at time $Z_1 = u$, thus $X_1 = \boldsymbol{j}_3$, $T_1 = T_0 + u$.*

Choose an initial state, say $\boldsymbol{i} \in E_{\boldsymbol{b}}$, by letting $T_0 = 0$ and $X_0 = \boldsymbol{i}$, and let

$$X_t = X_n \text{ for } t \in [T_n, T_{n+1}),$$

where $(X_n)_{n\geq 1}$ and $(T_n)_{n\geq 1}$ are given recursively as follows:

$$\begin{aligned} Z_{n+1} &= \inf_{u\geq 0}\left\{ \sum_{\boldsymbol{j}\in E_b} \theta_{T_n} N_{X_n,\boldsymbol{j}}(0,u] \;= 1 \right\} \\ T_{n+1} &= T_n + Z_{n+1} \\ X_{n+1} &= \{\boldsymbol{j}\colon \theta_{T_n} N_{X_n,\boldsymbol{j}}(0, Z_{n+1}] = 1\}\,. \end{aligned}$$

Consult Figure 3.1.

Let $q(\boldsymbol{i}) = \sum_{\substack{\boldsymbol{j}\in E_{\boldsymbol{b}} \\ \boldsymbol{j}\neq \boldsymbol{i}}} q_{\boldsymbol{ij}}$. The joint distribution of X_{n+1} and Z_{n+1}, conditioned on the values of $X_n = \boldsymbol{i}$ and $T_n = s$ is then

$$\begin{aligned} \mathsf{P}(X_{n+1} = \boldsymbol{j}, Z_{n+1} > t | X_n = \boldsymbol{i}, T_n = s) \;&= \\ = \;& \mathsf{P}(X_{n+1} = \boldsymbol{j} | X_n = \boldsymbol{i}) \mathsf{P}(Z_{n+1} > t | X_n = \boldsymbol{i}) \qquad (3.1) \\ = \;& \frac{q_{\boldsymbol{ij}}}{q(\boldsymbol{i})} \mathrm{e}^{-tq(\boldsymbol{i})}. \end{aligned}$$

Note that the distribution does not depend on T_n, that X_{n+1} and Z_{n+1} are conditionally independent, and that $Z_{n+1}|X_n$ is exponentially distributed with mean $1/q(X_n)$. See Brémaud[4] for details.

By the construction, these properties imply that the evolution of the process from time s, $(X_{t+s})_{t\geq 0}|X_s$, and the process up to time s, $(X_t)_{t=0}^{s}|X_0$, are independent, and

$$(X_t)_{t+s\geq 0}|X_s \overset{\mathscr{D}}{=} (X_t)_{t\geq 0}|X_0.$$

Thus, $(X_t)_{t\geq 0}$ is a homogeneous Markov chain and the intensities of transitions, for $\boldsymbol{i} \neq \boldsymbol{j}$, are

$$\lim_{t\searrow 0} \frac{1}{t}\mathsf{P}(X_s = \boldsymbol{i}, X_{s+t} = \boldsymbol{j}) = \lim_{t\searrow 0} \frac{1}{t}\mathsf{P}(X_0 = \boldsymbol{i}, X_t = \boldsymbol{j}) = q_{\boldsymbol{ij}}.$$

Identifying the intensities $q_{\boldsymbol{ij}}$ as those given in (2.1) we see that $(X_t)_{t\geq 0}$ have the right transition intensities.

We finish this section by noting that according to (3.1), X_{n+1} can be written

$$X_{n+1} = \phi(X_n, U_{n+1}), \quad n \geq 0,$$

for some deterministic function ϕ and $(U_n)_{n\geq 1}$ a sequence of random variables such that

$$\mathsf{P}(\phi(X_n, U_{n+1}) = \boldsymbol{j}|X_n) \ = \ \frac{q_{X_n,\boldsymbol{j}}}{q(X_n)}. \tag{3.2}$$

In the next section we shall derive an expression for ϕ and construct the Poisson processes $N_{\boldsymbol{ij}}$.

3.2 THE TWO-DIMENSIONAL POISSON EMBEDDING

Our next task is to construct the Poisson processes corresponding to those $N_{\boldsymbol{ij}}$ with non-zero intensities. To this end we shall use a two-dimensional Poisson process. Take $\lambda > 0$ and let N be a Poisson process on $\mathbb{R}_+ \times [0, \lambda]$ with expectation measure l, the Lebesgue measure. This means that for disjoint sets $A, A_1, \ldots, A_n \in \mathscr{B}(\mathbb{R}_+ \times [0, \lambda])$, the Borel σ-field,

1. the random variables $N(A_1)$,. . . ,$N(A_n)$ are independent, and
2. $N(A)$ is Poisson distributed with expected value $l(A)$.

The process $N_{(a,b]}$ defined by

$$N_{(a,b]}(s,t] = N((s,t] \times [a,b]), \quad 0 < a < b \leq \lambda, \quad 0 < s < t$$

is a one-dimensional Poisson process with intensity $(b - a)$. See Figure 3.2.

By dividing the region $\mathbb{R}_+ \times (0, \lambda]$ into horizontal strips we can construct several independent Poisson processes with constant intensities.

We want λ, the total intensity of transitions from state $\boldsymbol{i} \in E_{\boldsymbol{b}}$, to be constant for every $\boldsymbol{i}$. To achieve this we allow the chain $(X_t)_{t\geq 0}$ to have *quasi-transitions*, i.e., transitions from $\boldsymbol{i}$ to itself, so that the total

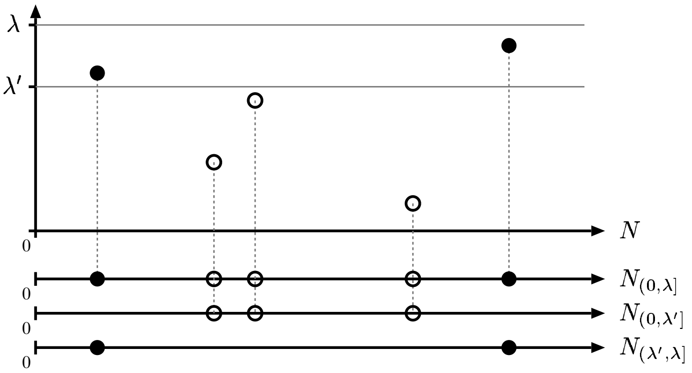

Figure 3.2. *The two-dimensional Poisson process N is used to construct two dependent, one-dimensional Poisson processes, $N_{(0,\lambda]}$ and $N_{(0,\lambda']}$ with intensities λ and λ' respectively. The processes $N_{(0,\lambda']}$ and $N_{(\lambda',\lambda]}$ are independent since they are derived from two disjoint sets of N.*

intensity out of $\boldsymbol{i}$ is fixed. Consider a bounded Jackson network of Type 1, $X = (X_t)_{t\geq 0}$. Examination of the transition intensities in (2.1) and (2.5) tells us that we can write the transitions out from $\boldsymbol{i}$ as follows:

$$\begin{array}{lll} \text{from} & \text{to} & \text{with intensity} \\ \boldsymbol{i} & \boldsymbol{i} + e_k \mathbf{1}_{\{i_k<b_k\}} & \beta_k \\ \boldsymbol{i} & \boldsymbol{i} - e_k \mathbf{1}_{\{i_k\geq n\}} & \nu_k(n)p_k \\ \boldsymbol{i} & \boldsymbol{i} - (e_k - e_m)\mathbf{1}_{\{i_k\geq n\}}\mathbf{1}_{\{i_m<b_m\}} & \nu_k(n)p_{km}, \end{array} \tag{3.3}$$

for $n = 1, \ldots, b_k$, $k, m \in \{1, \ldots, d\}$. To see this, consider for instance the intensity with which X_t goes from $\boldsymbol{i}$ to $\boldsymbol{i} - e_k$. This is

$$\sum_{n=1}^{b_k} \mathbf{1}_{\{i_k\geq n\}}\nu_k(n)p_k = p_k \sum_{n=1}^{i_k} \nu_k(n) = p_k\mu_k(i_k) = \delta_k(\boldsymbol{i}).$$

The other cases are similar. Equation (3.3) is the same thing as extending the transition intensities of (2.5) with the pseudo-transition intensities

$$\begin{aligned} c(\boldsymbol{i}) = \sum_{k=1}^{d} \Big(& \beta_k \mathbf{1}_{\{i_k=b_k\}} \\ & + \sum_{n=1}^{b_k} \Big(\mathbf{1}_{\{i_k<n\}}\nu_k(n)p_k + \sum_{m=1}^{d} \mathbf{1}_{\{i_k<n \cup i_m=b_m\}}\nu_k(n)p_{km} \Big)\Big) \end{aligned}$$

for all $\boldsymbol{i} \in E_{\boldsymbol{b}}$. Note that $c(\boldsymbol{i})$ is not an element of the intensity matrix Q, since the value of $c(\boldsymbol{i})$ does not affect the stationary distribution. Now,

$q(\boldsymbol{i}) = c(\boldsymbol{i}) + \sum_{\substack{\boldsymbol{j}\in E\\ \boldsymbol{j}\neq \boldsymbol{i}}} q_{\boldsymbol{ij}} = \lambda, \forall \boldsymbol{i} \in E_{\boldsymbol{b}}$, where $\lambda = \beta + \delta + \gamma$ which are defined as

$$
\begin{array}{lll}
\delta_k = \sum_{i=1}^{b_k} \nu_k(n) p_k & \gamma_{km} = \sum_{i=1}^{b_k} \nu_k(n) p_{km} & \gamma_k = \sum_{m=1}^{d} \gamma_{km} \\
\beta = \sum_{k=1}^{d} \beta_k & \delta = \sum_{k=1}^{d} \delta_k & \gamma = \sum_{k=1}^{d} \gamma_k .
\end{array}
$$

By dividing the region $\mathbb{R}_+ \times (0, \lambda]$ into horizontal strips of width β_k, δ_k, and γ_{km}, for $k, m \in \{1, \ldots, d\}$, we can construct independent Poisson processes N_{β_k}, N_{δ_k} and $N_{\gamma_{km}}$, $k, m \in \{1, \ldots, d\}$, with corresponding intensities. We can use these processes in the same manner as the processes $N_{\boldsymbol{ij}}$ earlier to govern the evolution of the Markov chain X. From (3.1) we see that

$$
\mathsf{P}(X_{n+1} = \boldsymbol{i} + e_k | X_n = \boldsymbol{i}) = \frac{q_{\boldsymbol{i}, \boldsymbol{i}+e_k}}{q(\boldsymbol{i})} = \frac{\beta_k}{\lambda}, \tag{3.4}
$$

if $i_k < b_k$, and

$$
\mathsf{P}(Z_{n+1} > t | X_n = \boldsymbol{i}) = \mathrm{e}^{-tq(\boldsymbol{i})} = \mathrm{e}^{-t\lambda}.
$$

Thus, we can construct the chain X by letting the time increments $(Z_n)_{n\geq 1}$ be i.i.d. exponentially distributed with expected value $1/\lambda$, i.e., letting $(T_n)_{n\geq 0}$ be a Poisson process with intensity λ, and by letting an i.i.d. sequence $(U_n)_{n\geq 1}$, uniform on $[0, \lambda]$, independent of $(T_n)_{n\geq 0}$, determine the jumps of $(X_n)_{n\geq 1}$. If $X_{n-1} = \boldsymbol{i}$ and U_n falls in the interval corresponding to β_k, say, then

$$
X_n = \boldsymbol{i} + e_k \mathbf{1}_{\{i_k < b_k\}},
$$

and analogously if U_n falls in an interval corresponding to some δ_k, γ_{km}, all in accordance to (3.3). Note that $(T_n, U_n)_{n\geq 1}$ is nothing but the enumeration of the points of N, where $T_k < T_{k+1}$ for $k \geq 1$.

With the philosophy that one should not try to explain with words what is best understood as a picture, we encourage the reader to consult Figure 3.3 for the construction of $(T_n, U_n)_{n\geq 1}$ and consider an illustration in Figure 3.4 of how this can be used to model a simple, bounded Jackson network.

For a bounded Jackson network of Type 2 the situation is completely analogous, but instead of (3.3) we have:

$$
\begin{array}{lll}
\text{from} & \text{to} & \text{with intensity} \\
\boldsymbol{i} & \boldsymbol{i} + e_k \mathbf{1}_{\{i_k < b_k\}} & \beta_k \\
\boldsymbol{i} & \boldsymbol{i} - e_k \mathbf{1}_{\{i_k \geq n\}} & \nu_k(n) p_k \\
\boldsymbol{i} & \boldsymbol{i} - (e_k - e_m \mathbf{1}_{\{i_m < b_m\}}) \mathbf{1}_{\{i_k \geq n\}} & \nu_k(n) p_{km},
\end{array} \tag{3.5}
$$

for $n = 1, \ldots, b_k$, $k, m \in \{1, \ldots, d\}$.

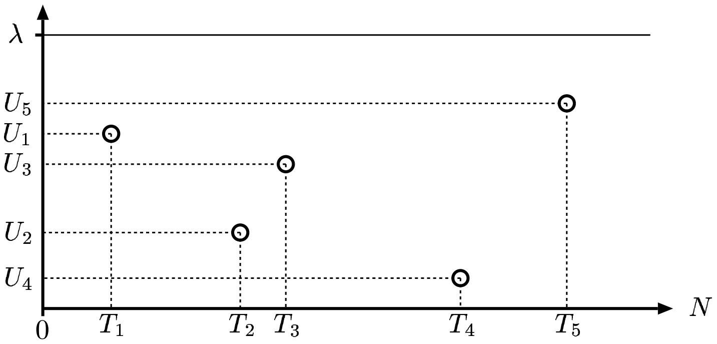

Figure 3.3. *The correspondence between the Poisson process N and the two independent i.i.d. sequences $(U_n)_{n\geq 1}$ and $(T_n)_{n\geq 1}$. Note the equivalence: N generates such sequences (T_n) and (U_n), and (T_n) and (U_n) can construct N.*

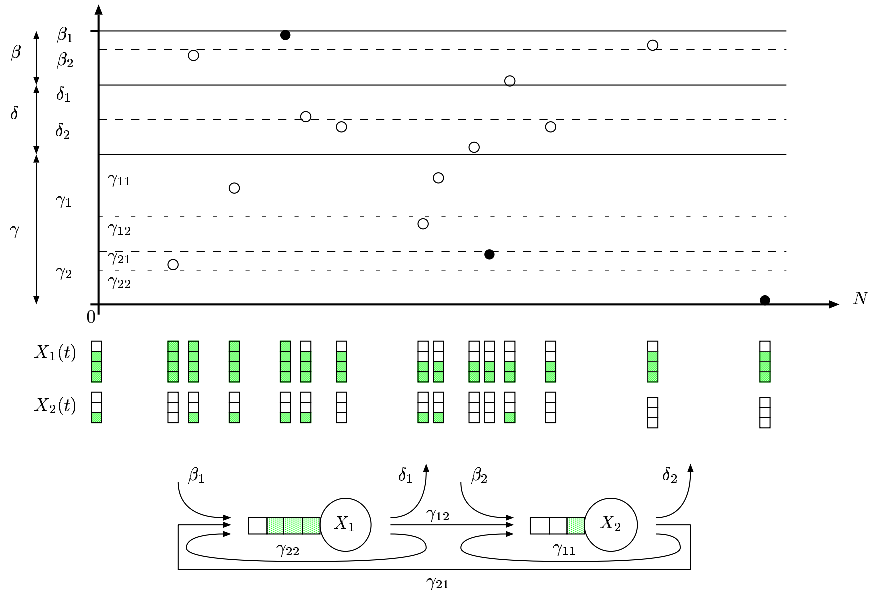

Figure 3.4. *How the two-dimensional Poisson process N governs the evolution of a bounded, 2-station Jackson network with $b = (4,3)$ when started in $X_0 = (3,1)$. Black dots symbolise events that trigger pseudo-transitions.*

Since we are going to use it in the following chapter, we establish the following notation.

Definition 3.1 For a bounded Jackson network X on $E_{\boldsymbol{b}}$ we define a *transition rule* as a deterministic function $\phi : E_{\boldsymbol{b}} \times [0, \lambda] \to E_{\boldsymbol{b}}$ that maps (X_n, U_{n+1}) on X_{n+1}, $n \geq 0$, where $(U_n)_{n \geq 1}$ is an i.i.d. sequence of random variables uniform on $[0, \lambda]$.

To summarise this section, we can construct the birth and death process $(X_t)_{t \geq 0}$ starting at $\boldsymbol{i}$ at time $T_0 = 0$, by generating the independent i.i.d. sequences $(Z_n)_{n \geq 1}$ and $(U_n)_{n \geq 1}$, where Z_n is exponentially distributed with expected value $1/\lambda$ and U_n is uniform on $[0, \lambda]$, $n \geq 1$. We let

$$X_t = X_n \quad \text{for} \quad T_n \leq t < T_{n+1} \tag{3.6}$$

and

$$\begin{aligned} T_{n+1} &= T_n + Z_{n+1} \\ X_{n+1} &= \phi(X_n, U_{n+1}). \end{aligned}$$

3.3 COUPLING OF TWO BIRTH AND DEATH PROCESSES

We will show how the results in the previous section can be used to produce the coupling in (2.10). Let $X = (X_t)_{t \geq 0}$ and $X' = (X'_t)_{t \geq 0}$ be two independent identical bounded Jackson networks on $E_{\boldsymbol{b}}$. We want to construct a coupling $(\widetilde{X}_t, \widetilde{X}'_t)_{t \geq 0}$ of X and X' such that when $\widetilde{X}_0 \preceq \widetilde{X}'_0$ then

$$\widetilde{X}_t \preceq \widetilde{X}'_t, \text{ for all } t \geq 0.$$

Let $(\widetilde{X}_t)_{t \geq 0}$ be constructed from $(\widetilde{X}_n)_{n \geq 0}$ according to (3.6), and the aim is to couple X and X' by a coupling $(\widetilde{X}_n, \widetilde{X}'_n)$ of $(X_n)_{n \geq 0}$ and $(X'_n)_{n \geq 0}$.

Suppose that $\widetilde{X}_0 = \boldsymbol{i} \preceq \boldsymbol{j} = \widetilde{X}'_0$. Then $\widetilde{X}_1 = \phi(\widetilde{X}_0, U)$ for some U uniform on $[0, \lambda]$, where ϕ is the transition rule of Definition 3.1 for this particular bounded Jackson network. Thus, the probability of a transition of $(\widetilde{X}_n)_{n \geq 0}$ from $\boldsymbol{i}$ to $\boldsymbol{i} + e_k \mathbf{1}_{\{i_k < b_k\}}$, say, is β_k / λ by (3.4). If that transition occurred, then $\phi(\boldsymbol{j}, U) = \boldsymbol{j} + e_k \mathbf{1}_{\{j_k < b_k\}}$ since ϕ is deterministic. The same is true for the other transitions, so by letting $\widetilde{X}'_{n+1} = \phi(\widetilde{X}'_n, U)$ we have produced the coupling:

from	to	with probability
$(\boldsymbol{i},\boldsymbol{j})$	$(\boldsymbol{i}+e_k\mathbf{1}_{\{i_k<b_k\}}, \boldsymbol{j}+e_k\mathbf{1}_{\{j_k<b_k\}})$	β_k/λ
$(\boldsymbol{i},\boldsymbol{j})$	$(\boldsymbol{i}-e_k\mathbf{1}_{\{i_k\ge n\}}, \boldsymbol{j}-e_k\mathbf{1}_{\{j_k\ge n\}})$	$\nu_k(n)p_k/\lambda$

and

$$(\boldsymbol{i},\boldsymbol{j}) \quad \begin{array}{l}(\boldsymbol{i}+(e_m-e_k)\mathbf{1}_{\{i_k\ge n\}}\mathbf{1}_{\{i_m<b_m\}},\\ \qquad \boldsymbol{j}+(e_m-e_k)\mathbf{1}_{\{j_k\ge n\}}\mathbf{1}_{\{j_m<b_m\}})\end{array} \quad \nu_k(n)p_{km}/\lambda, \tag{3.7}$$

if the network is of Type 1, and

$$(\boldsymbol{i},\boldsymbol{j}) \quad \begin{array}{l}(\boldsymbol{i}+(e_m\mathbf{1}_{\{i_m<b_m\}}-e_k)\mathbf{1}_{\{i_k\ge n\}},\\ \qquad \boldsymbol{j}+(e_m\mathbf{1}_{\{j_m<b_m\}}-e_k)\mathbf{1}_{\{j_k\ge n\}})\end{array} \quad \nu_k(n)p_{km}/\lambda,$$

if the network is of Type 2. It is clear from (3.3) or (3.5) that $\widetilde{X}_n \overset{\mathscr{D}}{=} X_n$ and $\widetilde{X}'_n \overset{\mathscr{D}}{=} X'_n$. Examination of the possible transitions reveals that

$$\widetilde{X}_n \preceq \widetilde{X}'_n \ \text{ for all } n > 0, \ \text{ when } \widetilde{X}_0 \preceq \widetilde{X}'_0.$$

Generate independent i.i.d. sequences $(Z_n)_{n\ge1}$ and $(U_n)_{n\ge1}$, as in (3.6), and let

$$\widetilde{X}_t = \widetilde{X}_n, \ \widetilde{X}'_t = \widetilde{X}'_n, \ \text{ for } T_n \le t < T_{n+1}, \text{ where } T_{n+1} = T_n + Z_{n+1},$$

and

$$\widetilde{X}_{n+1} = \phi(\widetilde{X}_n, U_{n+1}), \ \widetilde{X}'_{n+1} = \phi(\widetilde{X}'_n, U_{n+1}).$$

The probability of a transition of $(\widetilde{X}_n, \widetilde{X}'_n)$ from $(\boldsymbol{i},\boldsymbol{j})$ to $(\boldsymbol{i}+e_k, \boldsymbol{j}+e_k)$, say, is

$$\frac{\beta_k}{\lambda}\mathbf{1}_{\{j_k<b_k\}}\mathbf{1}_{\{i_k<b_k\}} = \frac{1}{\lambda}\beta_k\mathbf{1}_{\{j_k<b_k\}} = \frac{1}{\lambda}\beta(\boldsymbol{j}),$$

since $\boldsymbol{i} \preceq \boldsymbol{j}$. For a transition from $(\boldsymbol{i},\boldsymbol{j})$ to $(\boldsymbol{i}+e_k, \boldsymbol{j})$ the corresponding probability is

$$\frac{\beta_k}{\lambda}\mathbf{1}_{\{j_k=b_k\}}\mathbf{1}_{\{i_k<b_k\}} = \frac{\beta_k}{\lambda}(1-\mathbf{1}_{\{j_k<b_k\}})\mathbf{1}_{\{i_k<b_k\}} = \frac{1}{\lambda}(\beta(\boldsymbol{i})-\beta(\boldsymbol{j})).$$

From (3.2) we see that the transition intensity of $(\widetilde{X}_t)$ from $\boldsymbol{i}$ to $\boldsymbol{j} \ne \boldsymbol{i}$ is

$$q_{\boldsymbol{ij}} = q(\boldsymbol{i})\mathsf{P}(\phi(\boldsymbol{i}, U_{n+1}) = \boldsymbol{j}),$$

so that the transition intensity of $(\widetilde{X}_t, \widetilde{X}_t')$ from $(\boldsymbol{i}, \boldsymbol{j})$ to $(\boldsymbol{i}+e_k, \boldsymbol{j}+e_k)$, say, is

$$q((\boldsymbol{i},\boldsymbol{j}))\mathsf{P}((\phi(\boldsymbol{i},U_{n+1}),\phi(\boldsymbol{j},U_{n+1})) = (\boldsymbol{i}+e_k,\boldsymbol{j}+e_k)) = \lambda\frac{1}{\lambda}\beta(\boldsymbol{j}) = \beta(\boldsymbol{j}).$$

For a transition from $(\boldsymbol{i}, \boldsymbol{j})$ to $(\boldsymbol{i}+e_k, \boldsymbol{j})$ the corresponding intensity is

$$\lambda\frac{1}{\lambda}(\beta(\boldsymbol{i})-\beta(\boldsymbol{j})) = \beta(\boldsymbol{i})-\beta(\boldsymbol{j}).$$

Identifying these intensities with the ones given in (2.10) tells us that we have established our desired coupling.

To conclude, we can construct the coupling $(\widetilde{X}_t, \widetilde{X}_t')$ such that

$$\widetilde{X}_t \preceq \widetilde{X}_t', \quad t \geq 0, \text{ if } \widetilde{X}_0 \preceq \widetilde{X}_0,$$

This coupling corresponds to two bounded Jackson networks subject to the same arrivals of customers.

Chapter 4

Coupling from the past

4.1 FUNDAMENTALS

Let us retain the notation of our earlier chapters and construct the continuous time Markov chain, $(X_t)_{t\geq 0}$ from $(X_n)_{n\geq 0}$ and $(T_n)_{n\geq 0}$, according to (3.6); recall that for a sequence of i.i.d. variables $(U_n)_{n\geq 1}$ uniform on $[0, \lambda]$

$$X_{n+1} \;=\; \phi(X_n, U_{n+1}),$$

for some transition rule ϕ: $E_b \times [0, \lambda] \to E_b$. The objective is to find the stationary distribution of $(X_t)_{t\geq 0}$ when it is ergodic, which will always be the case for the queueing networks we consider. Basically, the coupling from the past algorithm relies on two intuitive ideas and a clever trick to make things work. First, consider a setting where several Markov chains all obey the same transition rule, ϕ, and suppose that at time T_0 they are simultaneously started from every possible state in E_b. Letting all the chains have the same sequence of transition times $(T_n)_{n\geq 1}$ while the chains evolve through time, we couple two chains whenever their trajectories intersect:

$$(X_{n+1}, X'_{n+1}) \;=\; \begin{cases} (\phi(X_n,\, U_n),\, \phi(X'_n,\, U'_n)) & \text{when } X_n \neq X'_n \\ (\phi(X_n,\, U_n),\, \phi(X_n,\, U_n)) & \text{when } X_n = X'_n. \end{cases}$$

This yields a coupling of our continuous time Markov chains $(X_t)_{t\geq 0}$, and the key point is that eventually all chains will have coalesced and all initial effects, i.e., dependence on initial state, have worned off. One would think that the common state of the chains at some time after the last coalescence is a sample from the stationary distribution, but, alas, this is not true.

Before that problem is resolved, we note that things could be simplified significantly by imposing some restrictions on $E_{\boldsymbol{b}}$ and ϕ. First we equip $E_{\boldsymbol{b}}$ with the partial ordering in (1.2), and recall that for $\mathbf{0}$ and $\boldsymbol{b}$,

$$\mathbf{0} \prec \boldsymbol{i} \prec \boldsymbol{b}, \quad \forall \boldsymbol{i} \in E_{\boldsymbol{b}} \setminus \{\mathbf{0},\, \boldsymbol{b}\}.$$

Now, with the coupling (3.7) the chain is monotone, i.e., for all $X_n \preceq X'_n$ we have

$$\widetilde{X}_{n+1} \preceq \widetilde{X}'_{n+1},$$

when we let $(X_n)_{n\geq 0}$ and $(X'_n)_{n\geq 0}$ have the same driving sequence $(U_n)_{n\geq 1}$. In that case, only two chains have to be run: $(X_n)_{n\geq 0}$ starting in $\mathbf{0}$, and $(X'_n)_{n\geq 0}$ starting in $\boldsymbol{b}$. When these two chains have coalesced, by necessity, so must all the others.

Here is the promised neat trick. Instead of running the chains forward in time to get a sample of the stationary distribution at some random time X_{T_n}, we start the chains $(X_t)_{t\geq s}$ at time points $s < 0$, to get a sample X_0 at the fixed time 0. Done properly, we will obtain a sample X_0 distributed according to π, from which the chain can continue.

In order to maintain a natural enumeration of events and time-points, we have to go through some trouble of notation. What we want to do is to run a Poisson process N on $\mathbb{R} \times [0,\, \lambda]$ and enumerate the points according to $(T_n, U_n)_{n=-\infty}^{\infty}$ glance at Figure 4.1.

Let N be a Poisson process on $\mathbb{R}_- \times [0, \lambda]$ and let $(T_n, U_n)_{n<0}, T_{n-1} < T_n$, be the corresponding enumeration of points. Let $(-s_{-n})_{n\geq 1}$ be an increasing sequence of real numbers such that $s_n \searrow -\infty$ as $n \to -\infty$, and put $s_0 = 0$. Let N_{s_n}, $n < 0$, denote the independent Poisson processes $N \cap [s_n,\, s_{n+1}) \times [0,\, \lambda]$ and M_{s_n} be the number of events of N_{s_n}. Finally, let $M_0 = 0$ and define $M_n = M_{s_n} + M_{n+1}$, the number of events of $N \cap [s_n, 0] \times [0, \lambda]$; see Figure 4.1.

Denote

$$\Phi_k^k(\boldsymbol{i}) = \boldsymbol{i}, \quad \Phi_n^k(\boldsymbol{i}) = \Phi_{n+1}^k(\phi(\boldsymbol{i}, U_n)), \quad n < k \leq 0, \forall \boldsymbol{i} \in E_{\boldsymbol{b}}.$$

Then for an initial state X_{s_n} at time s_n, $X_{s_k} = \Phi_{-M_n}^{-M_k}(X_{s_n}), n < k \leq 0$.

From the monotonicity of the chain, and since $\mathbf{0} \preceq \boldsymbol{i} \preceq \boldsymbol{b}, \forall \boldsymbol{i} \in E_{\boldsymbol{b}}$, we have by induction that

$$\Phi_{-n}^0(\mathbf{0}) \preceq \Phi_{-n}^0(\boldsymbol{i}) \preceq \Phi_{-n}^0(\boldsymbol{b}), \quad \forall \boldsymbol{i} \in E_{\boldsymbol{b}}, n > 0,$$

and if $\Phi_{-n}^0(\mathbf{0}) = \Phi_{-n}^0(\boldsymbol{b})$, we immediately see that $\Phi_{-n}^0(\boldsymbol{i})$ is the constant mapping for every $\boldsymbol{i} \in E_{\boldsymbol{b}}$. If we try larger and larger values of n until $\Phi_{-M_n}^0(\mathbf{0}) = \Phi_{-M_n}^0(\boldsymbol{b}) = X_0$, then the sample X_0 has exactly the distribution π.

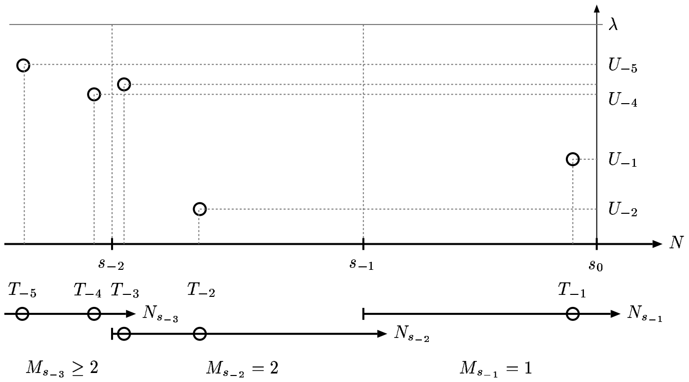

Figure 4.1. *The construction of* $(T_n, U_n)_{n<0}$. *In the interval* $[s_n, s_{n+1})$, *we have* M_{s_n} *events, and the total number of events from* s_n *to* s_0 *is* M_n.

Theorem 4.1 *The algorithm described above makes* $\Phi^0_{-M_n}$ *the constant mapping almost surely, as* $n \to \infty$, *i.e.,*

$$\exists c \in E_{\boldsymbol{b}}: \quad \mathsf{P}\left(\lim_{n\to\infty} \Phi^0_{-M_n}(\boldsymbol{i}) = c\right) = 1, \quad \forall \boldsymbol{i} \in E_{\boldsymbol{b}}.$$

Proof: This proof mimics the one found in Propp and Wilson's original paper [10] for discrete time Markov chains. First notice that for $n < 0$, $M_n = M_{s_n} + M_{n+1}$. Thus, M_{s_n} is a stopping time with respect to $(T_n)_{n<0}$, and since $(T_n)_{n<0}$ and $(U_n)_{n<0}$ are independent we conclude that M_n is independent of $(U_n)_{n<0}$. Further, if $\lambda > 0$ then $M_n \to \infty$ almost surely. The function $\Phi^0_{-n}(\,\cdot\,)$ depends only on $(U_k)^0_{k=-n}$, so $\Phi^0_{-M_n}$ is a constant mapping almost surely if Φ^0_{-n} is a constant mapping almost surely, as $n \to \infty$.
Since the chain is ergodic there exists a m such that for all states $\boldsymbol{i}, \boldsymbol{j}$ in $E_{\boldsymbol{b}}$, $\mathsf{P}(X_n = \boldsymbol{j} | X_{-(n+m)} = \boldsymbol{i}) > 0$. Thus, the probability of $\Phi^n_{-(n+m)}(\,\cdot\,)$ being a constant mapping is positive, say, $\epsilon > 0$. The chain run from time $-n \cdot m$ to 0 gives the mapping

$$\Phi^0_{-nm}(\,\cdot\,) \;=\; \Phi^0_{-m}(\Phi^{-m}_{-nm}(\,\cdot\,)) \;=\; \Phi^0_{-m} \circ \cdots \circ \Phi^{-(n-1)m}_{-nm}(\,\cdot\,).$$

The events

$$\left\{\Phi^{-(k-1)m}_{-km}(\,\cdot\,) = \text{constant}\right\}, \quad k = 1, \ldots, n,$$

are independent, and each of them has a positive probability of occurring, so the probability that at least one of them is constant is 1 by Borel-Cantelli's second lemma. But if one of them is constant, at time $-k$, say, then

$$\Phi^0_{-n}(\,\cdot\,) \;=\; \Phi^0_{-k}(\,\cdot\,) \;=\; \text{constant}, \quad \text{for all } n > k.$$

□

Theorem 4.2 *If the algorithm terminates with probability* 1*, then the output state* X_0 *is distributed according to* π*.*

Proof: This proof is a continuous-time version of the proof found in [10]. Let $(U_n)_{n<0}$ and $(T_n)_{n<0}$ be independent sequences of i.i.d. random variables constructed as the enumeration of points of a Poisson process on $\mathbb{R}_- \times [0, \lambda]$. Let $Y \overset{\mathscr{D}}{=} \pi$ and define

$$\begin{aligned} X_{-n} &= \Phi^0_{-n}(Y) = \phi(\Phi^{-1}_{-n}(Y)), \quad n = 1, 2, \dots \\ X_t &= X_{-n} \text{ if } T_{-n} < t < T_{-n+1}, \quad n = 1, 2, \dots \end{aligned}$$

By construction we have $X_t \sim \pi$ for every $t < 0$. Now, since $\Phi^0_{-n}(\,\cdot\,)$ is with probability one a constant mapping for large n, X_{-n} converges almost surely to this constant, say, X_0. The sequence $(-T_n)_{n<0}$ is almost surely increasing, so $X_{-t} \to X_0$ [a.s.] as $t \to \infty$, where X_0 by construction has distribution π. □

4.2 SIZE OF TIME-STEPS

Let k be the smallest value of n such that the two chains (X_t) and (X'_t) have coalesced when started from $(X_{s_{-n}}, X'_{s_{-n}}) = (\mathbf{0}, \boldsymbol{b})$. A question that remains to be settled is how to choose the sequence of starting times $(s_n)_{n<0}$ so that k is as small as possible, and $-s_{-k}$ is not too large. It turns out that letting s_{-n} follow a geometric sequence is a good strategy, and we will in the following use

$$s_{-n} = -(2^n - 1), \quad n \geq 0.$$

This gives an upper bound on the total number of transitions of (X_t, X'_t), $M_{-1} + \cdots + M_{-k}$, in the sense that the expected number of transitions is never larger than four times the number of transitions used if the sequence $(T_n, U_n)_{n<0}$ was known beforehand.

A schematic view of the algorithm is shown in Figure 4.2 and a realisation of a simulation of a 5-station, bounded Jackson network illustrated in Figure 4.4 is shown in Figure 4.3. The sequence of starting points used is $s_{-n} = -(2^n - 1)$.

4.3 OTHER ALGORITHMS

The number of iterations for the algorithm before it terminates depends on the observation X_0. It is important to note that an impatient user who interrupts the algorithm and only collects observations with short running times, i.e., a low number of iterations, introduces bias. Interruptible

```
$k \leftarrow 0$
repeat
    $k \leftarrow k + 1$
    $T \leftarrow s_{-k}$
    $n \leftarrow M_{s_{-k}}$
    $major \leftarrow \boldsymbol{b}$
    $minor \leftarrow \mathbf{0}$
    while $T < 0$ do
        $major \leftarrow \phi(major, U_{-n})$
        $minor \leftarrow \phi(minor, U_{-n})$
        $n \leftarrow n - 1$
        $T \leftarrow T + Z_{-n}$
    end
until $major = minor$
return $major$
```

Figure 4.2. *Algorithm for CFTP sampling.*

algorithms for perfect simulation have been developed; see for instance Fill[6].

Wilson[15] has formulated an enhancement of the coupling from the past algorithm which removes the necessity of storing the random numbers $(U_n)_{-s_k}^{0}$ (or the random seeds generating them) for later reuse. This modified algorithm can be used with a single stream of for instance hardware generated random numbers, and is thus called the *read once coupling from the past algorithm.*

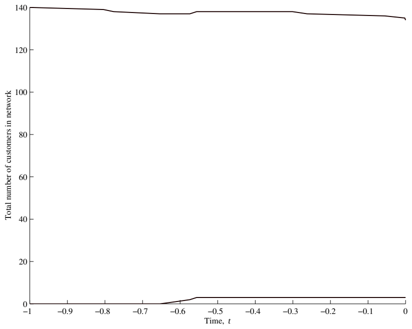

Figure 4.3.a.

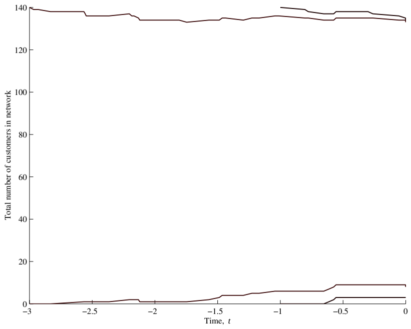

Figure 4.3.b.

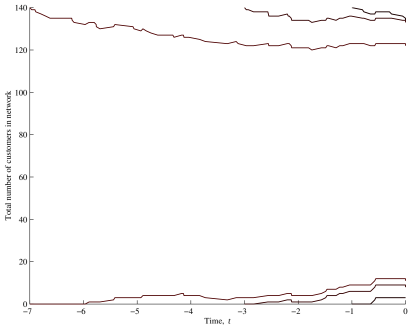

Figure 4.3.c.

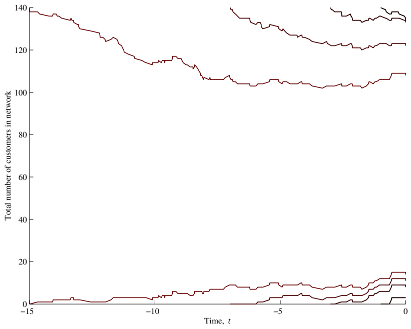

Figure 4.3.d.

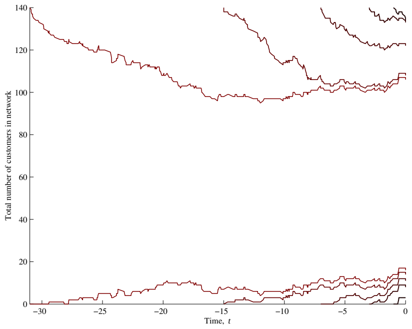

Figure 4.3.e.

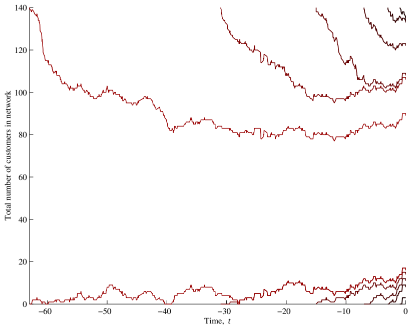

Figure 4.3.f.

Figures 4.3.a–j show realisations of a CFTP simulation of the 5-station, bounded Jackson network in Figure 4.4, for starting points $s_{-n} = -(2^n - 1)$, $n \geq 1$. *The trajectories show the total number of customers in the system for* (X_t) *and* (X'_t), *where the latter start in* $\boldsymbol{b}$, *which equals* $100 + 4 \times 10 = 140$ *customers in the system. The simulation when started at* s_{-10} *made the two chains* $(X_t)_{t \geq s_{-10}}$ *and* $(X'_t)_{t \geq s_{-10}}$, *started in* $\mathbf{0}$ *and* $\boldsymbol{b}$ *respectively, coalesce by time 0, and the output state* $X_0 = X'_0 = (6\,0\,4\,4\,2)$ *is a sample from the stationary distribution* π.

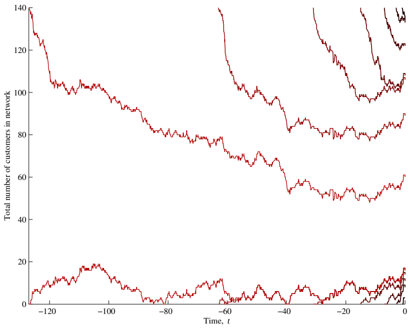

Figure 4.3.g.

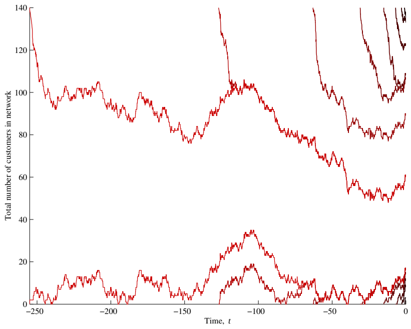

Figure 4.3.h.

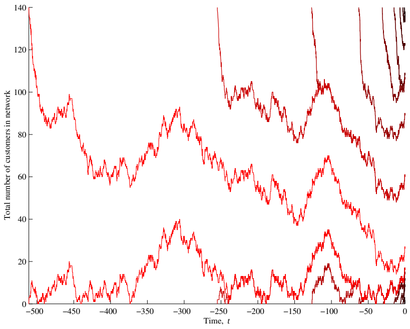

Figure 4.3.i.

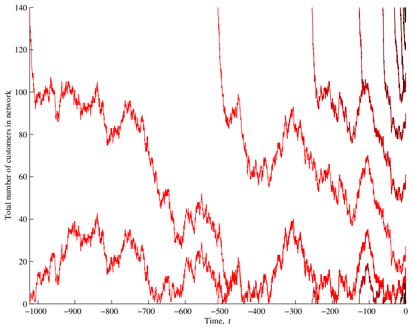

Figure 4.3.j.

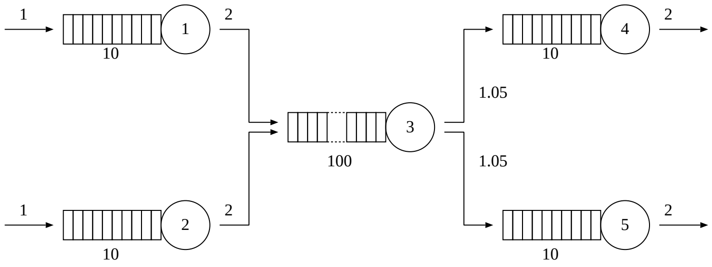

Figure 4.4. *An example of a bounded Jackson network with 5 stations, where* $\beta = (1, 1, 0, 0, 0)$, $\mu = (2, 2, 2.1, 2, 2)$, *and the only nonzero routing probabilities are* $p_{13} = p_{23} = 1$ *and* $p_{34} = p_{35} = 0.5$.

Chapter 5

Simulations

This chapter is devoted to the study of the CFTP algorithm when applied to bounded Jackson networks. In the first section we examine the performance of the sampling algorithm in order to determine how large a network can be simulated in a reasonable amount of time. Networks of different size, but of certain structures, are simulated and the number of iterations is plotted for each network. The objective of the following section is to simulate various Jackson networks with different arrangements and parameter settings to show the richness of the model and characteristics of the bounded Jackson networks. The networks are not to be considered as real models of computer networks, but more as an indication of what can be achieved by simulation. The section consists of several subsections, where the first treats a central server system with the objective of finding how well the internal flow of customers in the network can be approximated by Poisson processes. The next subsection concerns bottleneck behaviour for a 5-station Jackson network, and the last one concerns simulation and evolution of the general birth and death process $(X_t)_{t \geq 0}$ when it is stationary at time 0.

5.1 PERFORMANCE

The performance of the algorithm is measured in terms of simulation time for various queueing networks with a certain structure. To find a way of systematically testing different kinds of networks, let a d-station, bounded Jackson network be as follows:

- We have equally sized buffers, i.e., $b_1 = b_2 = \cdots = b_d$ in the bounds vector $\boldsymbol{b} = (b_1, \ldots, b_d)$.

- ○ Every node is a single-server queue and has the same arrival intensity of customers, and every customer has the same service rate, i.e., $\beta_1 = \cdots = \beta_d$ and $\mu_1 = \cdots \mu_d$.
- ○ The network corresponds to a complete graph with uniform routing, i.e., $p_{ij} = 1/(d+1)$, for all $i, j \in \{1, \ldots, d\}$. This means that a customer, upon leaving a queue, has a uniform probability of going to any other queue or leaving the network entirely.

As we increase the number of stations d in the network, we measure the simulation time of the algorithm before it terminates and outputs an observation from the stationary distribution. The actual simulation time is not a good measure since it is too dependent on computer architecture and changes in work load during the simulation. Instead, as a first performance measure we take the number of iterations, or transitions of the Markov chain $(X_t, X_t')^0_{t \geq s_{-k}}$, where k is the smallest value of n such that $(X_t, X_t')^0_{t \geq s_{-n}}$ has coalesced by time 0. As d increases, the total intensity of transitions in the network, λ, increases, so the number of iterations should be of magnitude $\lambda |s_{-k}|$. The particular choice of s_{-n} we shall use is

$$s_{-n} = -(2^n - 1),$$

and if k does not depend on d then we should see a linear growth in the number of iterations as d increases. In Figure 5.1 we see the total number of iterations of the algorithm for d-station, bounded Jackson networks, where d ranges from 2 to 200. The intensities and buffer sizes are given by

$$\beta_k = 2,\ \mu_k = 2,\ p_{km} = \frac{1}{d+1}, \text{ and } b_k = 100 \text{ for all } k, m \in \{1, \ldots, d\}.$$

We also plot the mean value of k as a function of d in Figure 5.2.

From the figures we find rather surprisingly that the number of successive starting-points s_{-n} stabilises to a fixed value, regardless of the number of stations in the queueing network. As a consequence, the number of transitions grows as λ which is linear in the number of stations.

To get an understanding of how the number of iterations before termination of the algorithm depends on the connectivity of the network, we adjust the routing probabilities as follows:

- ○ Fix a probability p.
- ○ For each station $k \in \{1, \ldots, d\}$ take d independent Bernoulli variables

$$I_{k1}, \ldots, I_{kd} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

- ○ Let $p_{ij} = I_{ij} / \sum_{k=1}^{d} I_{ik}$.

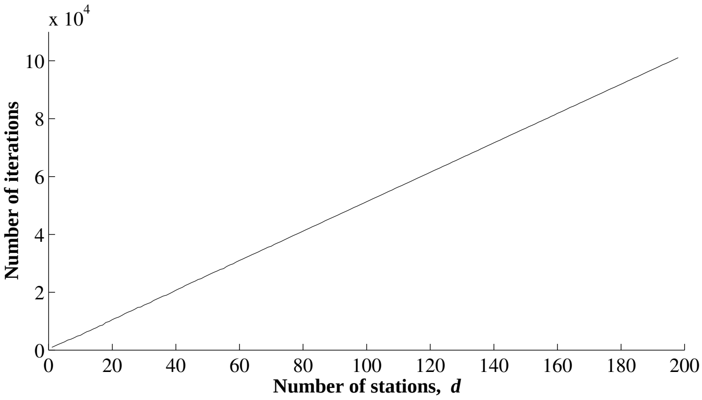

Figure 5.1. *For a d-station network, we measure the mean number of transitions of $(X_t, X'_t)_{t \geq s_{-k}}$, where k is the smallest number of n such that the chains (X_t) and (X'_t) have coalesced when started from s_{-n}. The mean value is taken for a sample of 100 observations.*

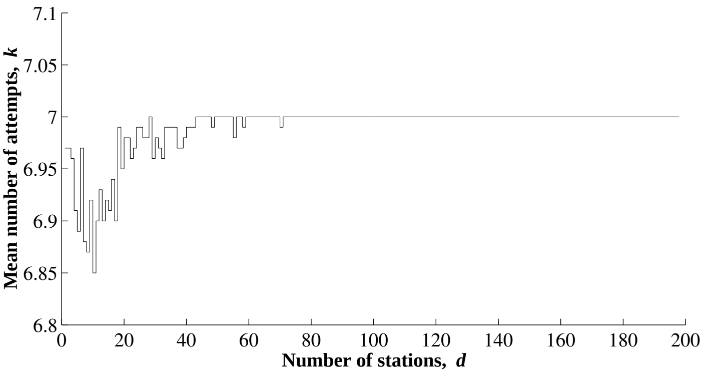

Figure 5.2. *The same simulation as in Figure 5.1, but we plot the mean value of k for the 100 observations taken. It seems that for $d > 100$, the number of attempts, k, has stabilised and the chains (X_t) and (X'_t) coalesce when started from $s_{-7} = -127$.*

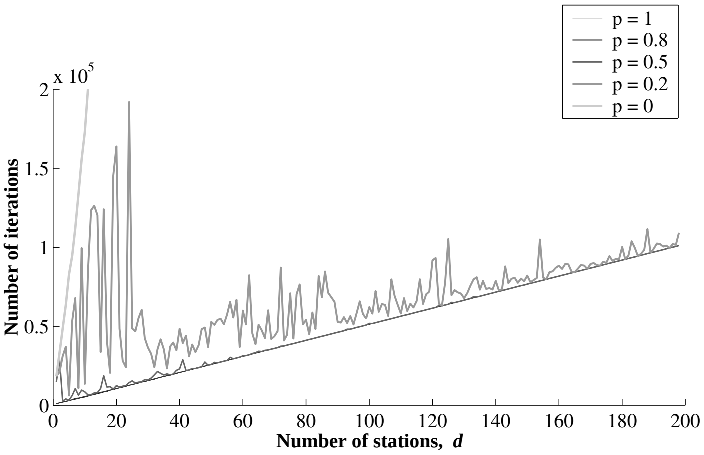

Figure 5.3. *For a d-station network, we measure the mean number of transitions of* $(X_t, X'_t)_{t\geq s_{-k}}$*, and again the mean is taken for a sample of 100 observations. The simulation corresponding to* $p = 0$ *is only run for values of d up to 168, and most of them fall outside the frame of the picture.*

We examine the total number of iterations of the algorithm for some values of p and d, and compare the result with the one obtained earlier for the same values of d.

Figure 5.3 shows the total number of iterations and Figure 5.4 the mean value of k. Apart from when $p = 0$, all simulations tend to need the same number of time-steps to make the two processes X_t and X'_t coalesce, for large d. When $p = 0$ the network consists of d queues with no connections between them, and since $\beta_k = 2$ and $\mu_k = 2$, we have $\rho_k = 1$ for every queue.

Suppose that we could calculate the traffic intensity, ρ_k, at every node in the network. When the net input rate λ_k into a station k is greater than the service rate μ_k we have a drift towards a large number of customers in the queue, and intuitively, the coalescence of the minimal and the maximal processes should occur early, which yields a small value of n and a low number of iterations before termination of the algorithm. The same should be true when the input rate is smaller than the service rate, and the worst-case scenario would be when $\lambda_k = \mu_k$ so that $\rho_k = 1$. For a standard Jackson network these traffic intensities could be calculated by first solving the balance equations (2.7) and using the solution λ_k in $\rho_k = \lambda_k/\mu_k$. A bounded Jackson network should have a smaller traffic intensity at each node, since the arrival intensity at each node is at most β_k and we have departures from the network due to failed transfers between queues. For

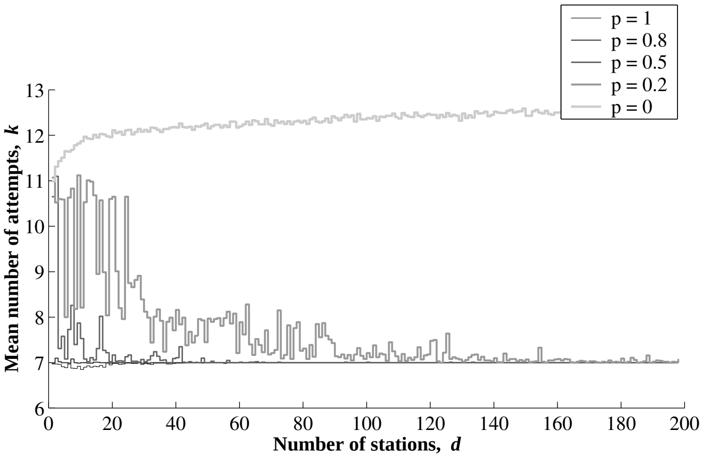

Figure 5.4. *The same simulations as in Figure 5.3, and we see that the number of attempts made, k, for simulations corresponding to $p = 0.8$, $p = 0.5$ and $p = 0.2$, settles at the same value as in the case $p = 1$. When $p = 0$ the network consists of d parallel M/M/1-queues and the parameter settings are such that $\rho_k = 1$ for $k = 1, \ldots, d$.*

a standard, d-station Jackson network corresponding to the bounded ones above, the traffic intensity is

$$\rho_k = \frac{\beta(d+1)}{\mu(d(1-p)+1)},$$

and for symmetry reasons, the traffic intensity is the same at all the nodes in the network. We note that the traffic intensity is increasing in p, and since $p = 0$ implied $\rho = 1$, the heuristic argument above tells us that we should expect fewer attempts for networks with large values of p.

We now adjust the arrival intensity to the system, let $\beta_k = 1$, $k = 1, \ldots, d$, and perform the same simulation as given above. We examine the total number of iterations of the algorithm for some values of p and d, and compare the result with the one obtained earlier for the same values of d.

Figures 5.5–5.7 show the mean number of iterations, the mean number of attempts and the traffic intensity for a corresponding standard Jackson network. Here $p = 0$ yields a queueing network consisting of parallel queues with traffic intensity $\rho_k = 0.5$, and for $p = 0$, $p = 0.5$, $p = 0.8$ and $p = 1$, the number of attempts seems to settle at a common level as d increases. The asymptote of case $p = 0.2$ is unknown. The value $p = 0.5$ makes the traffic intensity approach 1 as d increases, and yet the case $p = 0.2$ required the largest number of attempts, so the traffic intensity cannot by itself be

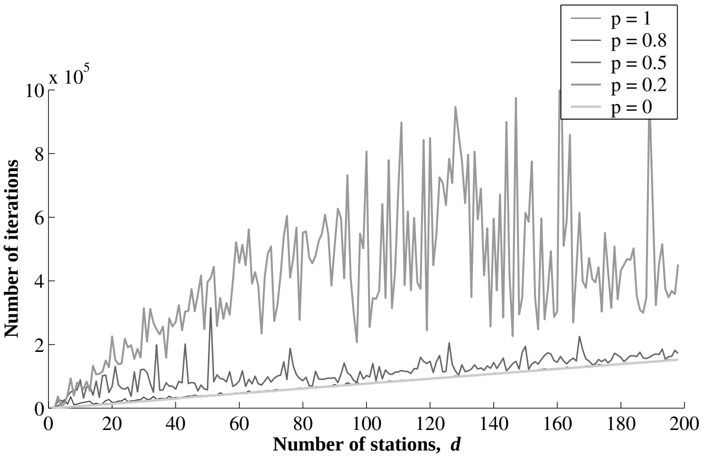

Figure 5.5. *This plot shows the mean number of transitions of* $(X_t, X'_t)_{t \geq s_{-k}}$*, for a sample size of 100. The simulation corresponding to* $p = 0$ *makes the queueing network consist of d parallel queues with traffic intensity* 0.5.

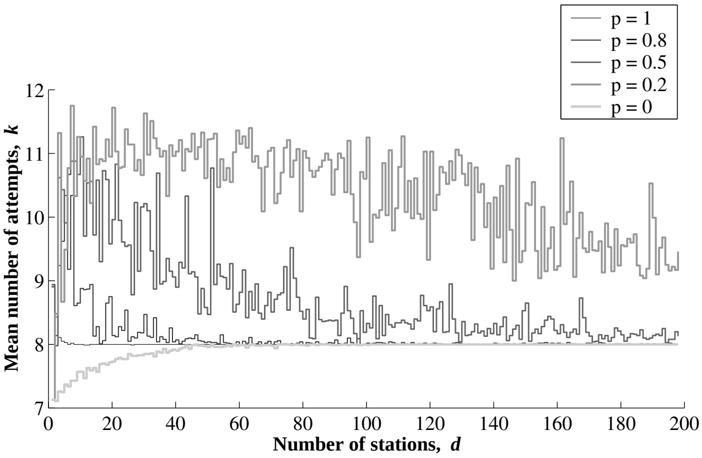

Figure 5.6. *The same simulations as in Figure 5.5, and we see that the number of attempts made, k, for simulations corresponding to* $p = 1$, $p = 0.8$, $p = 0.5$, *and* $p = 0$, *settles at a common value, which is greater than the value when* $\beta = 2$.

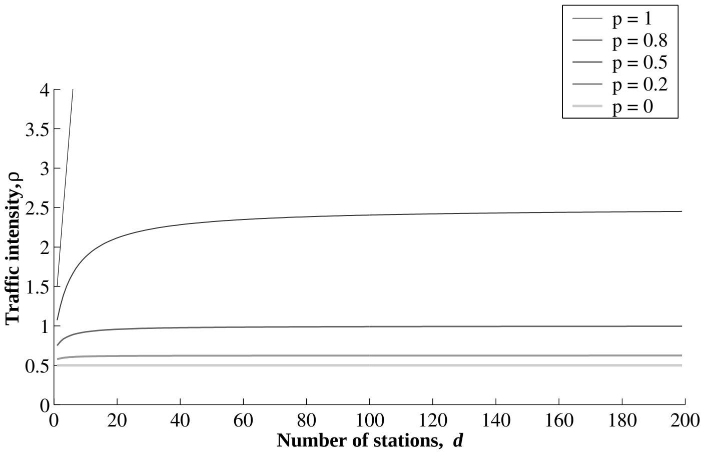

Figure 5.7. *The traffic intensity for the standard Jackson network when $\beta = 1$, $\mu = 2$ and each node is connected with $d \cdot p$ others, on the average. When $p = 0.5$ the traffic intensity tends to 1 as d increases.*

used to explain the number of attempts before termination. We also note that apart from the case $p = 0$, the number of iterations (and attempts) was smaller when $\beta = 2$.

We conclude, however, that the number of time-points, $s_{-n} = -(2^n - 1)$, attempted before the chains $(X_t)_{t \geq s_{-n}}$ and $(X_t')_{t \geq s_{-n}}$ coalesced before time 0, seems to settle down as d increases, and thus, the time before termination of the CFTP algorithm, as measured in the number of transitions, grows only linearly for large d. The conclusion is that it is feasible to use the CFTP algorithm to obtain samples from the stationary distribution even for large, bounded Jackson networks.

5.2 EXAMPLE OF NETWORKS

5.2.a CENTRAL SERVER SYSTEMS

A special family of bounded Jackson networks with a certain structure are those we call *central server systems*, where we imagine $d - 1$ parallel queues, which are connected to a central server of high capacity. This is assumed to be a model of a computer network where each queue corresponds to a physical machine and the customers are jobs or processes at that computer. Some jobs at a computer are forwarded to the central server; see Figure 5.8.

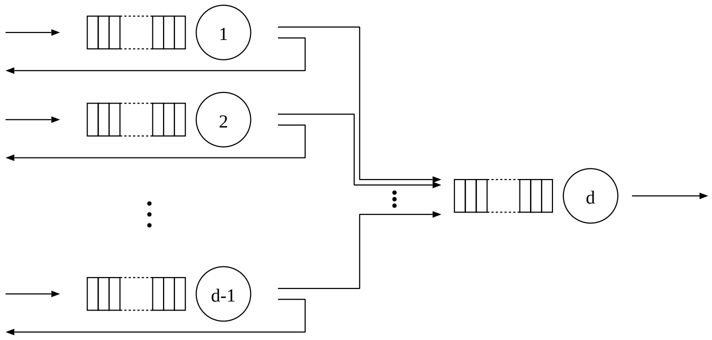

Figure 5.8. *This network consists of $d-1$ computers and a central server which is supposed to take care of special requests the computers cannot handle.*

Let the network be of Type 2 and β_k the arrival intensity at computer k, $k = 1, \dots, d-1$. The stationary distribution of the number of customers in the queue at computer k is obtained from (2.9) with $\rho_k = \beta_k/\mu_k$, and μ_k is the service rate at computer k. When the buffer size, b_k, at computer k, $k = 1, \dots, d-1$, is large compared to the expected number of customers, the departure process from the computer should be Poisson-like with intensity $\lambda_k = \beta_k$ when the network is stationary. Small buffers as compared to the intensities of arrivals would reject jobs from entering the network and thus yield a smaller value of λ_k, and also a less Poisson-like departure process. The net flow of jobs to the central server is a superposition of the departure processes from the computers, and we shall investigate how the distribution of customers at the central server queue depends on the buffer sizes $b_1 \dots, b_{d-1}$ and d, and compare the distribution with the distribution when the arrival process at the central server is a Poisson process with intensity $\sum_{k=1}^{d-1} \beta_k p_{kd}$.

If the departure process from station k is Poisson, then the routing probabilities p_{kd} correspond to Bernoulli sampling of the process, which makes the arrival process from queue k to the central server still a Poisson process with intensity $\beta_k p_{kd}$. Thus it suffices to investigate the case $p_{kd} = 1$ for $k = 1, \dots, d-1$.

To start slowly, let $d = 3$ and

$$\beta_k = 1 \quad \mu_k = 2 \quad p_{kd} = 1 \quad \text{and} \quad b_k = 30, \quad \text{for} \quad k = 1, \dots, d-1. \tag{5.1}$$

Let the server be given by the parameters

$$\beta_d = 0 \quad \mu_d = 2.1 \quad p_d = 1 \quad \text{and} \quad b_d = 100. \tag{5.2}$$

Let $(X_t)_{t \geq 0}$ denote the bounded Jackson network of Type 2 and $X^{(n)}(0)$ the n^{th} sample, obtained with the CFTP algorithm, from the stationary

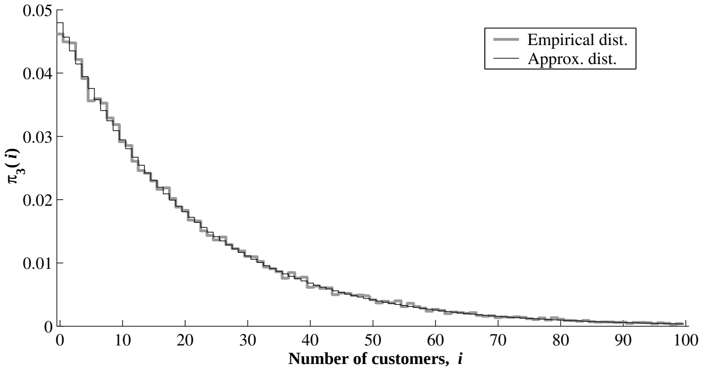

Figure 5.9. *The correspondence between the estimated stationary distribution, obtained from 50,000 perfect samples, and the distribution when the arrival process is Poisson at intensity 2.*

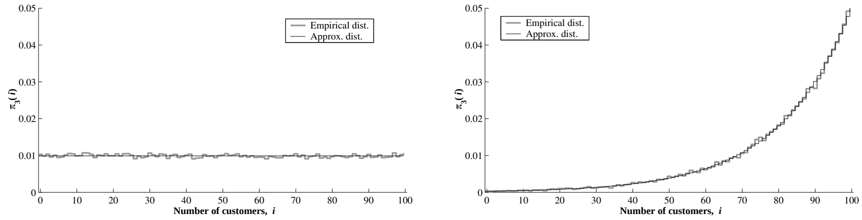

Figure 5.10. *This figure is the same as Figure 5.9, but where the service rate at the central server is* $\mu_3 = 2.0$ *(left) and* $\mu_3 = 1.9$ *(right). The goodness of fit measure is* $S = 89.4$ *and* $S = 101$ *respectively.*

distribution, π. We estimate π as

$$\widehat{\pi}(\boldsymbol{i}) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X^{(k)}(0)=\boldsymbol{i}\}},$$

and we denote the estimated marginal probabilities $\pi_m(i)$ by $\widehat{\pi}_m(i)$; we have

$$\widehat{\pi}_m(i) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_m^{(k)}(0)=i\}}, \quad m = 1, 2, 3.$$

Figure 5.9 shows $\widehat{\pi}_3$, the estimated, stationary distribution of customers at the central server, based on $n = 50,000$ observations, as compared with the distribution when the arrival process is Poisson with intensity $\beta_1 + \beta_2 = 2$.

The correspondence is good, as should be expected since equation (2.9) applied to queue k, $k = 1, 2$, yields $\pi_k(b_k) = 4.7 \cdot 10^{-10}$. Now, $\pi_k(b_k)$

is the fraction of time queue k is full, and since the arrival process is Poisson, it is also, by the PASTA-property, the fraction of arrivals who find queue k full. Thus, almost every arrival enters the system and under stationarity the departure process *from* station k is almost Poisson with intensity $\beta_k(1-\pi_k(b_k)) \approx \beta_k$. The superposition of the departure processes from stations 1 and 2 is thus close to a Poisson process with intensity $\beta_1 + \beta_2 = 2$. The χ^2-statistics of goodness of fit between the estimated distribution and the approximate distribution is

$$S = n \sum_{k=0}^{b_k} \frac{(\widehat{\pi}_d(k) - p_k)^2}{p_k} = 88.54,$$

where p_k is the probability of (2.9) with $\lambda_d = \beta_1 + \beta_2 = 2$. Since $S \overset{\mathscr{D}}{\approx} \chi^2$ with 100 degrees of freedom, the observed value of S corresponds to a p-value of 0.7870. The behaviour of the queue at the central server is thus essentially the same as if the arrival process were a Poisson process.

Now, let the parameters be as in (5.1) and (5.2), but with $b_1 = b_2 = 3$. Then $\pi_k(b_k) = 1/15 = 6.7\%$, and the fraction of arrivals who cannot enter the system is not negligible. The departure process from one of the first $d-1$ stations under stationarity is thus a stuttering process with an intensity approximately $\beta_k(1-\pi_k(b_k))$. The superposition of many such departure processes should form a Poisson-like process; Figure 5.11 shows the distribution of customers at the central server for different values of d. The parameters for the first $d-1$ stations are as in (5.1), but the central server has parameters

$$\beta_d = 0 \quad \mu_d = 2.1(d-1)/2 \quad p_d = 1 \quad \text{and} \quad b_d = 100.$$

The χ^2-statistics of goodness of fit, S, for comparison between the estimated distributions and (2.9) when the arrival process is Poisson at intensity $\lambda_d = \sum_{k=1}^{d-1} \beta_k(1-\pi_k(b_k))$ are shown in the table below:

d	μ_d	λ_d	S
3	2.1	28/15	654
20	19.95	266/15	361
50	51.45	686/15	217

The p-values are absurdly small, so the fit between the distributions is bad. One may argue that the discrepancy is only due to the huge number of observations made, and that the distributions are essentially the same when examining the graphs in Figure 5.11. But nevertheless, the arrival process is not a Poisson process. We note, though, that the more computers connected to the server the more similar to a Poisson process is the superposition of departure processes, and that the correspondence is worse when the central

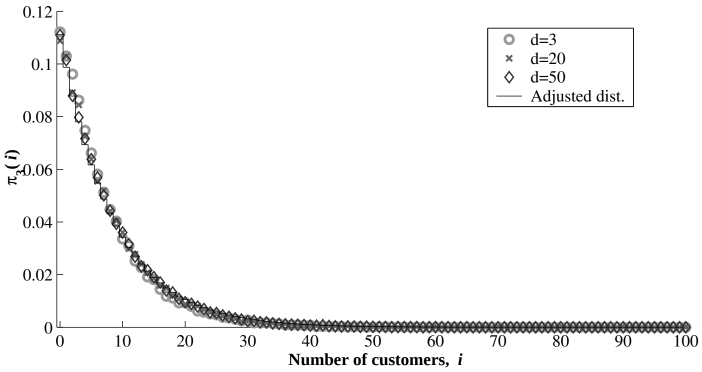

Figure 5.11. *The correspondence between $\widehat{\pi}_3$ for various numbers of computers connected to the server, together with the distribution when the arrival process is Poisson with intensity $\lambda_3 = \sum_{k=1}^{d-1} \beta_k(1 - \pi_k(b_k))$ (labelled 'Adjusted distribution'). The sample size used was $n = 50,000$.*

server is overloaded. The last feature is expected, since the arrival process has long inter-arrival times when the central server is congested.

5.2.b BOTTLENECK BEHAVIOUR

Recall the right-hand plot of Figure 5.10, where the service rate at the central server was smaller than the average arrival intensity. Had the network been a bounded Jackson network of Type 1, then a congestion of customers would have occurred at the computers connected to the server, yielding an even lesser inflow of customers to the system. This short subsection is an illustration of how bottlenecks in a network affect the overall performance. For Type 1 networks the impact could be severe.

Consider the 5-station, bounded Jackson network summarised by the parameters

$$(\beta_1, \ldots, \beta_5) = (1, 1, 0, 0, 0) \quad (\mu_1, \ldots, \mu_5) = (2, 2, 2, 2, 2)$$

and routing probabilities given by the matrix

$$[p_{km}] = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0. \end{bmatrix}$$

Station 3 is a bottleneck and for a vector of bounds $\boldsymbol{b} = (10, 10, 10, 10, 10)$, say, we compute an estimate of the stationary distribution. A Type 2 network will dispose of those departures from stations 1 and 2 that cannot be transferred to station 3, but in a Type 1 network the first two stations will be clogged by those customers.

We concentrate on two scenarios. In the first one, every node in the network has an average offered load smaller or equal to the service capacity, and we compare the difference between networks of Type 1 and of Type 2. This comparison is also made in a slightly overloaded situation, where $\mu_3 = 1.8$. In the second scenario we re-examine the first scenario but change the service rate μ_3 to

$$\mu_3(i) = \sum_{n=1}^{i} \nu_3(n), \text{ where } \nu_3(1) = \nu_3(2) = 1, \nu_3(n) = 0, \, n \geq 3$$

i.e., we have two servers at the bottleneck station, such that the maximum service capacity at the station is the same as for the single server in the first scenario. In the overloaded setting we have $\nu_3(1) = \nu_3(2) = 0.9$.

In Figure 5.10 we saw what distribution of customers we can expect at the bottleneck for a network of Type 2. It turns out that for a Type 1 network the bottleneck node is more congested. The reason for this is that no customers are released from the first two stations until there is a vacant position at the bottleneck. Thus, if the bottleneck node becomes full, customers will clog the primary stations and as soon as a departure occurs at the bottleneck one of those customers will fill the vacant position. The result is that the bottleneck will have long runs where it is saturated and it cannot settle down until the primary stations by chance become empty or receive customers with exceptionally long service times. Figure 5.12 shows the estimated stationary distribution of customers at station 3 obtained from 50,000 samples, and Figure 5.13 contains a trace of queue lengths for the bounded Jackson network of Type 1.

The network of Type 1 contains more customers on the average than the network of Type 2 as Theorem 2.5 stated. The network of Type 2 can manage the offered load whereas the network of Type 1 has a drift towards a large number of customers. When the service rate decreases to $\mu_3 = 1.8$, the overloaded situation, things go from bad to worse.

Simulating the second scenario where we have a double server at station 3, we obtain Figure 5.14, and we note that when the number of servers changes the networks tend to be even more congested by customers.

5.2.c ANALYSIS OF TRACES

This short section will illustrate the type of analysis that can be made by analysing traces of the Markov chain $(X_t)_{t \geq 0}$ describing a bounded Jackson

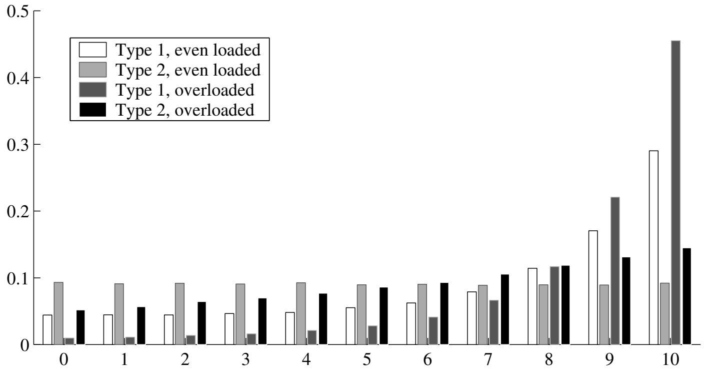

Figure 5.12. *Estimated distributions of the stationary numbers of customers at the bottleneck station in a 5-station bounded Jackson network.*

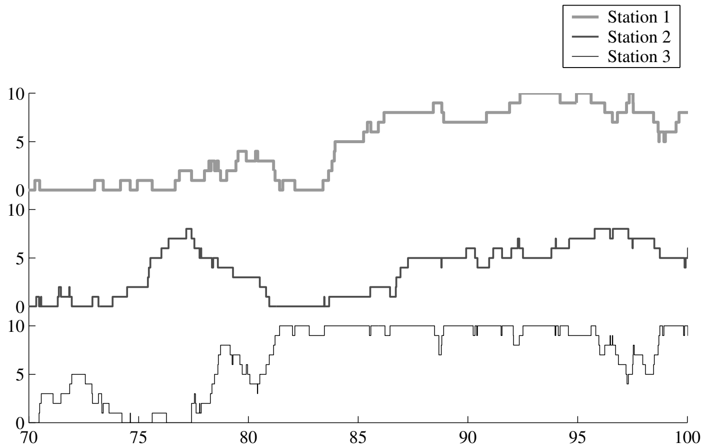

Figure 5.13. *A trace of queue lengths $(X_1(t))$, $(X_2(t))$ and $(X_3(t))$ for a bounded Jackson network (X_t) of Type 1 when $X_0 \overset{\mathcal{D}}{=} \pi$.*

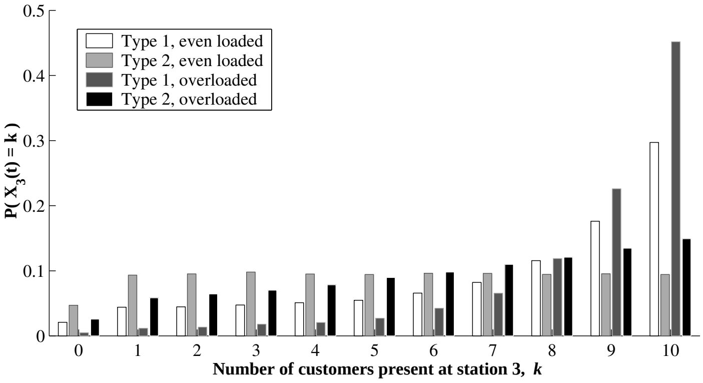

Figure 5.14. *Estimated distributions of the stationary numbers of customers at the bottleneck station in a 5-station bounded Jackson network with a double server at station 3.*

network, where $X_0 \overset{\mathscr{D}}{=} \pi$, a sample obtained by the CFTP algorithm. Consider the network of Type 2 treated in the previous subsection, where the service rate μ_3 was chosen so that it balanced the offered load of arriving customers. The arrival process depends on μ_3 since the service rate determines the distribution of customers at the station, on which in turn the arrival process depends.

Let $(X_t)_{t \geq 0}$ be the trace for the single-server system and $(X'_t)_{t \geq 0}$ be the trace when station 3 has two servers, both started in (different) samples from their respective stationary distributions.

Let S_n denote the sojourn time for customer n, i.e., the time spent in queue plus time spent when receiving service. Measuring consecutive sojourn times for customers at station 3 up to time $u > 0$ in both processes gives us sequences of samples $(S_n)_{n=1}^{N}$ and $(S'_n)_{n=1}^{N'}$ for $(X_t)_{t=0}^{u}$ and $(X'_t)_{t=0}^{u}$. Assuming that customers are served in the order of their arrival times, which in queueing theory is called FIFO* discipline, we may also calculate the samples of waiting times $(W_n)_{n=1}^{N}$ and $(W'_n)_{n=1}^{N'}$, i.e., the sequences of times that consecutive customers spend queueing.

In Figure 5.15 we let $u = 10,000$ and obtained $N = 18302$ samples of (S_n) and (W_n) for chain (X_t). For the second chain we got $N' = 18201$. The plots show the empirical sojourn time distributions and the empirical waiting time distributions for both chains based on these observations. We note that the waiting time for the double-server case is stochastically smaller

*First In First Out

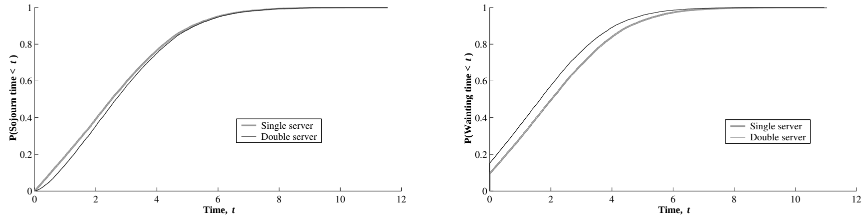

Figure 5.15. *Empirical distribution of sojourn time of customers at station 3 (left) and waiting time in queue at the same station (right). Both plots contain the distributions in the single-server and the double-server cases for the bounded Jackson network of Type 2.*

than in the single-server case, i.e., $W' \overset{\mathscr{D}}{\leq} W$, whereas for the sojourn time we have $S \overset{\mathscr{D}}{\leq} S'$. This is in fact a general result.

Now, consider again the 5-station networks of Section 5.2.b. By analysis of traces for the different networks, each started in a stationary initial state, we can estimate the blocking probabilities, i.e., the probability that an arriving customer finds a station full. Since we do not have an arrival theorem for our networks, we will notice that the fraction of arrivals that finds the bottleneck station full is not the fraction of time it is full, i.e., $\pi_3(b_3)$. One can argue that in a Type 1 network *no* arrivals find b_3 customers at station 3, but we count a customer as blocked each time a routing attempt is made and failed. Running simulations for the two scenarios we obtain Figure 5.16.

We note that in the Type 2 networks the estimated blocking probability is close to $\widehat{\pi}_3(b_3)$, while in Type 1 networks $\widehat{\pi}_3(b_3)$ seem to underestimate the blocking probability.

5.3 COMMENTS ON THE SIMULATIONS

In stochastic simulation studies there is always a note on generation of pseudo-random numbers, and now it is time for ours. The program generating all simulations was implemented in C making use of the SPRNG (Scalable Parallel Pseudo Random Number Generator) library for generating pseudo-random numbers. More specifically, what was used was the *48 bit linear congruential generator with prime addend*, which has period 2^{48} and has passed all of a series of standard statistical tests for randomness. The generator constructs sequences of pseudo-random numbers $(x_n)_{n\geq 1}$ according to the recursive relation

$$x_n = ax_{n-1} + p \pmod{2^{48}},\ n \in \mathbb{N}.$$

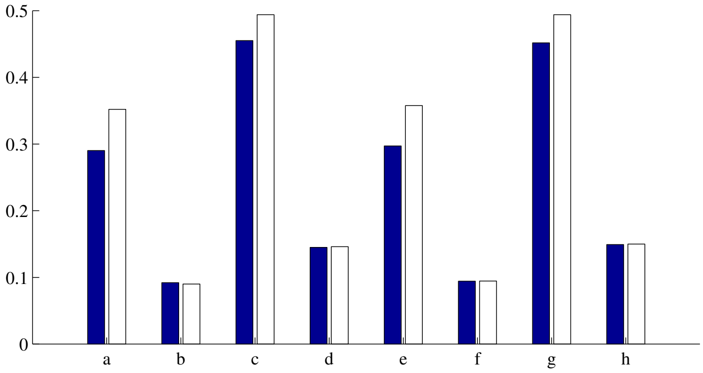

Figure 5.16. *Estimates of the blocking probability of an arriving customer at the bottleneck station (white bars), versus the estimates* $\widehat{\pi}_3(b_3)$ *(dark bars).*

Single server	*even load*	*overload*
Type 1	*a*	*c*
Type 2	*b*	*d*

Double server	*even load*	*overload*
Type 1	*e*	*g*
Type 2	*f*	*h*

where a is a multiplier and p is a prime. Different streams of random numbers use different primes p. Further information and documents on the performance are available at the www-site [12].

Conclusions

To conclude, the two models of open bounded queueing networks, given by Definition 2.5 and Definition 2.6, have different properties for instance in the number of customers at different stations. They can readily be simulated with the coupling from the past algorithm, and samples from their stationary distributions can be obtained even for quite complex queueing networks in reasonable amounts of time. In Chapter 5 some simulations indicated that the number of attempts of starting times rather surprisingly may be independent of the size and connectivity of the network. It was determined in those cases by the arrival and service rates.

It is our firm belief that these models can be successfully used to model various physical queueing networks and help in predicting qualitative performance of interest.

The two models can be combined to form queueing networks where the stations obey the routing scheme of either a Type 1 station or a Type 2.

Future research should include some extension of the models to cover closed networks and possibly different routing schemes. It should be fruitful to relax the assumptions made on the arrival processes and service times.

During a workshop in perfect simulation held in Denmark in 1997, a joint work between Lund and Wilson [8] was presented. They considered a storage system, a dam, which was exposed to stochastic amounts of rainfall at random times according to a Poisson process. The dam can be thought of as a single queue with a continuous amount of customers. Their generalisation to an infinite dam would lead us back to the ordinary Jackson network, but their work indicates that some relaxation of the exponential service times should be feasible.

References

[1] Aarssen, K. & van Dijk, N. M. A note on Jackson networks with state-dependent rejection probabilities. *Operation Research Letters* **16**, 177–181, 1994.

[2] Barndorff-Nielsen, O. E., Jensen, J. L., & Kendall, W. S., (editors). *Networks and chaos—statistical and probabilistic aspects*. Chapman & Hall, London, 1993.

[3] Boucherie, R. J. & van Dijk, N. M. On the arrival theorem for product form queueing networks with blocking. *Performance Evaluation* **29**, 155–176, 1997.

[4] Brémaud, P. *Markov Chains : Gibbs fields, Monte Carlo simulation and queues*. Springer-Verlag New York, Inc., NY, 1999.

[5] Dimakos, X. K. A guide to exact simulation. *International Statistical Review* , 1999. To appear.

[6] Fill, J. A. An interruptible algorithm for perfect sampling via Markov chains. *The Annals of Applied Probability* **8**(1), 131–162, 1998.

[7] Lindvall, T. *Lectures on the coupling method*. John Wiley & Sons Inc., New York, 1992.

[8] Lund, R. B. & Wilson, D. B. Exact sampling algorithms for storage systems, 1997. Manuscript.

[9] Norris, J. R. *Markov chains*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

[10] Propp, J. G. & Wilson, D. B. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**(1–2), 223–252, 1996.

[11] RAO, S. S., GUNASEKARAN, A., GOYAL, S. K., & MARTIKAINEN, T. Waiting line model applications in manufacturing. *Int. J. Production Economics* **54**, 1–28, 1998.

[12] Scalable Parallel Pseudo Random Number Generators Library, 15 Dec. 1999. `http://daniel.scri.fsu.edu/RNG/` (Retrieved 29 Feb. 2000.)

[13] WALRAND, J. *An Introduction to Queueing Networks*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1988.

[14] WILSON, D. B. Annotated bibliography of perfectly random sampling with Markov chains. In ALDOUS, D. & PROPP, J., (editors), *Microsurveys in Discrete Probability*, volume 41 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, page 209–220. American Mathematical Society, 1998. Updated versions to appear at `http://dimacs.rutgers.edu/~dbwilson/exact`.

[15] WILSON, D. B. How to couple from the past using a read-once source of randomness. *Random. Struct. Algorithm* **16**(1), 85–113, 2000.

[16] WOLFF, R. W. *Stochastic modeling and the theory of queues*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1989.