THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

On Perfect Simulation of Markovian Queueing Networks with Blocking

DAN MATTSSON

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY and GÖTEBORG UNIVERSITY
Göteborg, Sweden, 2002

On Perfect Simulation of Markovian Queueing Networks with Blocking
DAN MATTSSON
ISBN 91–7291–149–2

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 1831
ISSN 0346–718X

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31–772 1000

Printed and bound at School of Mathematical Sciences
Chalmers University of Technology and Göteborg University
Göteborg, Sweden, 2002

Abstract

Perfect simulation of a class of Markovian queueing networks with finite buffers is examined in this monograph. The network processes defined are modelled by restricted multivariate birth and death processes with intensities depending on a random environment. The random environments are mainly a set of independent ON/OFF processes. Several simulation experiments illustrate the use of the perfect sampling algorithm and properties of the networks.

The class of networks is analysed by coupling techniques, and monotonic properties for network trajectories are established. The networks are simulated by embedding a coupling, which imposes an ordering of network trajectories, in a two-dimensional Poisson process, and by using a monotone Propp–Wilson algorithm. The effectiveness of the algorithm is examined by a series of simulation experiments and is found to be satisfactory. The defined networks can be extended to handle batch arrivals, rerouting of blocked units, buffer sizes varying over time, and other types of environments.

The class of networks devised can be used as a model for a wide range of applications in which the Markovian assumptions are satisfied. Even for complex and large networks, samples are obtained in a reasonable amount of time although the simulation time is hard to predict beforehand.

Keywords: Markov chain Monte Carlo methods, Propp–Wilson algorithm, Perfect simulation, Queueing theory, Jackson networks, Birth and Death processes.
MSC 2000 subject classifications: 60K20, 60K25, 68M20, 90B22, 65C05, 60E15, 60J80

Contents

Acknowledgements

No man is an island...

and thus I am grateful for the good nature and grace of many people: those who 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 them all.

I would like to mention Olle Häggström for his valuable comments on a draft of this manuscript and, above all, my inspiring adviser Torgny Lindvall, who was always ready to provide support, encouragement and optimism.

Dan Mattsson
Göteborg, April 2002

Introduction

This thesis is devoted to the study of perfect simulation of queueing networks. A perfect sampling algorithm could be a useful tool for analysing and optimising queueing networks. A class of queueing networks was devised, similar to the Jackson networks, but with restrictions placed on the number of units present at the nodes. The similarity to the Jackson networks is that, basically, Poisson arrival streams, exponential service times and Markovian routing are used. These quantities may, however, be functions of a random environment.

The main objective is to show how, in this class of networks, samples from the stationary distribution of the number of units can be obtained by simulation, that is, using terminology from the theory of Monte Carlo simulation, to obtain *perfect* samples, or to develop a perfect or *exact* simulation algorithm. It is shown that the class of networks devised has some monotonic properties which enable one to couple realisations of networks and to apply the monotone Propp–Wilson algorithm to obtain perfect samples. Simulation experiments and samples from the stationary distributions should yield information about the features of any network under consideration; examples are information about bottlenecks, i.e., nodes in the network where congestion is likely to occur, and the prediction of network performance when some fundamental parameters change, such as increased waiting spaces, higher arrival rates, or possibly server crashes.

The networks presented here are Markov processes which can be embedded in a two-dimensional Poisson process. It is the embedding technique which enables one to simulate and couple the trajectories in a simple way, making the Propp–Wilson algorithm computationally effective.

The goal is achieved in several steps which form the outline of the thesis. Chapter 1, which can be skipped by most readers, contains some basic theory presented for completeness. The first parts, Sections 1.1 and 1.2, treat basic Markov theory and the theory of point processes. The Poisson point process is fundamental, and extensive use is made of it. Other processes used are multivariate birth and death processes, in particular ON/OFF processes, and Markov modulated Poisson processes; the latter form a foundation for the random environments of the networks presented. The birth and death processes serve to keep track of the number of units at different nodes of a network.

By observing that the birth and death processes are uniformisable, in Section 1.3, one can construct uniform variants of them with corresponding discrete subordinate chains. These chains are constructed, in Section 1.4, by embedding them in a two-dimensional Poisson process. Once embedded, it is a simple task to couple their trajectories; this is dealt with in Chapter 3, after the model of queueing networks is established.

Chapter 1 also includes a brief description of orderings and coupling theory, Section 1.5.

In Chapter 2 the new queueing networks are defined. A brief introduction to queueing networks in general is followed by Jackson networks in particular. Section 2.3 defines the class of networks, the SBE networks, followed by 2.4 on some queueing quantities of interest. For example, it is shown how samples of the time that units wait, at a service facility without actually being served, can be obtained from trajectories of a network process; this enables us to collect not only samples on the stationary distribution, but also samples on these quantities as well. The chapter ends with Section 2.5, where coupling techniques are used to show some analytically tractable properties of the new networks.

Chapter 3 begins with a section showing the construction of a subordinate chain of a network from the two-dimensional Poisson embedding. Section 3.2 describes the Propp–Wilson algorithm and how it is used on the networks to obtain perfect samples. The chapter ends with a section containing simulation experiments showing the effectiveness of the algorithm for different parameter settings, such as number of nodes, buffer sizes, and network load.

Chapter 4 shows how perfect samples can be used to analyse the networks, as well as to highlight selected properties and characteristics of them. Many of the theoretical results of Section 2.5 are also illustrated.

Next, Chapter 5 gives some of the possible model extensions.

Chapter 6 is a discussion on limitations of the queueing network model and possible feasible extensions of the model for future research.

The intended reader is an engineer with a background in basic probability and Markov theory. As a result of this, an effort was made to illustrate theorems and statements with examples and to describe alternative interpretations of the models devised. I hope readers of deeper background will also find this beneficial.

Chapter 1

Fundamentals

To establish the notations and conventions used throughout the thesis, as well as to formally present basic theorems and definitions, are the aims of this chapter. A list of symbols can be found at the end of the thesis.

Here n shall always 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, $(X_n)_{n\geqslant 0}$ denotes a Markov chain in discrete-time and $(X_t)_{t\geqslant 0}$ one in continuous time. Further, d as a constant always denotes a dimension, $d \in \mathbb{N}$, and $\boldsymbol{b}$ is a d-dimensional vector of bounds,

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

The vectors $\mathbf{0}_r = (0, \dots, 0)$ and $\mathbf{1}_r = (1, \dots, 1)$ are of length r, and are denoted by $\mathbf{0}$ and $\mathbf{1}$ when $r = d$.

For convenience, the somewhat unorthodox notation $\mathbb{Z}_+^{\boldsymbol{b}}$ is used 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, $k = 1, \dots, d$. For $A \subseteq \{1, \dots, d\}$ let $e_A = \sum_{k\in A} e_k$, and let $<x, e_A>$ denote $(x_k)_{k\in A}$.

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space. When a random variable X with state space $(E, \mathscr{E})$ is governed by probability law P, that is, $\mathsf{P}(X \in A) = P(A)$ for $A \in \mathscr{E}$, it is written $X \overset{\mathscr{D}}{=} P$. When two random variables X and Y in $(E, \mathscr{E})$ have the same law, it is written $X \overset{\mathscr{D}}{=} Y$.

A common probability distribution is the *geometric distribution*: a random variable X is geometrically distributed with parameter $p \in (0,1)$ if

$$\mathsf{P}(X = k) = p(1-p)^{k-1}, \quad k \in \mathbb{N}.$$

Some authors, however, call the distribution of $X' = X - 1$ the geometric distribution. The truncated geometric distribution is a generalisation of the distribution of $Y \overset{\mathscr{D}}{=} X'|X' \leqslant b$, where X' is geometrically distributed in the second sense.

To be precise, Y has a truncated geometric distribution with parameter ρ if

$$\mathsf{P}(Y = k) = \begin{cases} \dfrac{1-\rho}{1-\rho^{b+1}}\,\rho^k & \text{when } \rho \neq 1 \\ \dfrac{1}{b+1} & \text{when } \rho = 1 \end{cases} \qquad k = 0, 1, \ldots, b,$$

for some constant $b \in \mathbb{Z}_+$ and $\rho > 0$.

1.1 MARKOV THEORY

The statements on Markov theory made here, which can be found in any textbook on the subject, may be bypassed by readers familiar with it. For proofs of theorems and statements, see Norris [18].

One may define continuous-time Markov chains by starting from a semi-group of transition probabilities or from infinitesimal intensities of transitions. The latter approach was chosen, since intensity matrices are the natural framework for the work presented here.

Definition 1.1.1. For a finite or countably infinite state space E, an *intensity matrix*, $Q = [q_{xy}]_{x,y \in E}$, is a matrix with elements satisfying:

- $0 \leqslant q(x) < \infty$, where $q(x) = -q_{xx}$, for all $x \in E$. (Stability);
- $q_{xy} \geqslant 0$ for all $x, y \in E$ such that $x \neq y$; and
- $\sum_{y \in E} q_{xy} = 0$ for all $x \in E$. (Conservation).

In the following, $-q(x)$ is written for the diagonal elements of Q. In Section 1.3 a Markov chain $(X_t)_{t \geqslant 0}$ is constructed from a specified intensity matrix Q. To this end, assumptions must be made about Q, for example that it is non-explosive, for the Markov chains to be *regular*, i.e., to have piecewise constant trajectories and finitely many jumps in finite time. A necessary and sufficient condition is known as Reuter's Criterion, but observe that, for finite state spaces, the simpler sufficient condition

$$\sup_{x \in E} q(x) < \infty \tag{1.1}$$

is trivially satisfied; see Brémaud [5] for a thorough treatment of the subject. Note that some authors refer to intensity matrices as *infinitesimal generators*. Markov chains with intensity matrices satisfying (1.1) are called *uniformisable*.

Assume throughout that the Markov chains have $q(x) > 0$ for all $x \in E$. Thus, the Markov chains will eventually leave every state, i.e., there are not any absorbing (permanent) states.

Definition 1.1.2. Associated with the intensity matrix is a semigroup of *transition matrices*,

$$P_t = [p_{xy}(t)]_{x,y \in E}\,, \quad t \geqslant 0,$$

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

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

and I is the identity matrix $I = [\delta_{xy}]_{x,y \in E}$. The matrices P_t are stochastic matrices, i.e., the rows sum to unity, and they form a semigroup, since they satisfy $P_{s+t} = P_s P_t$ for all $s, t \geqslant 0$.

In the following, let μ be a probability measure on E.

Definition 1.1.3. The stochastic process $(X_t)_{t \geqslant 0}$ is a (continuous-time) homogeneous Markov chain on E with initial distribution μ, if $X_0 \overset{\mathscr{D}}{=} \mu$, and for all $m \in \mathbb{N}$, $0 \leqslant t_1 < \cdots < t_{m+1}$, $x_1, \ldots, x_{m+1} \in E$,

$$\mathsf{P}\left(X_{t_{m+1}} = x_{m+1} \,|\, X_{t_1} = x_1, \ldots, X_{t_m} = x_m\right) \;=\; p_{x_m, x_{m+1}}(t_{m+1} - t_m).$$

A (discrete-time) Markov chain, $(X_n)_{n \geqslant 0}$, on E is when $0 \leqslant t_1 < \cdots < t_{m+1}$ are chosen from the non-negative integers. Associated with a discrete-time Markov chain is an n-step transition matrix $P_n = [p_{ij}(n)]_{i,j \in E}$, where $p_{ij}(n) = \mathsf{P}(X_n = j | X_0 = i)$.

The next theorem states that well-behaved continuous Markov chains are determined by their intensity matrices.

Theorem 1.1.1. *Two regular homogeneous continuous Markov chains with the same intensity matrix, and the same initial distribution, have the same transition semigroup. Hence, the finite-dimensional distributions coincide and the chains are probabilistically equivalent.*

The next series of definitions leads to the definition of *ergodic* Markov chains, Definition 1.1.6. All queueing networks considered in Chapter 2 are described by ergodic processes, and the fundamental properties of such processes are stated in Theorems 1.1.5 and 1.1.6. The definitions and theorems are formulated only for continuous Markov chains. They hold for and are applicable to, unless otherwise stated, discrete chains as well, choosing the times t from the non-negative integers.

Definition 1.1.4. A state $y \in E$ is *accessible* from $x \in E$ if there exists a $t > 0$ such that $p_{xy}(t) > 0$. If every state x is accessible from every other state y, then the chain, and its intensity matrix, are said to be *irreducible*.

Fix $x \in E$ and let $\mu = \delta_x$.

Definition 1.1.5. The return time to state x is defined by

$$R_x = \inf\{t > 0 : X_t = x,\ \exists s < t : X_s \neq x\}.$$

A state x is *recurrent* if $\mathsf{P}_x(R_x < \infty) = 1$ and *transient* otherwise. If x is recurrent and $\mathsf{E}_x[R_x] < \infty$, then x is *positive recurrent*, otherwise *null recurrent*.

Theorem 1.1.2. *Let $(X_t)_{t\geqslant 0}$ be an irreducible Markov chain on E with intensity matrix Q. Then one of the following holds:*

- *all states are positive recurrent,*
- *all states are null recurrent, or*
- *all states are transient.*

Our attention is restricted to Markov chains with finite state spaces, typically $E = \mathbb{Z}_+^{\boldsymbol{b}}$, where, by the following theorem, positive recurrence is immediate from irreducibility.

Theorem 1.1.3. *If E is finite and $(X_t)_{t\geqslant 0}$ is irreducible, then $(X_t)_{t\geqslant 0}$ is positive recurrent.*

The following definition is applicable to continuous chains.

Definition 1.1.6. A continuous, irreducible and positive recurrent Markov chain is called *ergodic*.

For discrete chains an additional assumption is made in the definition of ergodicity.

Definition 1.1.7. For an irreducible, discrete-time Markov chain with n-step transition matrices $P_n = [p_{ij}(n)]_{ij\in E}$, let $d = \gcd\{n > 0 : p_{ii}(n)\}$. Here 'gcd' stands for 'greatest common divisor', and one can show that irreducibility implies that d does not depend on i. If $d = 1$ the Markov chain and the transition matrices are called *aperiodic*. An irreducible, positive recurrent, aperiodic discrete Markov chain is called *ergodic*.

Definition 1.1.8. Consider a semigroup of transition matrices P_t, $t \geqslant 0$. A probability measure π on E that satisfies

$$\pi = \pi P_t, \qquad \text{for all } t \geqslant 0,$$

where π is regarded as a row vector, is called a *stationary distribution*.

Suppose that the Markov chain $(X_t)_{t\geqslant 0}$ has a semigroup of transition matrices P_t that allows the stationary distribution π. If π is the initial distribution of X, $X_0 \stackrel{\mathscr{D}}{=} \pi$, then $X_t \stackrel{\mathscr{D}}{=} \pi$ for all $t \geqslant 0$. The chain is then said to be in *equilibrium*, and the crucial point is that the distribution of X_t does not depend on the time t.

For discrete chains, Definition 1.1.8 can be used in determining the stationary distribution, by solving $\pi = \pi P$, where P is the one-step transition matrix. For continuous chains, the following theorem relates π to Q.

Theorem 1.1.4. *Let Q be an intensity matrix with $\sup_{x\in E} q(x) < \infty$. A necessary and sufficient condition for π to be a stationary distribution of the induced semigroup of transition matrices P_t, is that π satisfies*

$$\pi(x) \sum_{\substack{y\in E\\ y\neq x}} q_{xy} = \sum_{\substack{y\in E\\ y\neq x}} \pi(y) q_{yx}, \quad x \in E, \tag{1.2}$$

or, more compactly,

$$\pi Q = \mathbf{0},$$

where π is regarded as a row vector. Further, the measure π exists (and is unique) if, and only if, the chain $(X_t)_{t\geqslant 0}$ is ergodic.

Equation (1.2) is sometimes referred to as the *full (or global) balance equation*. The left-hand side of (1.2) is simply $\pi(x)q(x)$. The statement on existence also applies to discrete-time Markov chains.

The next theorem links the concepts of ergodicity and stationarity, by stating that the limiting distribution of an ergodic Markov chain exists and is the stationary distribution.

Theorem 1.1.5. *Let $(X_t)_{t\geqslant 0}$ be an ergodic Markov chain on E with intensity matrix Q and stationary distribution π. Then, for the induced semigroups $(P_t)_{t\geqslant 0}$,*

$$\lim_{t\to\infty} p_{xy}(t) = \lim_{t\to\infty} \mathsf{P}_x(X_t = y) = \pi(y), \quad y \in E,$$

regardless of the initial state $x \in E$.

This implies that $\pi(y) = \lim_{t\to\infty} \mathsf{P}_\mu(X_t = y)$ for $y \in E$, regardless of the initial distribution μ.

Note that μP_t equals the distribution of the continuous chain at time t, X_t, when $X_0 \overset{\mathscr{D}}{=} \mu$. Thus, in light of Theorem 1.1.4, the distribution of X_t does not depend on t if, and only if, $\mu = \pi$, which means that there is no hope of obtaining a sample from the stationary distribution by simply observing $(X_t)_{t\geqslant 0}$ at a predefined time t, unless $(X_t)_{t\geqslant 0}$ is stationary from the start. This does not apply to discrete-time Markov chains, as can be seen by choosing one row of P_1 to equal π. The Propp–Wilson algorithm (Chapter 3) makes it possible, however, to obtain samples from π of a continuous chain, by an informed choice of stopping times of $(X_t)_{t\geqslant 0}$.

Theorem 1.1.6. *Let $(X_t)_{t\geqslant 0}$ be as in Theorem 1.1.5. Then, with $Y \overset{\mathscr{D}}{=} \pi$,*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X_s)\,ds = \sum_{y\in E} f(y)\pi(y) = \mathsf{E}[f(Y)] \quad \text{[a.s.]},$$

for every initial distribution $X_0 \overset{\mathscr{D}}{=} \mu$, and every function $f : E \to \mathbb{R}$ such that $\sum_{y\in E} |f(y)|\pi(y) < \infty$.

Theorem 1.1.6 captures the usefulness of the stationary distribution. It enables us to calculate the almost sure limiting time-average of a function of an ergodic Markov chain simply as an expectation.

The section concludes with a definition of a class of multivariate Markov processes.

Definition 1.1.9. Let $F = E \times G$ be a state space for a Markov chain $(Y_t)_{t \geqslant 0}$, and write $Y_t = (X_t, Z_t)$ where $X_t \in E$ and $Z_t \in G$. If $(Z_t)_{t \geqslant 0}$ is a Markov chain on G, then $(Z_t)_{t \geqslant 0}$ is a (random) environment process *of* $(Y_t)_{t \geqslant 0}$, and *to* $(X_t)_{t \geqslant 0}$.

1.2 POISSON POINT PROCESSES

This section defines formally a collection of basic point processes, in particular the Poisson process. A general definition is stated first, after which it is applied to the spaces of interest.

Let $\mathcal{S}$ be a complete separable metric space equipped with the σ-algebra $\mathscr{B}(\mathcal{S})$. The reader may regard the space as a Euclidian space such as $\mathbb{R}_+$, but the definition is stated in this general framework. Let $\mathcal{N}$ be the space of all integer-valued measures of $\mathcal{S}$ satisfying $N(A) \in \mathbb{Z}_+$ for every compact Borel set $A \in \mathscr{B}(\mathcal{S})$. The space $\mathcal{N}$ consists of the counting measures on $\mathcal{S}$. The attention is not placed on $\mathcal{N}$ directly, but, rather, on the properties of the measures it contains. For a detailed treatment, see Daley and Vere-Jones [6].

A point process, N, on $\mathcal{S}$ is a measurable mapping from a probability space $(\Omega, \mathscr{F})$ to $(\mathcal{N}, \mathscr{B}(\mathcal{N}))$; but essentially, $N(A)$, the number of points that fall in the compact Borel set $A \subseteq \mathcal{S}$, is an integer-valued random variable. A point process N, such that $N(\{t\}) \in \{0, 1\}$ for every $t \in \mathcal{S}$, is called simple. If $\mathsf{P}(N(\{t\}) \in \{0, 1\} \text{ for all } t \in \mathcal{S}) = 1$, it is said that N is simple almost surely.

Every realisation $N(\omega)$ of a simple point process can be expressed as

$$N(\omega, \cdot\,) = \sum_{k=1}^{\infty} \delta_{t_k(\omega)}, \tag{1.3}$$

where $\delta_t(A) = \mathbf{1}_{\{t \in A\}}$. The sequence $(t_k)_{k=1}^{\infty}$ of elements of $\mathcal{S}$ corresponds to the distinct locations of the points of the realisation. Fix ω and regard the realisation of a simple point process N as in (1.3). Let α be a probability distribution on $\mathbb{Z}_+$. Define the corresponding point process, N_α, with weight distribution α by

$$N_\alpha(\omega, \cdot\,) = \sum_{k=1}^{\infty} W_k(\omega) \delta_{t_k(\omega)}, \tag{1.4}$$

where W_k, $k \geqslant 1$, are i.i.d. random variables with distribution α. Note that this is actually a statement about individual realisations, but done rigorously, N_α is a point process. The weights, W_k, correspond to the number of multiple points at the same location, and the simple process N constitutes the support. If $\alpha(1) = 1$, then $W_k(\omega) = 1$ almost surely for $k \geqslant 1$, and hence, $N_\alpha = N$.

Definition 1.2.1. Let Λ be a measure on $\mathcal{S}$, finite on compact Borel sets. Then N is a *Poisson process* on $\mathcal{S}$ with parameter measure Λ, if, for every finite collection of disjoint compact sets $A_1, \ldots, A_n \in \mathscr{B}(\mathcal{S})$,

1. the random variables $N(A_1), \ldots, N(A_n)$ are independent, and

2. for $A \in \mathscr{B}(\mathcal{S})$, $N(A)$ is Poisson distributed with expected value $\Lambda(A)$, that is,

$$\mathsf{P}(N(A) = k) = \frac{\Lambda(A)^k}{k!} \mathrm{e}^{-\Lambda(A)}, \quad k = 0, 1, 2, \ldots$$

Note that the terminology of Daley and Vere-Jones [6] is used: Λ is a parameter measure. Another common term is *expectation measure*.

When $\mathcal{S} = \mathbb{R}_+$, the point processes N are one-dimensional, and $\mathbb{R}_+$ is interpreted as a time axis. Then $N((0,t])$ is written as N_t, $N(t)$, or possibly $N(0,t]$.

The parameter measures Λ of interest, can for every $A \in \mathscr{B}(\mathbb{R}_+)$ be expressed as

$$\Lambda(A) = \int_A \lambda(t)\, dl(t) \tag{1.5}$$

for a non-negative function λ, the *intensity process* of N. Equation (1.5) implies, among other things, that the one-dimensional Poisson process with expectation measure Λ is simple almost surely.

Suppose N is a one-dimensional Poisson process with parameter measure Λ satisfying (1.5) with constant intensity process λ. Then $\Lambda(A) = \lambda l(A)$ for every $A \in \mathscr{B}(\mathbb{R}_+)$, and N is a (time) homogeneous Poisson process. The constant λ is called simply the intensity, or rate, of the process. To use the results of Section 1.1, N may be regarded as the Markov process $(N_t)_{t \geqslant 0}$

with $N_0 = 0$ and intensity matrix Q given by,

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Using (1.4) where α is a probability distribution on $\mathbb{Z}_+$, one may from N define the point process N_α. Such a process is commonly referred to as a *compound* Poisson process. Note that the distribution α can, without loss of generality, be assumed to have no mass at 0.

The following theorem is quoted from Daley and Vere-Jones [6].

Theorem 1.2.1. *Let $\{N_k\}_{k\geqslant 1}$ be a family of independent one-dimensional Poisson processes with parameter measures $\Lambda_1, \Lambda_2, \ldots$ satisfying* (1.5) *for intensities $\lambda_1(t), \lambda_2(t), \ldots$, respectively. Let $\lambda(t) = \sum_{k=1}^{\infty} \lambda_k(t)$. If $\lambda(t) < \infty$ for all $t \geqslant 0$, then $N = \sum_{k=1}^{\infty} N_k$ is a Poisson process with parameter measure Λ satisfying* (1.5) *with intensity λ.*

Hence, since (1.5) is satisfied, N is simple almost surely.

As an example in two dimensions, choose $\lambda > 0$, take $\mathcal{S} = [0, \lambda]\times\mathbb{R}_+$ and let N be the Poisson process with the Lebesgue measure as parameter measure. Then it is clear that the process $N_{(a,b]}$, $0 \leqslant a < b \leqslant \lambda$, defined by

$$N_{(a,b]} = N((a, b]\times \cdot\,) \tag{1.6}$$

is a one-dimensional homogeneous Poisson process with intensity $(b - a)$. Further, for $0 \leqslant a < b < c < d \leqslant \lambda$, the processes

$$N_{(a,b]} \text{ and } N_{(c,d]}$$

are independent, since the strips $(a, b]\times\mathbb{R}_+$ and $(c, d]\times\mathbb{R}_+$ are disjoint Borel sets.

Theorem 1.2.2 (Competition Theorem). *Let $\{N_k\}_{k\geqslant 1}$ be a family of independent homogeneous one-dimensional Poisson processes with intensities $\lambda_1, \lambda_2, \ldots$, respectively. Let*

$$N = \sum_{k\geqslant 1} N_k, \qquad \lambda = \sum_{k\geqslant 1} \lambda_k,$$

and suppose that $\lambda < \infty$. Let Z be the time of the first event of N, that is,

$$Z = \inf\{t \geqslant 0 : N(t) > 0\}.$$

Denote with K the index of the process producing the event, that is, for $k \neq K$, $N_k(Z) = 0$, and $N_K(Z) = 1$. Then

$$\mathsf{P}(K = k, Z > t) = \mathsf{P}(K = k)\mathsf{P}(Z > t) = \frac{\lambda_k}{\lambda} \cdot \mathrm{e}^{-\lambda t}.$$

Specifically, Z is exponentially distributed.

By Theorem 1.2.1, N is a homogeneous Poisson process satisfying (1.5), and thus, simple. Hence, the random variable K in Theorem 1.2.2 is well defined.

These are the definitions of a Poisson process. However, for convenience, Poisson processes are here sometimes said to have intensity $\lambda = 0$, or parameter measure $\Lambda \equiv 0$. They are simply the point processes without points.

1.2.1 MARKOV MODULATED POISSON PROCESSES

A Markov modulated Poisson process, abbreviated MMPP, is a Poisson process with an intensity λ changing over time according to an underlying Markov chain. This is a special case of a Cox process.

Let $X = (X_t)_{t \geqslant 0}$ be a Markov chain on a finite state space E and fix $\lambda : E \to \mathbb{R}_+$. For every $x \in E$, let $\lambda(x)$ be an intensity, where one allows $\lambda(x) = 0$ for some $x \in E$. Here, λ is referred to as the *conditional intensity (process)* of N with respect to X.

The random parameter measure Λ, associated with the conditional intensity process, is defined for every bounded $A \in \mathscr{B}(\mathbb{R}_+)$ and $\omega \in \Omega$, by

$$\Lambda(A, \omega) = \int_A \lambda(X_t(\omega))\, dl(t). \tag{1.7}$$

Compare equations (1.7) and (1.5).

The Markov modulated Poisson process $N = (N_t)_{t \geqslant 0}$ is defined as the Poisson process on $\mathbb{R}_+$ with random parameter measure Λ as in (1.7). The process X is called the driving chain of the MMPP N. Note that the random measure Λ is defined not on sets but for every $\omega \in \Omega$, that is, for every realisation of X.

An explicit construction of an MMPP N is the following. Let the intensity matrix of X be given by $Q = [q_{xy}]_{x,y\in E}$, and define $N^* = (N_t^*)_{t\geqslant 0}$ to be the Markov chain on $\mathbb{Z}_+ \times E$ with intensity matrix Q_N where the possibly non-zero elements are for transitions

from	**to**	**with intensity**
(n, x)	$(n+1, x)$	$\lambda(x)$
(n, x)	(n, y)	q_{xy},

(1.8)

for $n \in \mathbb{Z}_+$, $x, y \in E$, $x \neq y$. The diagonal elements of Q_N satisfy

$$q^{(N)}_{(n,x),(n,x)} = -\sum_{\substack{y\in E\\ y\neq x}} q_{xy} - \lambda(x),$$

so that the rows of Q_N sum up to zero. The MMPP N at $t \geqslant 0$ is given by $N_t = <N_t^*, e_1>$, the first coordinate of N_t^*. The intensity of N at time t is simply $\lambda(<N_t^*, e_2>)$. Note that the process $(X_t)_{t\geqslant 0} = (<N_t^*, e_2>)_{t\geqslant 0}$, and that $(X_t)_{t\geqslant 0}$ is a random environment of $(N^*(t))_{t\geqslant 0}$. If E only has a single element, then X_t is constant and N_t reduces to a time homogeneous Poisson process.

1.2.2 BIRTH AND DEATH PROCESSES

Definition 1.2.2. Fix $d \in \mathbb{N}$ and let the state space $E = \mathbb{Z}_+^d$ or some finite subset thereof. A *multivariate birth and death* process, $X = (X_t)_{t\geqslant 0}$, on E, is a Markov chain with intensity matrix $Q = [q_{xy}]_{x,y\in E}$, where the elements satisfy, for $x, y \in E$, $x \neq y$,

$$q_{xy} > 0 \quad \text{only if} \quad x + e_B = y + e_D$$

for subsets $B, D \subseteq \{1, \ldots, d\}$ satisfying

$$B \cap D = \varnothing \qquad B \cup D \neq \varnothing.$$

Let $\gamma_{D,B}(x)$ denote the intensity of transitions from x to $x - e_D + e_B$, for disjoint sets D and B, and introduce the special notation

$$\gamma_{\varnothing,B}(x) \text{ is denoted by } \beta_B(x)$$
$$\gamma_{D,\varnothing}(x) \text{ is denoted by } \delta_D(x).$$

The space $E = \mathbb{Z}_+^d$ denotes the space of all possible configurations of a network with d nodes. For a configuration $x \in E$, x_k denotes the number of units at node k. Multivariate birth and death processes with sets B and D, either empty or containing only single elements, are the main interest here.

Suppose that $k, m \in \{1, \ldots, d\}$, $k \neq m$, and $x \in E$. Then $\beta_{\{m\}}(x)$, $\delta_{\{k\}}(x)$ and $\gamma_{\{k\},\{m\}}(x)$ are written as $\beta_m(x)$, $\delta_k(x)$ and $\gamma_{km}(x)$, respectively.

The multivariate birth and death process is used both as a model of queueing networks, and as a driving chain for a Markov modulated point processes which, in turn, serve as e.g. arrival processes. In the queueing theory context, the interpretation for transitions

1. from x to $x + e_m$ is an arrival (a birth) of a unit to (in) node m;
2. from x to $x - e_k$, corresponds to a departure (a death) of a unit from (in) node k; and (1.9)
3. from x to $x + (e_m - e_k)$, the transfer of a unit from node k to m.

1.2.3 ON/OFF PROCESSES AND IPP

Assume that N is a Markov modulated Poisson process on $\mathbb{R}_+$, where the underlying Markov chain X has a state space E with two elements labelled $E = \{\text{off}, \text{on}\}$. Let λ_{on} and λ_{off} be two non-negative constants. From the reasoning in Section 1.2.1, N is the simplest non-trivial example of an MMPP. The intensity matrix Q of X is

$$Q = \begin{bmatrix} -\sigma & \sigma \\ \nu & -\nu \end{bmatrix},$$

for some constants $\sigma > 0$ and $\nu > 0$. The 'off' state corresponds to the first row of Q. When enumerating the states as $E = \{0, 1\}$, one finds that X is a special case of a one-dimensional birth and death process.

When, for example, $\lambda_{\text{off}} = 0$, then as X oscillates between its two states, the MMPP N have intervals during which there are no arrivals, followed by intervals for which the arrival intensity is positive. For intervals $I = \{t : X_t = \text{on}\}$, N_t, $t \in I$, is a Poisson process with intensity λ_{on}. This particular kind of MMPP is commonly referred to as an interrupted Poisson process (IPP). Similarly, $(X_t)_{t \geqslant 0}$ is referred to as an ON/OFF process. The parameter ν is the intensity with which the chain leaves the on-state, hence, the expected length of an 'on' period is $1/\nu$, and similarly, $1/\sigma$ is the expected length of an 'off' period. An example is shown below.

To achieve the final generalisation, $r \in \mathbb{N}$, a positive integer, is introduced.

Definition 1.2.3. If $X_1, \ldots, X_r$ are r independent ON/OFF processes, then $X = (X_1, \ldots, X_r)$ is a multivariate birth and death process, with $\gamma_{km} = 0$ in (1.9), for $k, m \in \{1, \ldots, r\}$, $k \neq m$. Similarly, $\beta_k(x) = \mathbf{1}_{\{x_k=0\}}\sigma_k$ and $\delta_k(x) = \mathbf{1}_{\{x_k=1\}}\nu_k$. The process X is referred to as a *Poisson environment.*

For convenience, the notations $\nu_k(x) = \nu_k \mathbf{1}_{\{x_k=1\}}$, $\sigma_k(x) = \sigma_k \mathbf{1}_{\{x_k=0\}}$, are used, where the parameters are regarded as functions of the current state of the environment. Since the multivariate birth and death process X plays a specific rôle as a driving chain for the conditional intensity λ, and since it has a certain structure, it is given a special name: Poisson environment. It should be noted, however, that there exist not any standard terminology for the processes in this section; they are well-studied but known under many different names.

The MMPP processes N of interest have parameter measures Λ constructed from a multivariate version of (1.7) for some conditional intensity process $\lambda : \{0,1\}^r \to \mathbb{R}_+$ and Poisson environment X. This is a rather general setting, since λ may be defined quite arbitrarily. If $\lambda_k(x)$ are functions, $\lambda_k : \{0,1\} \to \mathbb{R}_+$, $k = 1, \ldots, r$, and

$$\lambda(x_1, \ldots, x_r) = \sum_{k=1}^{r} \lambda_k(x_k), \tag{1.10}$$

then the process N can be regarded as a superposition of independent interrupted Poisson point processes $N_1, \ldots, N_r$, with conditional intensities $\lambda_1(x_1), \ldots, \lambda_r(x_r)$, respectively.

The final step in the generalisation is when α is a probability distribution on $\mathbb{Z}_+$, and N_α is defined from N as in (1.4).

Example 1.2.1. Consider a Poisson environment consisting of two interrupted Poisson process. Let $E = \{0,1\}$ and X, X' be two ON/OFF processes on E with expected on- and off-lengths

$$\begin{aligned} \nu &= 1/10 & \nu' &= 1/2 \\ \sigma &= 1/10 & \sigma' &= 1/5. \end{aligned}$$

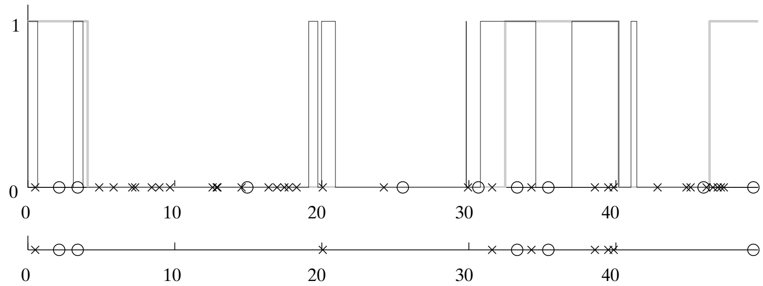

Figure 1.1. *Example of a Markov modulated Poisson process obtained from two interrupted Poisson processes. The thick grey line corresponds to the* ON/OFF *process X, and the circles to events of a homogeneous Poisson process with intensity $\lambda = 1/5$. The thin black line shows the trajectory of X', and the crosses events of a homogeneous Poisson process with intensity $\lambda' = 1/2$. The resulting point process N^* is shown separately with the marks of the individual processes N and N' preserved. This illustrates Example 1.2.1.*

The conditional intensities for the interrupted Poisson processes N and N', when the ON/OFF processes are on, are $\lambda = 1/5$ and $\lambda' = 1/2$, respectively. Figure 1.1 shows a realisation of the resulting Markov modulated Poisson process N^* obtained as a superposition of the interrupted Poisson processes, that is, $N^* = N + N'$.

1.3 UNIFORMISATION

This section is devoted to the construction and simulation of Markov chains using the Poisson processes just discussed. This will eventually lead to the construction of the couplings of Section 3.1.

Take $d \in \mathbb{N}$, $\boldsymbol{b} \in \mathbb{N}^d$ and let $E = \mathbb{Z}_+^{\boldsymbol{b}}$. To begin, note that there is a one-one correspondence between a continuous-time Markov chain $(X(t))_{t\geqslant 0}$ on E, and two stochastic processes $(X_n)_{n\geqslant 0}$ and $(\tau_n)_{n\geqslant 0}$. The discrete-time chain $(X_n)_{n\geqslant 0}$ on E keeps track of the successive states visited by $X(t)$, while $(\tau_n)_{n\geqslant 0}$ are the times of these transitions. Note that $X(t)$ is the continuous chain at time $t \geqslant 0$, and X_n is the state of the discrete chain after the n^{th} transition. It is assumed that the Markov chain $(X(t))_{t\geqslant 0}$ is uniformisable.

The sequences, $(X_n)_{n\geqslant 0}$ and $(\tau_n)_{n\geqslant 0}$, induced by $(X(t))_{t\geqslant 0}$, can be expressed as $X_0 = X(0)$, $\tau_0 = 0$, and

$$\begin{aligned}\tau_{n+1} &= \inf\{t > \tau_n : X(t) \neq X(\tau_n)\}\\ X_{n+1} &= X(\tau_{n+1}),\end{aligned}$$

for $n \geqslant 0$. The sequence of times between successive transitions is denoted

$$Z_{n+1} = \tau_{n+1} - \tau_n, \quad n \geqslant 0,$$

hence, $(Z_n)_{n\geqslant 1}$ is a sequence of exponentially distributed random variables with intensities depending on the sequence $(X_n)_{n\geqslant 0}$. It can be shown that

$$\begin{aligned}\mathsf{P}(X_{n+1} = y, Z_{n+1} > t | X_n) &= \mathsf{P}(X_{n+1} = y | X_n)\mathsf{P}(Z_{n+1} > t | X_n)\\ &= \frac{q_{X_n,y}}{q(X_n)} \cdot \mathrm{e}^{-tq(X_n)},\end{aligned} \tag{1.11}$$

for $t \in \mathbb{R}_+$, $y \in E\backslash\{X_n\}$, $n \geqslant 0$. That is, the next state, X_{n+1}, and the time to the transition, Z_{n+1}, are conditionally independent, when conditioned on the previous state X_n. The process $(X_n)_{n\geqslant 0}$ is called the *embedded* process of $(X_t)_{t\geqslant 0}$.

Going in the other direction is as simple. Take $(Z_n)_{n\geqslant 1}$ and $(X_n)_{n\geqslant 0}$ such that (1.11) is satisfied. Let $\tau_0 = 0$ and $\tau_{n+1} = \tau_n + Z_{n+1}$ for $n \geqslant 0$, and construct the continuous-time Markov chain $(X(t))_{t\geqslant 0}$, by letting

$$X(t) = X_n \quad \text{for } t \in [\tau_n, \tau_{n+1}), n \geqslant 0. \tag{1.12}$$

The main point is to show how sequences $(\tau_n)_{n\geqslant 0}$ and $(X_n)_{n\geqslant 0}$, satisfying (1.11), can be constructed. Or, rather, how two stochastic processes, $(\widehat{X}_n)_{n\geqslant 0}$ and $(T_n)_{n\geqslant 0}$, can be used to construct a continuous-time process $(\widehat{X}(t))_{t\geqslant 0}$ with the same intensity matrix as $(X(t))_{t\geqslant 0}$.

Let $(\widehat{X}_n)_{n\geqslant 0}$ be a discrete-time homogeneous Markov chain on E with transition probabilities given by the matrix $\widehat{P} = [\widehat{p}_{xy}]_{x,y\in E}$. Further, let $\widehat{N}$ be a one-dimensional homogeneous Poisson process with intensity λ and enumerate the points of $\widehat{N}$ from (1.3) as $(T_n)_{n\geqslant 0}$, such that $T_0 = 0, T_n < T_{n+1}$, for $n \geqslant 0$. Suppose that $(\widehat{X}_n)_{n\geqslant 0}$ and $\widehat{N}$ are independent, and define the continuous-time process on E

$$\widehat{X}(t) = \widehat{X}_{\widehat{N}(t)}, \quad t \geqslant 0. \tag{1.13}$$

Then $(\widehat{X}(t))_{t\geqslant 0}$ is a Markov chain. The Poisson process $\widehat{N}$ is referred to as the *clock*, and the discrete-time chain $(\widehat{X}_n)_{n\geqslant 0}$ as the *subordinate* Markov chain of $(\widehat{X}(t))_{t\geqslant 0}$.

As a first observation, $\widehat{X}(T_n) = \widehat{X}_n$ for all $n \geqslant 0$. Suppose that $\widehat{X}(0) = \widehat{X}(T_0) = \widehat{X}_0 = x$. The transition probabilities of $\widehat{X}(t)$ are

$$\begin{aligned}\widehat{p}_{xy}(t) &= \mathsf{P}(\widehat{X}(t) = y) = \mathsf{P}(\widehat{X}_{\widehat{N}(t)} = y) = \sum_{n=0}^{\infty} \mathsf{P}(\widehat{X}_n = y, \widehat{N}(t) = n) \\ &= \sum_{n=0}^{\infty} \mathsf{P}(\widehat{X}_n = y)\mathsf{P}(\widehat{N}(t) = n) = \sum_{n=0}^{\infty} \widehat{p}_{xy}(n) \cdot \frac{(\lambda t)^n}{n!} \mathrm{e}^{-\lambda t},\end{aligned}$$

or, expressed more compactly,

$$\widehat{P}_t = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \mathrm{e}^{-\lambda t} \cdot \widehat{P}_n,$$

where $\widehat{P}_n = \widehat{P}^n$ is the n-step transition matrix with elements $[\widehat{p}_{xy}(n)]_{x,y \in E}$. Thus, the continuous-time chain $(\widehat{X}(t))_{t \geqslant 0}$ has transition intensities

$$\begin{aligned}\widehat{q}_{xy} = \lim_{t \to 0} \frac{1}{t} \widehat{p}_{xy}(t) &= \lambda \widehat{p}_{xy}, \quad x, y \in E, x \neq y, \\ \widehat{q}(x) = \lambda(1 - \widehat{p}_{xx}).\end{aligned}$$

Written as matrices, this becomes

$$\widehat{Q} = \lambda(\widehat{P} - I), \tag{1.14}$$

where I is the identity matrix $I = [\delta_{xy}]_{x,y \in E}$.

Suppose Q is the intensity matrix of $(X_t)_{t \geqslant 0}$. By fixing $\lambda \geqslant \sup_{x \in E} q(x)$, and letting the transition probabilities for the discrete chain $(\widehat{X}_n)_{n \geqslant 0}$ for $x, y \in E$ such that $x \neq y$, be,

$$\begin{aligned}\widehat{p}_{xy} &= q_{xy}/\lambda \\ \widehat{p}_{xx} &= 1 - q(x)/\lambda\end{aligned} \quad \text{or} \quad \widehat{P} = Q/\lambda + I, \tag{1.15}$$

it is seen from (1.14) that $\widehat{X}(t)$ has intensity matrix $\widehat{Q} = Q$. Theorem 1.1.5 then states that the constructed process $(\widehat{X}(t))_{t \geqslant 0} \overset{\mathscr{D}}{=} (X(t))_{t \geqslant 0}$. The process $(\widehat{X}(t))_{t \geqslant 0}$ is known as *uniform* and the equivalence between $(\widehat{X}(t))_{t \geqslant 0}$ and $(X(t))_{t \geqslant 0}$ justifies the name *uniformisable* for the latter.

The difference between $(X(t))_{t \geqslant 0}$ and the constructed $(\widehat{X}(t))_{t \geqslant 0}$ is that the latter has *pseudo-transitions*, i.e., transitions with positive intensity which leave the continuous chain in the same state. In other words, the discontinuity epochs of $(\widehat{X}(t))_{t\geqslant 0}$ correspond to events in $\widehat{N}$ (at some times T_n), but not all events of $\widehat{N}$ are discontinuity epochs of $(\widehat{X}(t))_{t\geqslant 0}$. The subordinate chain may have $\widehat{X}_n = \widehat{X}_{n+1}$ for some $n \geqslant 0$.

Next, the embedded chain of $(X(t))_{t\geqslant 0}$ (and the constructed $(\widehat{X}(t))_{t\geqslant 0}$), $(X_n)_{n\geqslant 0}$, is related to the subordinate chain $(\widehat{X}_n)_{n\geqslant 0}$ as follows. Choose $x \in E$ and assume that $q(x) > 0$, so that $\widehat{p}_{xx} < 1$ according to (1.15). Using the subordinate chain $(\widehat{X}_n)_{n\geqslant 0}$, one can see that for $x \neq y$

$$p_{xy} = \mathsf{P}(X_1 = y|X_0 = x) = \sum_{k=0}^{\infty} \widehat{p}_{xx}^k \widehat{p}_{xy} = \frac{\widehat{p}_{xy}}{1-\widehat{p}_{xx}} = \frac{q_{xy}}{q(x)}.$$

The last step follows from (1.15). For $Z_n = \tau_n - \tau_{n-1}$, $n \geqslant 1$, one gets

$$\begin{aligned}\mathsf{P}(Z_1 > t|X_0 = x) &= \sum_{k=0}^{\infty} \mathsf{P}(N(t) = k, \widehat{X}_k = \cdots = \widehat{X}_0|X_0 = x) \\ &= \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!}\mathrm{e}^{-\lambda t} \cdot \widehat{p}_{xx}^k = \mathrm{e}^{-\lambda(1-\widehat{p}_{xx})t} = \mathrm{e}^{-tq(x)}.\end{aligned}$$

Hence $(X_n, Z_n)_{n\geqslant 0}$ satisfies (1.11).

Theorem 1.3.1. *Let $(\widehat{X}_n)_{n\geqslant 0}$ be the subordinate chain of a uniformisable ergodic Markov chain $(X_t)_{t\geqslant 0}$ with intensity matrix Q and stationary distribution π. Then π is the stationary distribution of $(\widehat{X}_n)_{n\geqslant 0}$.*

Proof. From Theorem 1.1.4, π satisfies $\pi Q = 0$. Using (1.14), this is equivalent to

$$\pi Q = 0 \quad \Leftrightarrow \quad \pi\lambda(\widehat{P} - I) = 0 \quad \Leftrightarrow \quad \pi\widehat{P} = \pi, \tag{1.16}$$

where $\widehat{P}$ is the transition matrix of $(\widehat{X}_n)_{n\geqslant 0}$. Since $(X_t)_{t\geqslant 0}$ is irreducible and ergodic, so is $(\widehat{X}_n)_{n\geqslant 0}$; the pseudo-transitions make $(\widehat{X}_n)_{n\geqslant 0}$ aperiodic. Thus, π is the stationary distribution of $(\widehat{X}_n)_{n\geqslant 0}$. □

Note that π is clearly *not* the stationary distribution of the embedded chain $(X_n)_{n\geqslant 0}$.

1.4 THE TWO-DIMENSIONAL POISSON EMBEDDING

A uniform Markov process $(X(t))_{t\geqslant 0}$ has been constructed with a prescribed intensity matrix Q, under the assumption that $\sup_{x\in e} q(x) < \infty$.

The construction made use of a clock $\widehat{N}$ and a subordinate discrete-time Markov process $(\widehat{X}_n)_{n\geqslant 0}$ independent of $\widehat{N}$. The next task is to construct $(\widehat{X}_n)_{n\geqslant 0}$ and $\widehat{N}$ in a sophisticated way, so that later (Section 3.1) families of processes $\{(X(t))_{t\geqslant 0}\}$ can be coupled. The intensity of the clock $\widehat{N}$, λ, can be chosen arbitrarily under the condition $\lambda \geqslant \sup_{x\in E} q(x)$. A particular uniformisation procedure is devised, where the subordinate chain and the clock are embedded in a two-dimensional Poisson process.

Choose $\lambda \geqslant \sup_{x\in E} q(x)$ and let N be a two-dimensional Poisson process on $\mathbb{R}_+\times[0,\lambda]$ with the Lebesgue measure as parameter measure Λ. By dividing the region $\mathbb{R}_+\times(0,\lambda]$ into horizontal disjoint strips, several independent homogeneous Poisson processes can be constructed, as noted in Section 1.2.

Let $(I_{x,y})_{y\in E}$ be a partition of the interval $(0,\lambda]$ for every $x\in E$, such that each $I_{x,y}$ has Lebesgue measure $l(I_{x,y}) = q_{xy}$ for $y\neq x$, and $I_{x,x} = \lambda - q(x)$. That is, for $x\in E$,

$$\bigcup_{y\in E} I_{x,y} = (0,\lambda] \quad \text{and} \quad I_{x,y}\cap I_{x,y'} = \varnothing \quad \text{for } y,y'\in E,\, y\neq y'.$$

Only state-combinations $x,y\in E$ such that $q_{xy}>0$ need to be considered. Construct homogeneous Poisson processes N_{xy}, $x,y\in E$, by letting

$$N_{xy}(A) = N(A\times I_{x,y}) \quad \text{and} \quad \widehat{N}(A) = N(A\times[0,\lambda])$$

for every bounded $A\in\mathscr{B}(\mathbb{R}_+)$. From (1.6), for fixed x, the processes $\{N_{xy}\}_{y\in E}$ are independent time homogeneous Poisson processes. The processes N_{xy} and $N_{x'y'}$ need not be independent, since $I_{x,y}$ and $I_{x',y'}$ belong to different partitions; hence $I_{x,y}$ and $I_{x',y'}$ may overlap. Furthermore, $\widehat{N}$ is a homogeneous one-dimensional Poisson process with intensity λ. Enumerate the points of $\widehat{N}$ as $(T_n)_{n\geqslant 0}$ such that $T_0 = 0$ and $T_n < T_{n+1}$, $n\geqslant 0$. For every $x\in E$, let

$$K_x(n+1) = \{y : N_{xy}((T_n, T_{n+1}]) = 1\}, \quad n\geqslant 0.$$

Then $K_x(n)$ is the index of the process responsible for the n^{th} event of $\widehat{N}$ according to the partition $(I_{x,y})_{y\in E}$. Since the intervals $(T_n, T_{n+1}]$, $n\geqslant 0$, are disjoint and N_{xy} are homogeneous, $(K_x(n))_{n\geqslant 1}$ is an i.i.d. sequence for every $x\in E$. According to Theorem 1.2.2, $\widehat{N}$ is independent of the sequences $(K_x(n))_{n\geqslant 1}$, for every $x\in E$. Further, for $x\neq y$

$$\mathsf{P}(K_x(n) = y) = \frac{l(I_{x,y})}{\lambda} = \frac{q_{xy}}{\lambda},$$

and

$$\mathsf{P}(K_x(n) = x) = \frac{l(I_{x,x})}{\lambda} = \frac{\lambda - q(x)}{\lambda}.$$

Let $\widehat{X}_0 = X(0)$, and

$$\widehat{X}_{n+1} = K_{X_n}(n+1) \quad \text{for } n \geqslant 0,$$

Then $(\widehat{X}_n)_{n\geqslant 0}$ is a homogeneous Markov chain with transition probabilities $\widehat{p}_{xy} = q_{xy}/\lambda$, $x, y \in E$ for $x \neq y$, and $\widehat{p}_{xx} = 1 - q(x)/\lambda$. Hence, from (1.15), the generated sequence $(\widehat{X}_n)_{n\geqslant 0}$ constitutes a subordinate Markov chain which together with $(T_n)_{n\geqslant 0}$, constructs $(X(t))_{t\geqslant 0}$ as

$$X(t) = \widehat{X}_n \quad \text{for } t \in [T_n, T_{n+1}),$$

which has intensity matrix Q.

A two-dimensional process, N, is easily constructed. Let the time increments Z_n for $n \geqslant 1$, be independent and exponentially distributed with expected value $1/\lambda$, and let $(U_n)_{n\geqslant 1}$ be an i.i.d. sequence, uniform on $[0, \lambda]$, independent of $(Z_n)_{n\geqslant 0}$. Let $T_0 = 0, T_n = T_{n-1} + Z_n$ for $n \geqslant 1$, which makes $(T_n, U_n)_{n\geqslant 1}$ determine the points of N, and translated to $(\widehat{X}_n, T_n)_{n\geqslant 0}$, the sequence $(U_n)_{n\geqslant 1}$ determines the the jumps of $(\widehat{X}_n)_{n\geqslant 1}$. If $\widehat{X}_{n-1} = x$ and U_n falls in the interval corresponding to $I_{x,y}$, then $\widehat{X}_n = y$, that is,

$$\widehat{X}_{n+1} = \{y \in E : U_{n+1} \in I_{\widehat{X}_n, y}\} = \phi(\widehat{X}_n, U_{n+1}).$$

A *transition rule* is defined as the deterministic function $\phi : E \times [0, \lambda] \to E$ which maps $(\widehat{X}_n, U_{n+1})$ on $\widehat{X}_{n+1}$, $n \geqslant 0$, where $(U_n)_{n\geqslant 1}$ is an i.i.d. sequence of random variables uniform on $[0, \lambda]$. Specifying the transition rule is equivalent to specifying the partitions $(I_{x,y})_{y\in E}$ for all $x \in E$.

The constant λ is chosen such that $\lambda \geqslant \sup_{x\in E} q(x) < \infty$. When $\lambda = \sup_{x\in E} q(x)$, the uniformisation procedure is minimal, in the sense that the number of pseudo-transitions, $\sum_{x\in E} N_{x,x}$, is as small as possible. When E is finite, it is possible to let $\lambda = \sum_{x\in E} q(x) < \infty$ enabling a construction of partitions such that the $N_{x,y}$ all are independent. In Section 3.1 a coupling of a family of processes $(X(t))_{t\geqslant 0}$ is constructed by choosing a suitable family of partitions, and a λ satisfying

$$\sup_{x\in E} q(x) < \lambda < \sum_{x\in E} q(x).$$

1.5 ORDERINGS AND COUPLING THEORY

The results presented now are crucial to the analysis in subsequent chapters. To compare queueing networks with each other, and compare different states of a single queueing network, an ordering of the state spaces needs to be imposed. For a more detailed treatment of orderings and their properties, see Gotlieb and Gotlieb [9]. For coupling theory and stochastic domination, see Lindvall [11] and Thorisson [26]. Shaked and Shanthikumar [22] is a reference on stochastic orderings which has a chapter on stochastic comparisons of networks.

Definition 1.5.1 (Binary relations). A binary relation R on a space E is a non-empty subset of the product space $E \times E$, thus $R \subset E \times E$. Elements $i, j \in E$ such that $(i, j) \in R$ are said to be related by R.

Definition 1.5.2. Consider the space $E = \mathbb{Z}_+^d$. The standard partial ordering on E, is defined by

$$(i, j) \in R \text{ if } i_k \leqslant j_k, \text{ for all } k = 1, \ldots, d. \tag{1.17}$$

That $(i, j) \in R$ is written $i \preceq j$.

It is convenient to write $i \prec j$ when $i \preceq j$ and $i \neq j$. The term partial emphasises that the space may contain elements i and j which cannot be compared, i.e., are not related to each other.

Note that the partial ordering $\preceq$ causes the state space to have a unique minimal element $\mathbf{0} = (0, \ldots, 0)$. For $E = \mathbb{Z}_+^{\boldsymbol{b}}$ it is also true that

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

A stochastic ordering is now introduced. Consider a space E and a partial ordering, $\preceq$, of this space. Let $f : E \to \mathbb{R}$ be a bounded and non-decreasing function on E, that is, $x \preceq x'$ implies $f(x) \leqslant f(x')$.

Definition 1.5.3. Let P and P' be two probability measures on E. One says that P is *stochastically dominated* by P' if

$$\int_E f \, dP \leqslant \int_E f \, dP', \tag{1.18}$$

for all bounded and non-decreasing functions f. This is written $P \overset{\mathscr{D}}{\preceq} P'$.

When the probability measures are laws of random variables, the variables are said to dominate each other. To be precise, let X and X' be two random variables on E. Then X is stochastically dominated by X', written $X \overset{\mathscr{D}}{\preceq} X'$, when $P \overset{\mathscr{D}}{\preceq} P'$, $X \overset{\mathscr{D}}{=} P$ and $X' \overset{\mathscr{D}}{=} P'$. Definition 1.5.3 is equivalent to the following, which is stated as a theorem.

Theorem 1.5.1. *Let X and X' be two random variables on E. Then X is stochastically dominated by X' if, and only if,*

$$\mathsf{P}\left(X \preceq x\right) \geqslant \mathsf{P}\left(X' \preceq x\right), \textit{ for all } x \in E.$$

Definition 1.5.4. Let P and P' be two probability measures on a measurable space $(E, \mathscr{E})$. A *coupling* (of P and P') is a joint probability measure $\widehat{P}$ on $E \times E$ (equipped with the product σ-algebra $\sigma(E \times E)$) such that

$$\widehat{P}(A \times E) = P(A) \qquad \widehat{P}(E \times A) = P'(A)$$

for every A in $\mathscr{E}$.

There is a duality between a coupling of two random variables and stochastic domination. Suppose that X and X' are two random variables on E, coupled so that $X \preceq X'$ almost surely. Let f be a bounded non-decreasing function on E. Then

$$\int_E f\,dP = \mathsf{E}[f(X)] = \mathsf{E}[f(\widehat{X})] \leqslant \mathsf{E}[f(\widehat{X}')] = \mathsf{E}[f(X')] = \int_E f\,dP',$$

which shows $P \overset{\mathscr{D}}{\preceq} P'$, a stochastic dominance between the laws of X and X'. In fact, the wording of *Strassen's theorem*, see Lindvall [13] or Lyons and Peres [15], shows an *equivalence* between the existence of two probability measures P and P' on E, such that P is dominated by P' with respect to an ordering, $\preceq$, and the existence of a probability measure $\widehat{P}$ on $E \times E$ with P and P' as marginals and $\widehat{P}(\{(x, y) \in E \times E : x \preceq y\}) = 1$.

The main concern is construction of couplings of entire processes. Let $\preceq$ be a partial ordering of the state space E of two independent Markov chains $X = (X_t)_{t \geqslant 0}$ and $X' = (X'_t)_{t \geqslant 0}$; typically, the processes describe queueing networks. Subsequently, couplings $(\widehat{X}, \widehat{X}')$, of X and X', such that

$$\widehat{X}_0 \preceq \widehat{X}'_0 \quad \Rightarrow \quad \widehat{X}_t \preceq \widehat{X}'_t, \ \forall t \geqslant 0 \tag{1.19}$$

are constructed; specifically for network processes $X \overset{\mathscr{D}}{=} X'$.

There is a duality between the dominance relation (1.19) and a monotonic property.

Lemma 1.5.2. *Suppose that there is a coupling of two copies of a Markov process $(X_t)_{t\geqslant 0}$ on $\mathbb{Z}_+^{\boldsymbol{b}}$ such that* (1.19) *holds. Then $X_0 = \mathbf{0}$ implies*

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

A network process satisfying the conditions in Lemma 1.5.2 is said to be *stochastically increasing* in t.

Proof. Fix $s > 0$ and let $\widehat{X}_0 = \mathbf{0}$ and $\widehat{X}'_0 = X_s$. Then $\widehat{X}_0 \preceq \widehat{X}'_0$, and the coupling of the processes yields $\widehat{X}_t \preceq \widehat{X}'_t$. But $X_t \overset{\mathscr{D}}{=} \widehat{X}_t \preceq \widehat{X}'_t \overset{\mathscr{D}}{=} X_{t+s}$, which shows that $(X_t)_{t\geqslant 0}$ is stochastically increasing in t. □

Suppose that a family of processes $\{(X_t^\theta)_{t\geqslant 0}\}_{\theta\in\Theta}$ have probability laws governed by some parameter set Θ. To establish a concept of monotonicity in the parameters of the processes, consider the following.

Definition 1.5.5. A family of processes is said to be stochastically increasing in θ if there exists a family of processes $\{(\widehat{X}_t^\theta)_{t\geqslant 0}\}_{\theta\in\Theta}$ such that $\widehat{X}^\theta \overset{\mathscr{D}}{=} X^\theta$ and, for all $\theta \leqslant \theta'$, $\widehat{X}_t^\theta \preceq \widehat{X}_t^{\theta'}$ holds for all $t \geqslant 0$ [a.s.].

To establish stochastic monotonicity in a parameter, it suffices to construct a coupling $(\widehat{X}, \widehat{X}')$ satisfying (1.19), for processes $\widehat{X} \overset{\mathscr{D}}{=} X^\theta$ and $\widehat{X}' \overset{\mathscr{D}}{=} X^{\theta'}$ such that $\widehat{X}_0 = \widehat{X}'_0$, for all values $\theta \leqslant \theta'$.

Chapter 2

Queueing networks

2.1 INTRODUCTION TO QUEUEING NETWORKS

There is an immense literature on queueing. 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 units, are accepted only if sufficient resources are available to deal with them. The resources are modelled as queues with finite waiting spaces; each queue is served by multiple servers emptying the queue and thus freeing resources. Requests are visualised as units arriving at a queue, and they are accepted if the waiting space has vacant positions. As a model of a computer network, a node may be a computer or a switch, while a unit is a data packet, request or message. As a model of manufacturing lines, nodes could be work stations and units the objects produced. In this thesis, the words 'units' and 'nodes' are used regardless of the particular interpretation in the model under study. For a brief overview of queueing networks, consult Søren Asmussen's contribution to Barndorff-Nielsen *et al.* [3]. A detailed treatment of the topic can be found in Wolff [30] and Serfozo [21].

The objective is to compute quantities, such as the probability that an arriving unit is rejected, queue lengths, service times, average waiting 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 represented by units using the resources have other measures of quality of service.

Much research is directed to queueing networks that have 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.* [20]. Blocking occurs when a node does not have enough free resources to handle a request or an arriving unit; several blocking mechanisms have been suggested to model different kinds of realistic behaviour. The blocking schemes range from expunging blocked units from the system, to letting them wait at different nodes or placing them in a waiting orbit. Closed form expressions for the stationary distribution of units in queueing networks with blocking are available in only a limited number of models, and huge effort has been expended on approximative algorithms for calculating it and estimates of the errors in the calculations.

This chapter first defines the class of queueing networks under consideration and, after the basic definitions, a framework of couplings is set up to analyse the networks, as well as to simulate them to obtain samples from their stationary distributions (Chapter 3).

2.2 JACKSON NETWORKS

One distinguishes between open and closed networks. When every node of the network is directly or indirectly subject to arrivals of units from the outside, and all units eventually leave the network, the network is called *open*. As opposed to open, in a *closed* queueing network the total number of units in the system is constant. Note also that some authors define open to mean 'not closed' networks. Networks which are neither closed nor open are called *mixed* networks. Closed and mixed networks will not be covered in the thesis.

Definition 2.2.1. Let $d \in \mathbb{N}$ be the number of nodes in the network. A Jackson network is modelled as a multivariate birth and death process $(X_t)_{t \geqslant 0}$ on $E = \mathbb{Z}_+^d$, with transition intensities when $X_t = x \in E$,

$$\begin{aligned} \beta_k(x) &= \beta_k, \\ \delta_k(x) &= \delta_k(x_k) = \mu_k(x_k) p_k, \\ \gamma_{km}(x) &= \mu_k(x_k) p_{km}, \end{aligned} \tag{2.1}$$

for $k, m \in E$.

It is convenient to denote e.g. $\delta_k(x)$ by $\delta_k(x_k)$ to emphasise that it depends only on x_k. Similarly, $\beta_k(x)$ is constant and thus denoted β_k. It is natural to assume that the service intensity $\mu_k(x_k)$ increases in x_k, that is, it can for all k be expressed as

$$\mu_k(n) = \sum_{m=1}^{n} \nu_k(m), \quad \text{where } \nu_k(m) \geqslant 0 \text{ for } m \geqslant 1. \tag{2.2}$$

The *routing probabilities* p_{km} are the probabilities that a unit departing from node k goes to node m, and

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

is the probability that a unit departing from node k leaves the network. The interpretation of the intensities in (2.1) is that arrivals to node k form a time homogeneous Poisson process of rate β_k, and the service times of a unit are exponentially distributed.

Example 2.2.1. Consider a single node Jackson network with a constant service rate $\mu_1(x) = \mu_1 \mathbf{1}_{\{x_1>0\}}$ and $p_1 = 1$. In queueing theory contexts, this is simply the fundamental model of a queue, usually denoted $M/M/1$. Letting $\mu_1(x) = \sum_{m=1}^{x_1} \nu_1(m)$, one obtains the $M/M/k$ queue, where $\nu_1(m) > 0$, $m = 1, \ldots, k$, are the same constant, and $\nu_1(m) = 0$ for $m > k$. It is a well-known result that the departure process from a stationary $M/M/k$ node is a Poisson process.

Since only open networks are dealt with, restrictions must be laid on the service intensities and routing probabilities.

Definition 2.2.2. A Jackson network is *open* if 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

  $$\beta_k \cdot p_{kk_1} \cdot p_{k_1k_2} \cdots p_{k_lm} > 0,$$

  which means that every node m is supplied with units from the outside, either directly or indirectly; and

- For all $k \in \{1, \dots, d\}$ there exist $m, k_1, \dots, k_l \in \{1, \dots, d\}$ such that

$$p_{kk_1} \cdot p_{k_1k_2} \cdots p_{k_lm} \cdot \delta_m(e_m) > 0,$$

  which means that every node k has a way of expunging units from the network, either directly or indirectly.

The idea is simply that every unit arriving to the network will eventually leave. Units may cycle back to the same node when p_{kk} is positive, but for the network to be open, it is necessary that $p_{kk} < 1$. When the routing probabilities are such that no unit is allowed to visit the same node more than once, the network is a *feed-forward* network. In fact, any stationary feed-forward open Jackson network has independent Poisson arrival processes to the nodes.

Before the stationary distribution is stated, the following theorem from Walrand [27] regarding open Jackson networks, is observed.

Theorem 2.2.1. *A Jackson network on $\mathbb{Z}_+^d$ is irreducible if, and only if, it is open.*

Let $(\Lambda_k)_{k=1}^d$ be a non-negative solution to the following linear system of equations:

$$\Lambda_k = \beta_k + \sum_{m=1}^{d} \Lambda_m p_{mk}, \tag{2.3}$$

for $k = 1, \dots, d$. Define

$$\kappa_k = 1 + \sum_{n=1}^{\infty} \frac{\Lambda_k^n}{\mu_k(1) \cdots \mu_k(n)} \tag{2.4}$$

assuming that the sum converges for all $k = 1, \dots, d$. Equation (2.3) is sometimes referred to as the detailed balance equation. In the solution, Λ_k is the rate at which units leave node k, and it should balance the rate at which they enter node k.

The next remarkable theorem concerns the stationary distribution π of a Jackson network. For a proof, see Wolff [30].

Theorem 2.2.2 (Jackson). *If the sequence $(\Lambda_k)_{k=1}^d$ is a non-negative solution to* (2.3) *where κ_k from* (2.4) *satisfies $\kappa_k < \infty$, for all $k = 1, \dots, d$, then*

$$\pi(x) = \pi_1(x_1) \cdots \pi_d(x_d), \tag{2.5}$$

where

$$\pi_k(n) = \frac{1}{\kappa_k} \frac{\Lambda_k^n}{\mu_k(1) \cdots \mu_k(n)}, \quad n \in \mathbb{Z}_+ \tag{2.6}$$

for $k = 1, \ldots, d$, is the stationary distribution of an open Jackson network with parameters β_k, μ_k, and p_{km}, $k, m = 1, \ldots, d$. For $n = 0$ the product in the denominator is taken to be one.

Definition 2.2.3. When $\pi(x) = \pi_1(x_1) \cdots \pi_d(x_d)$, the stationary distribution is said to be in a *product form*. A network of queues with such a distribution is said to be a *product-form network*.

A remarkable feature of stationary open Jackson networks is that, at any given time t, $X_k(t)$ is independent of $X_m(t)$, $m \neq k$, due to the product form of π. From Definition 2.2.3 it might seem that all stationary product-form networks have this property. However, these distributions are often restrictions of a product-form distribution to a smaller state space, and dependencies may arise from this restriction. Take, for example, the closed Jackson network which also has a product-form stationary distribution; clearly the number of units at different nodes are dependent, since the sum of all units is constant.

In a stationary Jackson network the fraction of time during which queue k contains x_k units is $\pi_k(x_k)$. The fraction of *arrivals* to queue k who find x_k units there is also $\pi_k(x_k)$. For Poisson arrivals, this feature is known as the PASTA[1] property of a Poisson process, nevertheless it is also valid for a Jackson network, in equilibrium, where the net-arrival processes are *not* Poisson. This result is called the Arrival theorem for Jackson networks.

2.3 BOUNDED NETWORKS, SBE NETWORKS

2.3.1 INTRODUCTION

The network models devised here are similar in structure to the Jackson networks, but contain a lot more parameters. To introduce these parameters smoothly, it is first stated why and how the Jackson network needs to be generalised.

First, a bound is put on the number of units at each node. Typically, such a restriction is thought to be a more realistic model for an existing network

[1] Poisson Arrivals See Time Averages; see Wolff [30].

where resources are limited. Thus, buffer capacities at the nodes, need to be specified.

Since the nodes have a finite capacity, a blocking of units is introduced. When the buffer at a node is full, no more units can be allowed, either from the outside or from other parts of the network, until a unit has departed. Units arriving at a node with a full buffer are blocked, and the behaviour of blocked units must be specified. If they are in transit from another node, they may either remain at the originating node and try again at a later time, or they may be rerouted to another node, or possibly leave the network entirely. The behaviour is determined by a blocking mechanism which is modelled by assigning probabilities to alternatives.

In a Jackson network, units arrive one at a time, but the new networks must handle batch arrivals, i.e., when several units arrive at the same instant. One may think of the instant of a batch arrival as the arrival of a bus, carrying a fixed or random number of units. For random batches a probability distribution of the batch sizes must be specified; this may vary for different nodes of the network. Since there is a blocking network, one must also decide how to handle batch arrivals that have more units than will fit entirely.

The arrival processes to the nodes are not, as in the ordinary Jackson case, a time homogeneous Poisson process; they depend on a random environment. Actually, the arrival and service intensities at a node are functions of a Poisson environment. Essentially, the aim is to model phenomena such as server crashes, i.e., when a service facility of a node becomes inoperative for a period of time, or MMPP arrival processes.

2.3.2 DEFINITIONS

Let $Z = (Z_1, \dots, Z_r)$ be a Poisson environment of dimension $r \in \mathbb{N}$. Let $d \in \mathbb{N}$ be a dimension and $\boldsymbol{b} \in \mathbb{N}^d$ be a vector of bounds. The evolution of the network is governed by the Markov chain $(X, Z) = (X_t, Z_t)_{t \geqslant 0}$, where X is a process defined on $E = \mathbb{Z}_+^{\boldsymbol{b}}$.

Definition 2.3.1. A general simple bounded network in a random environment is modelled as the multivariate birth and death process $(X_t, Z_t)_{t \geqslant 0}$

with transition intensities for $(x, z) \in E \times \{0,1\}^r$ that satisfy:

$$\begin{aligned}
\beta_k(x, z) &= f_k(z)\beta_k(x)\mathbf{1}_{\{x_k < b_k\}},\\
\delta_k(x, z) &= g_k(z)\delta_k(x)\mathbf{1}_{\{x_k > 0\}},\\
\gamma_{km}(x, z) &= g_k(z)\gamma_{km}(x)\mathbf{1}_{\{x_k > 0\}}\mathbf{1}_{\{x_m < b_m\}},\\
\beta_{d+l}(x, z) &= \sigma_l(z),\\
\delta_{d+l}(x, z) &= \nu_l(z),
\end{aligned} \tag{2.7}$$

for $k, m \in \{1, \dots, d\}$ and $l = 1, \dots, r$. All of the functions, f_k, g_k, β_k, δ_k, and γ_{km}, are assumed to be non-negative and bounded. The functions ν_l and σ_l are those mentioned, after Definition 1.2.3, as belonging to the Poisson environment. Networks with intensities given by (2.7) are abbreviated GSBE networks, which is short for General Simple Bounded networks in a random Environment. The word simple denotes that batch arrivals are excluded.

It should be clear from (2.7) that the component $(Z_t)_{t \geqslant 0}$ constitutes a Poisson environment as in Definition 1.2.3. Also, $(Z_t)_{t \geqslant 0}$ is a random environment process of $(X_t, Z_t)_{t \geqslant 0}$.

Without loss of generality, the functions f_k and g_k are assumed to take values on the unit interval, i.e.,

$$f_k, g_k : \{0,1\}^r \to [0,1], \quad k = 1, \dots, r.$$

This is not a restriction, since a rescaling of f_k, for example, can be compensated by a suitable rescaling of β_k.

Although the network is described as a multivariate birth and death process, the interpretations of the first d components differ from the r last ones. The reader is encouraged to regard the first part, items one to three of the definition as a conditional intensity matrix Q_z of X when $Z_t = z$. In other words, regard it as if the environment process Z were specified, and then state the conditional intensity matrix of X, generated by Z. As seen from a networking perspective, the transition intensities are time-dependent; they are not time-dependent if, and only if, the functions of the environment, $f(Z_t)$ and $g(Z_t)$, are almost surely constant.

Example 2.3.1. Fix $d \in \mathbb{N}$ and let $\boldsymbol{b} \in \mathbb{N}^d$. The simplest translation of a d-node Jackson network to a simple bounded network is obtained by disregarding the state of the environment, i.e.,

$$f_k \equiv 1 \quad g_k \equiv 1 \qquad k = 1, \dots, d,$$

and letting

$$\beta_k(x) = \beta_k \quad \delta_k(x) = \mu_k(x_k)p_k \quad \gamma_{km}(x) = \mu_k(x_k)p_{km},$$

for $k, m \in \{1, \dots, d\}$, $k \neq m$. Compare these intensities to the ones in (2.1). Since f_k and g_k are constant, the evolution of Z_t does not affect the number of units in the network, X_t, thus $(Z_t)_{t \geqslant 0}$ can be disregarded entirely. The parameters β_k, μ_k, p_k and p_{km} are those of the Jackson network. It should be clear that (2.7) corresponds to a Jackson network restricted to the state space $\mathbb{Z}_+^{\boldsymbol{b}}$.

The restriction operates as a blocking mechanism which can be interpreted as units that fail in transferring from one node to another, remain at the originating node. See comments below.

The use of the functions f_k and g_k is to widen the class of arrival processes and service intensities. An example is given to illustrate this.

Example 2.3.2. Let $d = 1, b \in \mathbb{N}$ and take $r \in \mathbb{N}$. Let ν_k, σ_k be positive constants for $k = 1, \dots, r$ and assume that $f_1(z) = \frac{1}{r}\sum_{k=1}^{r} z_k$. Let the network parameters be $\beta_1(x) = \beta_1$, $\delta_1(x, z) = \delta_1(x) = \mu_1(x_k)$, $p_1 = 1$. Apart from the influence of f_1, this would be an $M/M/k$ queue restricted to $\{0, \dots, b\}$, in line with Examples 2.2.1 and 2.3.1. Such nodes are usually denoted $M/M/k/b$, where b is the bound on the number of units that may be present.

Here, $f_1(z)$ is the proportion of 'on' processes in the environment. The intensity of arrivals to the queue, when the environment is in state z, is $\beta_1 \frac{1}{r}\sum_{k=1}^{r} z_k$, which is the intensity of a Markov modulated Poisson process. Thus, the arrival process can be regarded as the superposition of r independent interrupted Poisson processes, with conditional intensities β_1/r.

Letting $f_1(z) = \frac{1}{C}\sum_{k=1}^{r} h_k(z_k)$ where $h_k(1) > 0$, $h_k(0) = 0$, and $C = \sum_{k=1}^{r} h_k(1)$, the same setting as in (1.10) is obtained. The arrival process is a superposition of the interrupted Poisson processes $N_1, \dots, N_r$, with conditional intensities $\beta_1 h_1(z_1)/C, \dots, \beta_1 h_r(z_r)/C$, respectively. As noted in Section 1.2, the superposition of such processes is almost surely simple, so the network is subject to unit arrivals.

> Even though the superposition of the departure process and the process of blocked arrivals from a stationary $M/M/k/b$ node is a Poisson process, neither process, by itself, is Poisson.

A particular class of GSBE networks will now be derived. Fix $d \in \mathbb{N}$, $\boldsymbol{b} \in \mathbb{N}^d$ and let $E = \mathbb{Z}_+^{\boldsymbol{b}}$. The arrival processes depend on the state of the network through functions $\lambda_k : \{0, \dots, b_k\} \to \mathbb{R}_+$, $k = 1, \dots, d$. It is assumed that for $x \preceq x'$,

$$\lambda_k(x_k) \geqslant \lambda_k(x'_k). \tag{2.8}$$

Each arrival may denote the arrival of a unit or a batch of units. For a node k, let

$$\alpha_k(i) : \mathbb{N} \to [0,1], \qquad \sum_{i \in \mathbb{N}} \alpha_k(i) = 1$$

denote the probability that the batch contains i units. Naturally, $\alpha_k(1) = 1$ is used to model non-random arrivals of single units.

For service rates, g_k is used to model the impact of the environment, and $\mu_k(x) = \mu_k(x_k)$ is the state dependent service rate at node k. In the same way as for Jackson networks, p_{km} denotes the routing probabilities.

Let $a_{km} : \mathbb{Z}_+ \to [0,1]$ be the probabilities for a unit from node k to node m, to be accepted when there are x_m units already present there. Usually, only $a_{km}(x_m) = \mathbf{1}_{\{x_m < b_m\}}$ is considered. To ease the notation, $a_{km}(x_m)$ is written for $a_{km}(x)$. Note that $p_{km} a_{km}(x)$ may be considered as state dependent routing probabilities. Assume that for $x_m \leqslant x'_m$

$$1 \geqslant a_{km}(x) \geqslant a_{km}(x') \geqslant 0 \tag{2.9}$$

and $a_{km}(b_m) = 0$. The *acceptance probabilities*, a_{km}, apply only to units moving within the network. Any rejection of arrivals from the outside should be modelled by an appropriate choice of rates $\lambda_k(x)$.

Further, $c_{km}(x)$ denotes the probability that a blocked k to m unit leaves the network when there are x units at the originating node. For $x \preceq x'$ it is required that

$$1 \geqslant c_{km}(x) \geqslant c_{km}(x') \geqslant 0. \tag{2.10}$$

Thus, a unit leaving node k heading for node m has a probability $1 - a_{km}(x)$ of being rejected. When blocked, it has a probability $c_{km}(x)$ of leaving the network. The probabilities $c_{km}(x)$ are referred to as *desistance* probabilities. When $\sum_{m=1}^{d} c_{km}(x) = 0$, units at that node are called *persistent*.

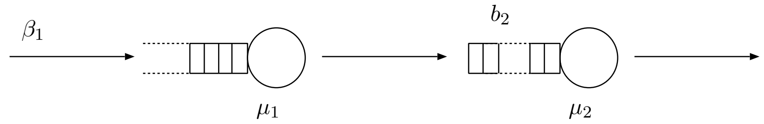

Figure 2.1. *This simple network (tandem queue) illustrates how blocking of units can occur when a node has a finite waiting space.*

Definition 2.3.2. An SBE network is a GSBE network with

$$\begin{aligned}\beta_k(x) &= \lambda_k(x)\mathbf{1}_{\{x_k<b_k\}}\\ \delta_k(x) &= \mu_k(x_k)\left[p_k + \sum_{m=1}^{d} p_{km}(1-a_{km}(x_m))c_{km}(x_k)\right]\\ \gamma_{km}(x) &= \mu_k(x_k)p_{km}a_{km}(x_m),\end{aligned} \tag{2.11}$$

for $k, m \in \{1, \ldots, d\}$, where λ_k, a_{km} and c_{km} satisfy (2.8), (2.9) and (2.10), respectively. The service rates μ_k satisfy (2.2).

Let us disregard the environment for the moment and let f_k, $k = 1, \ldots, d$, be constants. The state-dependent factors $a_{km}(x)$ and $c_{km}(x)$ impose a blocking mechanism that works in the following way. Consider the network in Figure 2.1 consisting of two nodes in series. The first node has Poisson arrivals of intensity $\beta_1(x) = \lambda_1$ and one server which works at rate μ_1. The second node has a finite waiting space that can accommodate, at most, b_2 units, and its server works at rate μ_2. The blocking mechanism for the first server works in such way that a unit, which has finished service at the first node, may be looped back to the first server and receive a new independent service time with probability $(1-a_{12}(x))(1-c_{12}(x))$, that is, the probability that it is rejected but does not leave the network. In queueing theory context, this is commonly referred to as *repetitive blocking*. When $c_{12}(x) \equiv 0$, another interpretation is that the server at the first node idles when the second node is full; this is commonly referred to as *communication blocking*.

To generalise, consider a network of d nodes where node k has a finite waiting space of size b_k. A unit that finishes service at node k chooses to transfer to node m with probability p_{km} and leaves the network with probability p_k. For a chosen node m, the unit has probability $(1 - a_{km}(x))(1 - c_{km}(x))$ to remain at node k an exponential amount of time before a new routing attempt is made. By independent trials, it is possible for a unit that tries again, to choose a node other than the original one. This is known as a *repetitive blocking mechanism* with *random destination*, as opposed to a 'fixed destination' which is determined after the first service time and cannot be altered later.

The condition that $a_{km}(x)$ is non-decreasing in x_m imposes the restriction that the acceptance probability is higher when fewer units already occupy a node. A goal for a network optimiser would be to find a suitable choice of acceptance probabilities.

In Mattsson [16] networks with $a_{km}(x) = \mathbf{1}_{\{x_m<b_m\}}$ were considered for the cases $c_{km}(x) \equiv 1$ and $c_{km}(x) \equiv 0$. Networks with $c_{km}(x) = 0$ were called 'Bounded Jackson networks of Type 1', while $c_{km}(x) = 1$ networks were of Type 2.

There is a symmetry between a_{kk} and c_{kk}, which can be seen in the definition. The routing works in such a way that a unit decides where to go *before* departing from the present node. Suppose that $p_{kk} > 0$ for some $k \in \{1, \ldots, d\}$. Since $a_{kk}(b_k) = 0$, no units will recycle from k to k, even though their own departure would leave room for their admittance. If this is not desirable, $c_{kk} \equiv 0$ can be used. The interpretation is that the unit may not be allowed to transfer to node k, since $a_{kk} < 1$; thus, it remains at node k, since it is not allowed to leave because of an unsuccessful transfer, $c_{kk} = 0$.

The stationary distribution of the number of units at a node k in an SBE network with $c_{km} = 1$ for all $m = 1, \ldots, d$, and a net-arrival process of units which is Poisson, is a truncated geometric distribution. The condition on the arrival process means that node k does not have any inflow of units from other parts of the network, not even from node k itself, that is, $p_{mk} = 0$ for all $m = 1, \ldots, d$. This result, which will be used in the simulation study (Chapter 4), is stated as a theorem.

Theorem 2.3.1. *A node k in an SBE network which satisfies $c_{km}(x) = 1$ and $p_{mk} = 0$ for all $m = 1, \ldots, d$, has stationary distribution π_k of the number of units, where*

$$\kappa_k = 1 + \sum_{n=1}^{b_k} \prod_{m=0}^{n-1} \frac{\beta_k(m)}{\mu_k(m+1)} \qquad \pi_k(n) = \frac{1}{\kappa_k} \prod_{m=0}^{n-1} \frac{\beta_k(m)}{\mu_k(m+1)}$$

for $n = 0, \ldots, b_k$.

In particular, with $\beta_k(x) = \beta_k \mathbf{1}_{\{x_k<b_k\}}$, $\mu_k(x) = \mu_k \mathbf{1}_{\{x_k>0\}}$ and $\rho = \beta_k/\mu_k$, π_k is a truncated geometric distribution with parameters ρ and b_k.

Proof. A node k under such conditions can be modelled as a truncated birth and death process on $\{0, \ldots, b_k\}$ by itself, with birth intensity $\beta_k(x)$ and death intensity $\mu_k(x)$. The balance equations (1.2) reduces to the detailed balance equations:

$$\pi_k(n)\beta_k(n) = \pi_k(n+1)\mu_k(n+1), \quad n = 0, \ldots, b_k - 1.$$

Solving for π_k and determining the normalising constant κ_k gives the statement. □

Definition 2.3.3. Fix $d \in \mathbb{N}$ and $H \subseteq \{1, \ldots, d\}$. Let α_k be a probability distribution on $\mathbb{N}$ for every $k \in H$. For $k \notin H$, α_k is degenerated, i.e., $\alpha_k(1) = 1$. Let $\beta_k(x, z)$, $\delta_k(x, z)$ and $\gamma_{km}(x, z)$ be the intensities in (2.7) belonging to an SBE network for which $\lambda_k(x) = \lambda_k$ in (2.11) for every $k \in H$. Define the bounded network with a batch arrivals as the Markov chain $(X_t, Z_t)_{t \geqslant 0}$ having conditional intensity matrices Q_z, $z \in \{0, 1\}^r$, with possibly non-zero elements of the form

$$\begin{aligned} q_{x,x+ie_k} &= \beta_k(x, z)\alpha_k(i), \text{ for } i = 1, 2, \ldots, b_k - x_k - 1, \\ q_{x,x+(b_k - x_k)e_k} &= \beta_k(x, z) \sum_{i=b_k - x_k}^{\infty} \alpha_k(i), \\ q_{x,x-e_k} &= \delta_k(x, z), \text{ and} \\ q_{x,x-e_k+e_m} &= \gamma_{km}(x, z), \end{aligned} \tag{2.12}$$

and $(Z_t)_{t \geqslant 0}$ is a Poisson environment of dimension r.

Note that this network is not described by a multivariate birth and death process, since it does not necessarily have unit increments. The set H is simply the set of nodes where there are batch arrivals. If $H = \varnothing$ then the network with batch arrivals reduces to an SBE network. From (2.12) it is clear that for batch arrivals to node k, as many units are allowed as will fit into the buffer of size b_k.

For SBE networks an explicit solution for the stationary distribution is hard to find since the balance equations are difficult to solve. Networks with finite waiting spaces at the nodes may have product-form stationary distributions if, for example, units that find a node full, immediately skip that node and continue as if they had received service there. Such a network is sometimes said to have a jump-over blocking scheme. According to Boucherie and van Dijk [4], a formal proof of the product form of π can be found in Aarssen and van Dijk [1]. Boucherie and van Dijk [4] includes a survey of situations for which the Arrival theorem is valid for queueing networks with blocking.

2.4 QUEUEING QUANTITIES

Let E be a state space and $(X_t)_{t\geqslant 0}$ a Markov chain on E governed by intensity matrix Q that has off-diagonal elements q_{xy}, $x, y \in E$, $x \neq y$, and diagonal elements $-q(x)$, $x \in E$. Let $A \subset E$ and assume that $(X_t)_{t\geqslant 0}$ is ergodic with stationary distribution π.

Definition 2.4.1. The flow of $(X_t)_{t\geqslant 0}$ into the set A on the time interval $[0, t]$ is

$$N_A(t) = \#\{\tau \in [0, t] : X_{\tau -} \in A^c,\ X_\tau \in A\}.$$

The *rate* at which $(X_t)_{t\geqslant 0}$ enters the set A is defined as

$$\lambda(A) = \lim_{t\to\infty} \frac{1}{t} N_A(t).$$

Note that for ergodic processes the limit above exists almost surely. It also follows that

$$\lambda(A) = \sum_{x\in A^c} \pi(x) \sum_{y\in A} q_{xy}.$$

This expression is used when calculating the throughput in (2.13).

2.4.1 QUEUEING DISCIPLINES

Although the Markov process $(X_t)_{t\geqslant 0}$ describes the evolution of the queueing network, the particular units are indistinguishable, i.e., without further assumptions one cannot tell which unit moves where. Thus, when estimating unit movements from trajectories of stationary processes, additional assumptions must be made as to how transitions should be interpreted. For example, a transition x to $x - e_k + e_m$ can be interpreted as the movement of a unit from node k to node m within the network, but another interpretation is that the departure from node k triggers an immediate arrival of another unit to node m. Hence, when calculating the time a unit spends within the network, the particular interpretation is crucial and must be determined.

Fix node k in an SBE network $X = (X_t)_{t\geqslant 0}$. Suppose that $(v_m)_{m\geqslant 1}$ is a sequence of probability distributions on $\{1, \ldots, m\}$, i.e.,

$$v_m(n) \geqslant 0, \quad \sum_{n=1}^{m} v_m(n) = 1,$$

for all $m \geqslant 1$. By the 'lack of memory' property of the exponential distribution, the service at node k can be regarded as given simultaneously to the x_k units present according to the distribution v_{x_k}. That is, unit n, $1 \leqslant n \leqslant x_k$, at node k receives service intensity $v_{x_k}(n)\mu_k(x_k)$; the service time is exponentially distributed and independent of the service time of the $x_k - 1$ other units at node k. The time until a departure from node k is the minimum of x_k independent exponential random variables, hence it is exponentially distributed with intensity

$$\sum_{n=1}^{x_k} v_{x_k}(n)\mu_k(x_k) = \mu_k(x_k).$$

Conditioned on a departure from node k, the particular unit finishing service at that time is chosen according to the distribution v_{x_k}. Every choice of distributions $(v_n)_{n\geqslant 1}$ yields the same process X, only the interpretations of unit movements differ. This captures the concept of units being indistinguishable. It is for tracking single units that assumptions must be placed on the probability measures $(v_n)_{n\geqslant 1}$. Even $(v_n)_{n\geqslant 1}$ may vary over time. Although the resulting network process $(X_t, (v_n(t))_{n\geqslant 1})_{t\geqslant 0}$ is not a Markov chain, since one may choose sequences of probability measures depending on the history of X, the trajectory of the number of indistinguishable units $(X_t)_{t\geqslant 0}$, is.

Some of the most common sequences of probability distributions $(v_m)_{m\geqslant 1}$ and their interpretations are now presented. When $v_m(1) = 1$, $m \geqslant 1$, the first unit in the queue is always the first to receive service (and to leave the network). This queueing discipline is commonly abbreviated FCFS, or FIFO for first–in, first–out. If $v_m(m) = 1$, $m \geqslant 1$, the last unit in the queue is always the first to receive service, or the node has LIFO, last–in, first–out, discipline. Note that the interpretation is that an arriving unit immediately gets the attention of a server. The name of the queueing discipline refers to the order in which the units leave *the queue*; this, apart from a single server case, could be different from the order in which they leave the node.

Atomic probability measures v_m select the same single unit at each departure instant, and the most natural interpretation is that there is a single server node. If $\mu_k(x_k) = \mu_k \mathbf{1}_{\{x_k>0\}}$, then this single server works at constant rate μ_k when attending units.

When $v_m(n) = 1/m$, for $1 \leqslant n \leqslant m$, $m \in \mathbb{N}$, and $\mu_k(x_k) = \mu_k \mathbf{1}_{\{x_k>0\}}$, one interpretation is that all units at the node receive service simultaneously at the rate μ_k/x_k from a single server. This discipline is known as (egalitarian) processor sharing, commonly abbreviated PS. Here, the particular unit departing is uniformly chosen from those units present at the node.

When the distributions $(v_m)_{m\geqslant 1}$ satisfy

$$v_m(n) = \frac{1}{m}, \quad 1 \leqslant m \leqslant c,\ 1 \leqslant n \leqslant m$$
$$v_m(n) = \frac{1}{c}, \quad c < m,\ 1 \leqslant n \leqslant m,$$

for some constant $c \in \mathbb{N}$, the departing units are chosen uniformly from the first c present. A common service intensity for these measures is

$$\mu_k(x_k) = x_k \mu_k \mathbf{1}_{\{x_k>0\}}, \quad x_k \leqslant c,$$
$$\mu_k(x_k) = c\mu_k \mathbf{1}_{\{x_k>0\}}, \quad x_k > c,$$

and the node can be interpreted as having c independent servers, each serving units at rate μ_k, on a FIFO basis.

2.4.2 NETWORKING QUANTITIES

Let us recall Definition 2.4.1 where 'rate' was defined and also recall the expression for calculating it.

Definition 2.4.2. The *throughput* ρ_k from a node k is the rate at which units leave node k.

One may calculate the throughput as

$$\rho_k = \sum_{x\in E} \pi(x)\mu_k(x). \tag{2.13}$$

or, analogously,

$$\rho_k = \sum_{x\in E} \pi(x)\beta_k(x) \tag{2.14}$$

since the rate at which units enter node k under stationarity is equal to the rate at which units depart.

When $\mu_k(x) = \mu_k \mathbf{1}_{\{x_k>0\}}$ or $\beta_k(x) = \beta_k \mathbf{1}_{\{x_k<b_k\}}$, the throughput is

$$\rho_k = \mu_k(1 - \pi_k(0)) \quad \text{or} \quad \rho_k = \beta_k(1 - \pi_k(b_k)), \tag{2.15}$$

respectively, where π_k is the marginal distribution of π. For a stationary Jackson network with $\mu_k(x) = \mu_k \mathbf{1}_{\{x_k>0\}}$, Equation (2.6) gives $\pi_k(0) = 1 - \Lambda_k/\mu_k$ where Λ_k is a part of the solution to (2.3). Thus, $\rho_k = \Lambda_k$, as expected. Equation (2.15) is used to estimate the throughput from samples on the stationary distribution.

Definition 2.4.3. The *sojourn time* of a unit is defined as the time the unit stays at a particular node on one visit.

Definition 2.4.4. The *waiting time* is the time a unit spends initially queueing, i.e., staying at the node without actually being served for the first time.

For a LIFO queueing discipline, all waiting times are zero, even though the service to units may be interrupted at some points.

When constructing the waiting and sojourn times of units from the trajectory of a network process $(X_t)_{t\geqslant 0}$, one should keep in mind that the queueing discipline is crucial when calculating these quantities.

Let S be the time of the first arrival at node k for a network process $(X_t)_{t\geqslant 0}$, i.e.,

$$S = \inf\{t > 0 : X_k(t) > X_k(t-)\}.$$

The trajectory is followed to the point at which the unit that arrived at time S leaves node k. Define the successive departure times of units from node k after time S as, $T_0 = S$ and

$$T_i = \inf\{t > T_{i-1} : X_k(t-) > X_k(t)\}, \quad i \in \mathbb{N}.$$

Assuming FIFO queueing discipline and a single server node, the time at which the unit arriving at time S starts to be served is $T_{X_k(S)-1}$, hence, the waiting time of the unit is $T_{X_k(S)-1} - S$.

Generally, with c servers attending units at node k on a FIFO basis, the waiting time of a unit is

$$T_{(X_k(S)-c)\wedge 0} - S.$$

Indeed, the number of units at node k at time S is $X_k(S)$, including the one arriving at time S: if $X_k(S) \leqslant c$, the waiting time of the unit is zero. For convenience, let $N = (X_k(S) - c) \wedge 0$ denote the number of units in front of the unit arriving at S.

Consider a node with $v_m(1) = 1$ for $m \geqslant 1$, a single server FIFO setting. The departure time of the unit arriving at time S is $T_{X_k(S)}$, while the sojourn time of the unit is

$$T_{X_k(S)} - S.$$

For c servers, let

$$v_m(n) = \frac{1}{m},\ 1 \leqslant n \leqslant m$$

for $1 \leqslant m \leqslant c$, and $v_m(n) = 1/c$ when $m > c$. Since the distributions are uniform, the index n is suppressed in the following. When the unit enters service, there are $X_k(T_N)$ units present. At time T_{N+1}, the probability that the unit will depart is $v_{X_k(T_{N+1}-)}$. Generally, for $i \geqslant 1$, at each time T_{N+i}, the probability that a given unit will depart is $v_{X_k(T_{N+i}-)}$. Construct

$$M = \inf\{i \geqslant N+1 : U_i \leqslant v_{X_k(T_i-)}\}$$

where $(U_i)_{i\geqslant 1}$ is an i.i.d. sequence of random variables uniform on $[0,1]$. Then M counts the number of potential departure times for the unit until it departs: the departure time is T_M. Thus, the sojourn time at node k of the unit is $T_M - S$.

Little's law relates the expected sojourn time of a unit to the expected number of units present.

Theorem 2.4.1 (Little's law). *Let $X \overset{\mathscr{D}}{=} \pi$ be a sample on the stationary distribution of an ergodic network. Fix k and let S_k be the sojourn time of a unit at node k, and ρ_k the throughput at node k. Then*

$$\mathsf{E}[X_k] = \rho_k\, \mathsf{E}[S_k].$$

Refer to Wolff [30] for a proof. A remarkable thing is that X_k and ρ_k do not concern the actual queueing discipline, consequently, neither does $\mathsf{E}[S_k]$. A similar result relates the average number in queue, i.e., at the node without being served, and the average queueing time.

2.5 PROPERTIES OF SBE NETWORKS

The algorithm for generating perfect samples of the stationary distribution of our network, relies heavily on the domination results shown is this section.

2.5.1 COUPLING OF ENVIRONMENTS

Let Z and Z' be two independent Poisson environments of dimension r and expected on- and off-periods that satisfy

$$\nu_i \geqslant \nu_i' \quad \text{and} \quad \sigma_i \leqslant \sigma_i', \tag{2.16}$$

for $i = 1, \dots, r$. Then Z_t is stochastically dominated by Z_t', or $Z_t \overset{\mathscr{D}}{\preceq} Z_t'$, for all $t \geqslant 0$ when $Z_0 \preceq Z_0'$. This result is immediate from the following theorem.

Theorem 2.5.1. *If Z and Z' are two independent Poisson environments satisfying* (2.16)*, then there exists a coupling $(\widetilde{Z}, \widetilde{Z}')$ of Z and Z', such that $Z \overset{\mathscr{D}}{=} \widetilde{Z}$, $Z' \overset{\mathscr{D}}{=} \widetilde{Z}'$ and $\widetilde{Z}_0 \preceq \widetilde{Z}_0'$ implies $\widetilde{Z}_t \preceq \widetilde{Z}_t'$, for all $t \geqslant 0$.*

Proof. Fix $z, z' \in \{0,1\}^r$ such that $\widetilde{Z}_0 = z \preceq z' = \widetilde{Z}_0'$. The coupling is achieved by letting the bivariate chain $(\widetilde{Z}_t, \widetilde{Z}_t')_{t \geqslant 0}$ have transition intensities from (z, z'),

to	**with intensity**
$(z + e_k, z' + e_k)$	$\sigma_k(z')$
$(z, z' + e_k)$	$\sigma_k'(z') - \sigma_k(z')$
$(z + e_k, z')$	$\sigma_k(z) - \sigma_k(z')$
$(z - e_k, z' - e_k)$	$\nu_k'(z)$
$(z, z' - e_k)$	$\nu_k'(z') - \nu_k'(z)$
$(z - e_k, z')$	$\nu_k(z) - \nu_k'(z)$.

(2.17)

Note that $\sigma_k(z) - \sigma_k(z') > 0$ and $\nu_k'(z') - \nu_k'(z) > 0$ only if $0 = z_k < z_k' = 1$. From (2.17) follows that the marginal transition intensities from state (z, z') are given by

from	**to**	**with intensity**
z	$z + e_k$	σ_k
$z + e_k$	z	ν_k
z'	$z' + e_k$	σ_k'
$z' + e_k$	z'	ν_k',

(2.18)

so, indeed, $\widetilde{Z}_t \stackrel{\mathscr{D}}{=} Z_t$ and $\widetilde{Z}'_t \stackrel{\mathscr{D}}{=} Z'_t$. Further, the possible transitions of $(\widetilde{Z}, \widetilde{Z}')$ by (2.17) ensure that $\widetilde{Z}_0 \preceq \widetilde{Z}'_0$ implies $\widetilde{Z}_t \preceq \widetilde{Z}'_t$ for all $t \geqslant 0$, and the proof is complete. □

A closer look at (2.17) reveals that this is simply a coupling of the pairs $(\widetilde{Z}_k(t), \widetilde{Z}'_k(t))$ for $k = 1, \dots, r$.

An important corollary follows from Theorem 2.5.1.

Corollary 2.5.2. *If Z and Z' are two independent Poisson environments of dimension r, and have the same expected on- and off-periods, i.e.,*

$$\sigma'_k = \sigma_k \quad \textit{and} \quad \nu'_k = \nu_k,$$

for $k = 1, \dots, r$, then there exists a coupling $(\widetilde{Z}, \widetilde{Z}')$ of Z and Z', such that $\widetilde{Z}_0 \preceq \widetilde{Z}'_0$ implies $\widetilde{Z}_t \preceq \widetilde{Z}'_t$, for all $t \geqslant 0$, and $\lim_{t\to\infty} |\widetilde{Z}_t - \widetilde{Z}'_t| = 0$ almost surely.

Proof. What remains to be proved is the convergence. Let

$$T_k = \inf\{t > 0 : \widetilde{Z}'_k(t) - \widetilde{Z}_k(t) = 0\}, \text{ and } T = \max(T_1, \dots, T_r).$$

The random variables T_k are stopping times with respect to $(\widetilde{Z}_k(t), \widetilde{Z}'_k(t))$, where $T_k = \infty$ if, and only if, $\widetilde{Z}'_k(t) > \widetilde{Z}_k(t)$ for all $t \geqslant 0$. Since the processes are independent, the stopping times are also independent. Suppose that $T_k < \infty$, then (2.17) ensures that $\widetilde{Z}'_k(t) = \widetilde{Z}_k(t)$ for all $t \geqslant T_k$. Further, T_k is the minimum of the first down-transition time for $\widetilde{Z}'_k$ and the first up-transition time for $\widetilde{Z}_k$, which are exponentially distributed; thus T_k is independent and exponentially distributed with intensity $\nu_k + \sigma_k$. It is clear that $T_k < \infty$ [a.s.] and, by independence,

$$\mathsf{P}(T > t) = 1 - \prod_{k=1}^{r} \left(1 - \mathrm{e}^{-t(\nu_k + \sigma_k)}\right)$$

which goes to 0 as $t \to \infty$. □

2.5.2 COMPARISONS OF NETWORKS

The couplings of SBE networks now to be established, lead to several inequalities for the network models and prove some properties of the networks. The inequalities will also constitute a foundation for the perfect sampling algorithm in the next chapter.

Let us say that $(X_t, Z_t) \preceq (X'_t, Z'_t)$ if $Z_t \preceq Z'_t$ and $X_t \preceq X'_t$. This is simply the partial ordering $\preceq$ on $E \times \{0,1\}^r$. Let $(X_t, Z_t)_{t \geqslant 0}$ and (X'_t, Z'_t) be two independent SBE networks with the same set of parameters except for the arrival rates, which, to some nodes in (X'_t, Z'_t), are higher. By denoting the parameters in the network described by (X'_t, Z'_t) with primed symbols, the assumption can be formulated as

$$\beta_k(x) \leqslant \beta'_k(x)$$

for all $x \in E$, and $k \in \{1, \dots, d\}$. Then, for $x \preceq x'$,

$$\begin{aligned} \beta_k(x,z) &= f_k(z)\beta_k(x)\mathbf{1}_{\{x_k < b_k\}} \\ &\geqslant f_k(z)\beta_k(x')\mathbf{1}_{\{x'_k < b_k\}} = \beta_k(x', z) \\ \beta_k(x', z) &\leqslant f_k(z)\beta'_k(x')\mathbf{1}_{\{x'_k < b_k\}} = \beta'_k(x', z) \end{aligned}$$

and

$$\delta_k(x,z) = \delta'_k(x,z) \quad \gamma_{km}(x,z) = \gamma'_{km}(x,z).$$

Since $Z \overset{\mathscr{D}}{=} Z'$, the coupling of Z and Z', $(\widetilde{Z}, \widetilde{Z}')$, is simply $\widetilde{Z} = \widetilde{Z}'$. Furthermore, the coupling of (X, Z) and (X', Z'), $(\widetilde{X}, \widetilde{Z}, \widetilde{X}', \widetilde{Z}')$, has transition intensities from (x, z, x', z) given by (2.19):

to	with intensity	
$(x, z, x' + e_k, z)$	$\beta'_k(x', z) - \beta_k(x', z)$	
$(x + e_k, z, x', z)$	$\beta_k(x, z) - \beta_k(x', z)$	(2.19a)
$(x + e_k, z, x' + e_k, z)$	$\beta_k(x', z)$	
$(x, z, x' - e_k, z)$	$g_k(z)[\mu_k(x') - \mu_k(x)]\left(p_k + \sum_{m=1}^{d} p_{km}(1 - a_{km}(x'))c_{km}(x')\right)$	
$(x - e_k, z, x', z)$	$g_k(z)\mu_k(x) \sum_{m=1}^{d} p_{km}(1 - a_{km}(x))[c_{km}(x) - c_{km}(x')]$	(2.19b)
$(x - e_k, z, x' - e_k, z)$	$g_k(z)\mu_k(x)\left(p_k + \sum_{m=1}^{d} p_{km}(1 - a_{km}(x))c_{km}(x')\right)$	
$(x, z, x' - e_k + e_m, z)$	$g_k(z)[\mu_k(x') - \mu_k(x)]p_{km}a_{km}(x')$	
$(x - e_k + e_m, z, x', z)$	$g_k(z)\mu_k(x)p_{km}[a_{km}(x) - a_{km}(x')](1 - c_{km}(x'))$	(2.19c)
$(x - e_k + e_m, z, x' - e_k, z)$	$g_k(z)\mu_k(x)p_{km}[a_{km}(x) - a_{km}(x')]c_{km}(x')$	
$(x - e_k + e_m, z, x' - e_k + e_m, z)$	$g_k(z)\mu_k(x)p_{km}a_{km}(x')$	

for $k, m \in \{1, \dots, d\}$. The transitions for the environment are given by

$(x, z + e_l, x', z + e_l)$	$\sigma_l \mathbf{1}_{\{z_l = 0\}}$	(2.19d)
$(x, z - e_l, x', z - e_l)$	$\nu_l \mathbf{1}_{\{z_l = 1\}}$	

for $l = 1, \dots, r$.

In the following z is suppressed when stating the transitions.

If $x \preceq x'$, then the transitions to

- $(x + e_k, x' + e_k)$
- $(x - e_k, x' - e_k)$
- $(x, x' + e_k)$
- $(x - e_k, x')$
- $(x - e_k + e_m, x' - e_k + e_m)$

will all preserve the order, for example, $x \preceq x'$ implies $x \preceq x' + e_k$. Further, if $x_k = x'_k$, the transitions to $(x + e_k, x')$ have intensity

$$\beta_k(x, z) - \beta_k(x', z) = f_k(z)[\beta_k(x)\mathbf{1}_{\{x_k<b_k\}} - \beta_k(x')\mathbf{1}_{\{x'_k<b_k\}}] = 0;$$

the transitions are possible only if $x_k < x'_k$, and $x + e_k \preceq x'$ still holds after such a transition.

Similarly, transitions to $(x, x' - e_k)$ occur with intensity

$$g_k(z)[\mu_k(x') - \mu_k(x)]\left(p_k + \sum_{m=1}^{d} p_{km}(1 - a_{km}(x'))c_{km}(x')\right)$$

which is positive only if $x'_k > x_k$, and consequently $x \preceq x' - e_k$. Finally,

$$g_k(z)[\mu_k(x') - \mu_k(x)]p_{km}a_{km}(x') = 0$$

when $x_k = x'_k$, and

$$g_k(z)\mu_k(x)p_{km}[a_{km}(x) - a_{km}(x')]c_{km}(x') = 0,$$
$$g_k(z)\mu_k(x)p_{km}[a_{km}(x) - a_{km}(x')](1 - c_{km}(x')) = 0$$

when $x_m = x'_m$. Thus, if $\widetilde{X}_0 = x \preceq x' = \widetilde{X}'_0$, and $\widetilde{Z} = \widetilde{Z}'$, then $(\widetilde{X}, \widetilde{Z}, \widetilde{X}', \widetilde{Z}')$ is a coupling of (X, Z) and (X', Z'), such that $(\widetilde{X}_t, \widetilde{Z}_t) \preceq (\widetilde{X}'_t, \widetilde{Z}'_t)$ for all $t \geqslant 0$ when $(\widetilde{X}_0, \widetilde{Z}_0) \preceq (\widetilde{X}'_0, \widetilde{Z}'_0)$. Indeed, $\widetilde{X}_t \preceq \widetilde{X}'_t$ for all $t \geqslant 0$.

Hence, from Lemma 1.5.2 and Definition 1.5.5 the following theorem is proved.

Theorem 2.5.3. *The number of units at a node in an SBE network, when $X_0 = \mathbf{0}$, is stochastically increasing in t. Also, the number of units is stochastically increasing in the arrival rate β_k.*

This result applies also to ordinary Jackson networks.

In the same manner, the following may be proved.

Theorem 2.5.4. *SBE networks with higher (lower) desistance probabilities will have stochastically fewer (more) units at each node of the network.*

Proof. Let $c_{km}(x)$ and $c'_{km}(x)$ be two sets of desistance probabilities with $c_{km}(x) \geqslant c'_{km}(x)$. Let $(X_t)_{t \geqslant 0}$ and $(X'_t)_{t \geqslant 0}$ be two independent SBE networks, identical apart from the desistance probabilities which are $c_{km}(x)$ and $c'_{km}(x)$, respectively. Replacing all occurrences of $c_{km}(x')$ with $c'_{km}(x')$ in (2.19b) and (2.19c) gives the correct marginals, and (2.19) is a coupling of $(X_t)_{t \geqslant 0}$ and $(X'_t)_{t \geqslant 0}$.

It is only the second line of (2.19b) that is affected by this change. Since the assumptions ensure that $c_{km}(x) - c'_{km}(x') \geqslant 0$ for all $x \preceq x'$, the expression of the intensity is valid. Furthermore, transitions from (x, x'), $x \preceq x'$, to $(x - e_k, x')$ preserve the ordering. Thus, the coupling $(\widetilde{X}, \widetilde{Z}, \widetilde{X}, \widetilde{Z}')$ ensures that $\widetilde{X}_t \preceq \widetilde{X}'_t$ for all $t \geqslant 0$, whenever $\widetilde{X}_0 \preceq \widetilde{X}'_0$ and $\widetilde{Z} = \widetilde{Z}'$. □

Thus, the number of units at a node in an SBE network is stochastically decreasing in the desistance probabilities c_{km}.

To start smoothly when considering the impact of the environment, suppose that (X, Z) describes an SBE network with g_k constant for every $k \in \{1, \dots, d\}$. Without loss of generality, $g_k \equiv 1$ for all k is assumed. Let (X', Z') be an SBE network independent of (X, Z) with the same parameters as (X, Z), apart from f'_k for some k, where $f_k(z) \leqslant f'_k(z)$ for all $z \in \{0,1\}^r$ is assumed. Then, $x \preceq x'$, $z \preceq z'$ implies

$$\begin{aligned} \beta_k(x, z) &= f_k(z)\beta_k(x)\mathbf{1}_{\{x_k < b_k\}} \\ &\geqslant f_k(z)\beta_k(x')\mathbf{1}_{\{x'_k < b_k\}} = \beta_k(x', z), \\ \beta_k(x', z) &\leqslant f'_k(z)\beta'_k(x')\mathbf{1}_{\{x'_k < b_k\}} = \beta'_k(x', z), \end{aligned}$$

which means that the coupling (2.19) still holds for $(\widetilde{X}, \widetilde{Z}, \widetilde{X}', \widetilde{Z}')$, a coupling of (X, Z) and (X', Z'), where $\widetilde{Z} = \widetilde{Z}'$. Hence, $\widetilde{X}_0 \preceq \widetilde{X}'_0$ implies

$$(\widetilde{X}_t, \widetilde{Z}_t) \preceq (\widetilde{X}'_t, \widetilde{Z}'_t), \quad \text{for all } t \geqslant 0.$$

Note that this is not really a generalisation of the first example, but merely another interpretation. An imaginary application illustrates this.

Example 2.5.1. Consider two telephone switching networks of the same size and capacities. During office hours, one of the networks

> is subject to a higher intensity of calls. Consequently, that network has a larger number of units within it, or formulated differently, more calls in progress. The environment plays the rôle of determining the start and end of office hours, with f_k and f_k as the relative intensities during those time periods.

Suppose now that f_k is non-decreasing in z. Then,

$$\begin{aligned}\beta_k(x,z) &= f_k(z)\beta_k(x)\mathbf{1}_{\{x_k<b_k\}}\\ &\geqslant f_k(z)\beta_k(x')\mathbf{1}_{\{x'_k<b_k\}} = \beta_k(x',z), \text{ and}\\ \beta_k(x',z) &\leqslant f'_k(z')\beta'_k(x')\mathbf{1}_{\{x'_k<b_k\}} = \beta'_k(x',z').\end{aligned}$$

Note that, since $g_k = g'_k = 1$,

$$\delta_k(x,z) = \delta_k(x,z') = \delta'_k(x,z') = \delta'_k(x,z)$$

and similarly for $\gamma_{km}(\,\cdot,\,\cdot\,)$.

Assume that the parameters of Z and Z' satisfy (2.16), and let $(\widetilde{Z},\widetilde{Z}')$ be the coupling of Z and Z' in Theorem 2.5.1. Modify (2.19) in the following way. Let $(\widetilde{X},\widetilde{Z},\widetilde{X}',\widetilde{Z}')$ be the coupling of (X,Z) and (X',Z') with transition intensities as in (2.17), (2.19b) and (2.19c), but where

to	**with intensity**
$(x+e_k,z,x'+e_k,z')$	$\beta_k(x',z)$
$(x,z,x'+e_k,z')$	$\beta'_k(x',z')-\beta_k(x',z)$
$(x+e_k,z,x',z')$	$\beta_k(x,z)-\beta_k(x',z),$

for $k\in\{1,\dots,d\}$, replaces (2.19a). Then $\widetilde{X}_0 \preceq \widetilde{X}'_0$, $\widetilde{Z}_0 \preceq \widetilde{Z}'_0$ implies

$$(\widetilde{X}_t,\widetilde{Z}_t) \preceq (\widetilde{X}'_t,\widetilde{Z}'_t), \quad \text{for all } t\geqslant 0.$$

> *Example* 2.5.2. Consider the setting in Example 2.5.1, but let one of the switching systems have stochastically longer office hours. Then the number of ongoing calls will be larger in that network. As a special case, this holds when $f_k = f'_k$, that is, only the on- and off-intervals of the environments differ between the networks.

There is also have a natural domination result between a class of SBE networks and bounded networks with batch arrivals. Let H denote the set of nodes with batch arrivals. Let (X, Z) describe an SBE network with β_k constant for every $k \in H$. Let (X', Z') be the batch network analog of (X, Z), with batch size distributions α_k, $k = 1, \dots, d$, $\alpha_k(1) = 1$ for $k \notin H$.

Take $\widetilde{X}_0 \preceq \widetilde{X}'_0$, $\widetilde{Z} = \widetilde{Z}'$ and let the transition intensities for $(\widetilde{X}_t, \widetilde{Z}_t, \widetilde{X}'_t, \widetilde{Z}'_t)$ from (x, z, x', z) be given by (2.20), (2.19b), (2.19c) and (2.19d), corresponding to arrivals, departures, inter-network and environmental transitions, respectively.

to	with intensity
$(x + e_k, x' + ie_k)$	$\beta_k(x', z)\alpha_k(i),\ i = 1, \dots, b_k - x'_k - 1$
$(x + e_k,$ $x' + (b_k - x'_k)e_k)$	$\beta_k(x', z) \sum_{i=b_k - x'_k} \alpha_k(i)$
$(x + e_k, x')$	$\beta_k(x, z) - \beta_k(x', z)$

(2.20)

Hence, the coupling now constructed preserves the ordering: $\widetilde{X}_0 \preceq \widetilde{X}'_0$ implies $\widetilde{X}'_t \preceq \widetilde{X}'_t$ for all $t \geqslant 0$, and the following theorem is proved.

Theorem 2.5.5. *For an SBE network (X, Z), let H denote the set of nodes where β_k is constant for $k \in H$. Let (X', Z') be the batch network analog of (X, Z) with arrival size distributions α_k, $k = 1, \dots, d$, $\alpha_k(1) = 1$ for $k \notin H$. Then, $X_0 \preceq X'_0$ implies $X_t \overset{\mathscr{D}}{\preceq} X'_t$ for all $t \geqslant 0$.*

Of course, since the blocking mechanism allows as many arriving units as will fit at every node, two batch networks with

$$\sum_{i=1}^{n} \alpha_k(i) \geqslant \sum_{i=1}^{n} \alpha'_k(i),$$

for every $k \in H$, $n \in \mathbb{N}$, can be coupled in the same manner. An example of the coupling at work can be found in Figure 4.33 in Section 4.1.8.

Such domination results cannot be established for networks that have, for example, different service rates. A faster service rate at node k, for example, leads to fewer units at that particular node. As a consequence, however, more units will transfer to other nodes, according to the routing probabilities p_{km}, hence, there will be more units at the destination nodes. In terms of queue lengths, transitions from (x, x') where $x \preceq x'$ to $(x - e_k + e_m, x')$ are possible albeit $x_m = x'_m$. If $x_m = x'_m$, then $x - e_k + e_m \npreceq x'$.

Example 2.5.3. Consider a tandem network X described by

$$\begin{array}{lll} \beta_1(x) = \beta_1 \mathbf{1}_{\{x_1 \leqslant b_1\}} & \mu_1(x) = 10 \cdot \mathbf{1}_{\{x_1 > 0\}} & p_{12} = 1 \\ \beta_2(x) = \beta_2 \mathbf{1}_{\{x_2 \leqslant b_2\}} & \mu_2(x) = \mathbf{1}_{\{x_2 > 0\}} & p_2 = 1 \end{array}$$

(see the network in Figure 2.1) which has a bound on the number of units at the first node. Let X' be the network with the same parameters as X, apart from $\mu'_1(x) = \mathbf{1}_{\{x_1>0\}}$, that is, the network described by X' has a lower service rate at the first node. If $X(0) = X'(0)$, a coupling of the service completion times, as before, will eventually yield a departure from node 1 in X but not in X', after which $X_2(t) \geqslant X'_2(t)$.

If, on the other hand, $p_{21} = 1$ and $p_1 = 1$, that is, the order of the nodes in the tandem queue is reversed, then a departure from node 1 is a departure from the network, and $X(t) \preceq X'(t)$ for all $t \geqslant 0$, if $X(0) = X'(0)$.

The above example should make it clear that a coupling similar to (2.19) can be constructed for networks with different service rates at terminating nodes, i.e., nodes from which units surely leave the network.

Theorem 2.5.6. *Let $(X_t, Z_t)_{t \geqslant 0}$ be an SBE network as in Definition 2.3.2, and let H be the set of terminating nodes*

$$H = \{k \in \{1, \dots, d\} : p_k = 1\}.$$

Let (X'_t, Z'_t) be an SBE network, independent of (X_t, Z_t), with the same set of parameters, except that the service rates at some nodes in H are higher. Denoting the parameters in the network described by (X'_t, Z'_t), by primed symbols, the above assumption is formulated

$$\mu_k(x) \geqslant \mu'_k(x)$$

for all $x \in E$, and $k \in H$. Then, $X(0) \preceq X'(0)$ implies that $X(t) \overset{\mathcal{D}}{\preceq} X'(t)$ for all $t \geqslant 0$.

Proof. As before, $(\widetilde{Z}, \widetilde{Z})$ is a coupling of $Z \overset{\mathcal{D}}{=} Z'$. Let $(\widetilde{X}, \widetilde{Z}, \widetilde{X}', \widetilde{Z})$ have transition intensities as in (2.19a) for $k = 1, \dots, d$, and as in (2.19b) and (2.19c) for $k \notin H$. For $k \in H$ the transition (departure) intensities from state (x, z, x', z) are

to	**with intensity**
$(x - e_k, z, x' - e_k, z)$	$g_k(z)\mu'_k(x)$,
$(x - e_k, z, x', z)$	$g_k(z)[\mu_k(x) - \mu'_k(x)]$,
$(x, z, x' - e_k, z)$	$g_k(z)[\mu'_k(x') - \mu'_k(x)]$.

Note that $[\mu'_k(x') - \mu'_k(x)] = 0$ if $x'_k = x_k$. Hence, if $\widetilde{X}_0 \preceq \widetilde{X}'_0$ then $\widetilde{X}_t \preceq \widetilde{X}'_t$ for all $t \geqslant 0$. □

That was not very exciting. Let $S : E \to \mathbb{Z}_+$ be the function

$$S(x) = \sum_{k=1}^{d} x_k$$

that counts the number of units in a network. Clearly, if $x \preceq x'$ then $S(x) \leqslant S(x')$ and the previous domination results apply to the total number of units in the network. The following theorem from Lindvall [12] states a domination result on the total number of units in an open Jackson network.

Theorem 2.5.7. *Let X and X' be two independent open Jackson network with the same set of parameters, except that $\mu_k(x) \geqslant \mu'_k(x)$ for some $k \in \{1, \dots, d\}$. Then $X(0) = X'(0)$ implies*

$$S(X_t) \overset{\mathscr{D}}{\leqslant} S(X'_t)$$

for all $t \geqslant 0$.

Proof. Suppose that $X(0) = X'(0)$. Couple X and X' with $(\widetilde{X}, \widetilde{X}')$ by letting $\widetilde{X}$ and $\widetilde{X}'$ have identical arrival processes, and route each arrival through the network in the same way for both processes; the successive visits of an arriving unit to different nodes, in both networks, are described by a routing process on $\{0, \dots, d\}$, where state 0 is the outside of the network. The routing process is a Markov chain with transition matrix P that has elements p_{km} for $k, m \in \{1, \dots, d\}$, $p_{k0} = p_k$ and $p_{00} = 1$. Couple the service times in the networks, Z_{kn} and Z'_{kn}, at the different nodes such that $Z_{kn} \leqslant Z'_{kn}$, $k = 1, \dots, d$, $n \in \mathbb{N}$. Here, Z_{kn} is the service time of the n^{th} unit at node k. Thus, the time an arriving unit spends in the network is shorter in network $\widetilde{X}$ than in network $\widetilde{X}'$, consequently $S(\widetilde{X}_t) \leqslant S(\widetilde{X}'_t)$. Hence, $S(X_t) \overset{\mathscr{D}}{\leqslant} S(X'_t)$ for all $t \geqslant 0$ if $X(0) = X'(0)$. □

In Shanthikumar and Yao [23] a similar proof of the statement can be found, as well as other domination results applicable to both open and closed networks. Unfortunately, the result does not apply to networks with state depending routing probabilities, such as the SBE networks; remember that $p_{km}a_{km}(x)$ can be regarded as a state depending routing probability. The blocking of units makes the statement in general false as the following example illustrates.

Example 2.5.4. Consider the two SBE networks (tandem queues), X and X', with $X_0 = X'_0 = \mathbf{0}$, given by the set of parameters

$$\begin{aligned} \beta_1(x) &= \beta'_1(x) \\ \beta_2(x) &= \beta'_2(x) = 0 \end{aligned} \qquad \begin{aligned} \mu_2(x) &= \mu'_2(x) = 0 \\ \mu_1(x) &= \mu_1 \mathbf{1}_{\{x_1>0\}} \\ \mu'_1(x) &= 0 \end{aligned} \qquad p_{12} = p'_{12} = 1.$$

Since $\mu_2 = \mu'_2 = 0$, the number of units at the second node is the number of units served at the first node, which, in the network X', is zero, since $\mu'_1 = 0$. Now, the number of units in X' at time t, $S(X'_t)$, is $X'_1(t)$, the number of arriving units accepted up to time t, which grows steadily toward b_1. The number of arriving units up to time t in the other network is $S(X_t)$, which grows steadily toward the bound $S(\boldsymbol{b}) > b_1$. Coupling the arrivals and service times gives $S(X'_t) \leqslant S(X_t)$ for all $t \geqslant 0$, and $S(X'_t) < S(X_t)$ for sufficiently large t.

Due to the buffer size limit at the arrival node, there may be arrivals to a network that has fewer units, which have no correspondence in a network with more units.

To summarise, if $\mu_k(x) \geqslant \mu'_k(x)$ for some $k \in \{1, \dots, d\}$ and $X(0) = X'(0)$ are the initial values of two SBE networks, then combining Example 2.5.3, where $S(X_t) \overset{\mathscr{D}}{\leqslant} S(X'_t)$ for all $t \geqslant 0$, with Example 2.5.4, where $S(X_t) \overset{\mathscr{D}}{\geqslant} S(X'_t)$, shows that there is no hope, under such general conditions, of being able to construct a coupling of X and X' such that $S(X_t)$ is dominated by $S(X'_t)$. Theorem 2.5.11, however, shows a domination of departure processes for tandem networks with different service rates.

The next theorem states a domination result in the acceptance probabilities for a particular class of SBE networks.

Theorem 2.5.8. *Let (X, Z) and (X', Z') be two independent d-node SBE networks with state space $E = \mathbb{Z}_+^{\boldsymbol{b}} \times \{0,1\}^r$, identical except for the acceptance* *probabilities which satisfy* $a_{km}(x) \leqslant a'_{km}(x)$ *for all* $x \in E$. *If* $c_{km}(x) \equiv 1$, $Z_0 = Z'_0$ *and* $X_0 \preceq X'_0$ *then* $X_t \overset{\mathscr{D}}{\preceq} X'_t$ *for all* $t \geqslant 0$.

Proof. The coupling $(\widetilde{X}, \widetilde{Z}, \widetilde{X}', \widetilde{Z}')$ of (X, Z) and (X', Z') has $\widetilde{Z} = \widetilde{Z}'$ almost surely. Denote

$$\widehat{a}_{km}(x, x') = a_{km}(x) \wedge a'_{km}(x') \quad \text{and} \quad \check{a}_{km}(x, x') = a_{km}(x) \vee a'_{km}(x').$$

Suppose that $x, x' \in \mathbb{Z}_+^{\boldsymbol{b}}$ are such that $x \preceq x'$. The transition rates of $(\widetilde{X}, \widetilde{Z}, \widetilde{X}', \widetilde{Z}')$ from a state (x, z, x', z), $z \in \{0, 1\}^r$, are given by (2.19a), (2.19d) and

to	**with intensity**
$(x, z, x' - e_k, z)$	$g_k(z)[\mu_k(x') - \mu_k(x)]\Big(p_k + \sum_{m=1}^{d} p_{km}(1 - a'_{km}(x'))\Big)$
$(x, z, x' - e_k + e_m, z)$	$g_k(z)[\mu_k(x') - \mu_k(x)]p_{km}a'_{km}(x')$
$(x - e_k, z, x' - e_k, z)$	$g_k(z)\mu_k(x)\Big(p_k + \sum_{m=1}^{d} p_{km}(1 - \check{a}_{km}(x, x'))\Big)$
$(x - e_k, z, x' - e_k + e_m, z)$	$g_k(z)\mu_k(x)p_{km}[a'_{km}(x') - \widehat{a}_{km}(x, x')]$
$(x - e_k + e_m, z, x' - e_k, z)$	$g_k(z)\mu_k(x)p_{km}[a_{km}(x) - \widehat{a}_{km}(x, x')]$
$(x - e_k + e_m, z, x' - e_k + e_m, z)$	$g_k(z)\mu_k(x)p_{km}\widehat{a}_{km}(x, x')$

These transition intensities do form a coupling of two SBE networks. For example, $\widetilde{X}$ has transition intensity, from state x to $x - e_k$:

$$\begin{aligned} & g_k(z)\mu_k(x)\Big(p_k + \sum_{m=1}^{d} p_{km}(1 - \check{a}_{km}(x, x'))\Big) \\ & + \sum_{m=1}^{d} g_k(z)\mu_k(x)p_{km}[a'_{km}(x') - \widehat{a}_{km}(x, x')] \\ & \qquad = \quad g_k(z)\mu_k(x)\Big(p_k + \sum_{m=1}^{d} p_{km}(1 - a_{km}(x))\Big) = \delta_k(x, z), \end{aligned}$$

since

$$\check{a}_{km}(x,x') + \widehat{a}_{km}(x,x') = a_{km}(x) + a'_{km}(x').$$

The other marginal intensities are similarly shown to be the desired ones. The transitions of $(\widetilde{X}, \widetilde{X}')$ from (x, x') to states $(x-e_k, x'-e_k)$, $(x-e_k, x'-e_k+e_m)$, $(x-e_k+e_m, x'-e_k+e_m)$ and the transitions in (2.19a) and (2.19d), all preserve the order. Only if $x_k < x'_k$ can transitions to $(x, x'-e_k)$ and $(x, x'-e_k+e_m)$ occur, hence $x \preceq x'-e_k$ after such transitions. Furthermore, transitions to $(x-e_k+e_m, x'-e_k)$ have positive intensity only if $a'_{km}(x') < a_{km}(x)$. This happens only if $x'_m > x_m$ since the acceptance probabilities are non-increasing in x_m and $a_{km}(x) \leqslant a'_{km}(x)$. Thus, $x-e_k+e_m \preceq x'-e_k$, as desired.

Consequently, if $x \preceq x'$ and $(\widetilde{X}_0, \widetilde{Z}_0, \widetilde{X}'_0, \widetilde{Z}'_0) = (x, z, x', z)$, then $\widetilde{X}_t \preceq \widetilde{X}'_t$ for all $t \geqslant 0$. □

From Theorem 2.5.8 follows a corollary on domination of trajectories for SBE networks with different buffer sizes. To construct two networks, similar but with different restrictions on the waiting spaces, the acceptance probabilities and arrival rates are modified. For $d \in \mathbb{N}$, let $\boldsymbol{b}, \boldsymbol{b}' \in \mathbb{N}^d$ be two vectors of bounds such that $\boldsymbol{b} \preceq \boldsymbol{b}'$. The acceptance probabilities satisfy

$$a_{km}(x) \leqslant a'_{km}(x)$$

for all $x \in \mathbb{Z}_+^{\boldsymbol{b}}$, and $a_{km}(x) = 0$ whenever $x_m = b_m$. Furthermore, the set of arrival rates

$$\beta_k(x_k) = \lambda_k(x_k)\mathbf{1}_{\{x_k < b_k\}} \qquad \beta'_k(x_k) = \lambda_k(x_k)\mathbf{1}_{\{x_k < b'_k\}}$$

for $k = 1, \dots, d$ and a function $\lambda_k(x_k)$ non-increasing in x_k, $x_k \in \mathbb{Z}_+$, satisfy $\beta_k(x_k) \leqslant \beta'_k(x_k)$ for all $x_k \in \mathbb{Z}_+$.

Corollary 2.5.9. *Let (X, Z) and (X', Z') be two independent d-node SBE networks with buffer restrictions $\boldsymbol{b}, \boldsymbol{b}' \in \mathbb{N}^d$, respectively. Assume that $Z \overset{\mathscr{D}}{=} Z'$, and that the networks have the same set of service rates, the same routing probabilities and that the desistance probabilities are identically 1. Suppose that the arrival rates and acceptance probabilities, $a_{km}(x)$, $a'_{km}(x)$, $\beta_k(x)$ and $\beta'_k(x)$, are defined as above. If $c_{km}(x) \equiv 1$, $Z_0 = Z'_0$ and $X_0 \preceq X'_0$ then $X_t \overset{\mathscr{D}}{\preceq} X'_t$ for all $t \geqslant 0$.*

Note that in Corollary 2.5.9 (Theorem 2.5.8), the process X $(\widetilde{X})$ is defined as a processes on $x \in \mathbb{Z}_+^{\boldsymbol{b}'}$, but has intensities which restrict it to $\mathbb{Z}_+^{\boldsymbol{b}}$. A natural choice of acceptance probabilities are, of course, $a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}}$ and $a'_{km}(x) = \mathbf{1}_{\{x_m < b'_m\}}$.

The wording of Corollary 2.5.9 is that the number of units in a network is stochastically increasing in the buffer sizes if the desistance probabilities are all one and the blocking mechanism allow all units that fit. Corollary 2.5.9, for $a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}}$ and $a'_{km}(x) = \mathbf{1}_{\{x_m < b'_m\}}$, is a special case of Theorem 1 in Sparaggis and Cassandras [24].

There is no domination result for the network processes when $c_{km}(x) \neq 1$ for some $x \in E$ and $k, m \in \{1, \dots, d\}$, but when $c_{km}(x) \equiv 0$, however, there is a domination result for the departure process in a tandem network.

Consider a tandem SBE network with d nodes, i.e., the routing probabilities are $p_{kk+1} = 1$ for $k = 1, \dots, d-1$ and $p_d = 1$. Only the first node is supplied with units from the outside, that is, $\beta_1 > 0$ and $\beta_k(x) = 0$ for $k = 2, \dots, d$. Suppose that all units are persistent; $c_{kk+1} \equiv 0$, and let $(X_t)_{t \geqslant 0}$, $(X'_t)_{t \geqslant 0}$ be the SBE networks, just described, with buffer sizes $\boldsymbol{b}$ and $\boldsymbol{b}'$, respectively, where $\boldsymbol{b} \preceq \boldsymbol{b}'$. Further, let $(D_k(t))_{t \geqslant 0}$, $k = 1, \dots, d$, be the counting processes which count the number of departures up to time t from nodes $k = 1, \dots, d$ in the tandem network, with $D_k(0) = 0$. Define $D_0(t)$ to be the cumulative number of accepted units at the first node. The number of units at node k, $X_k(t)$, satisfies

$$X_k(t) = x_k + D_{k-1}(t) - D_k(t) \quad \text{for } k = 1, \dots, d,\ t \geqslant 0, \tag{2.21}$$

where $x = (x_1, \dots, x_d) \in \mathbb{Z}_+^{\boldsymbol{b}}$ is the initial state $X(0) = x$.

For convenience, let $\gamma_{kk+1}(x) = \mu_k(x) a_{kk+1}(x_{k+1})$, for $k = 1, \dots, d$, where $a_{dd+1} \equiv 1$, and define $\widehat{\gamma}_k(x, x')$ as $\gamma_{kk+1}(x) \wedge \gamma_{kk+1}(x')$.

Recall the assumptions on the arrival rates and acceptance probabilities preceding Corollary 2.5.9 and consider the following coupling of $(X_t)_{t \geqslant 0}$ and $(X'_t)_{t \geqslant 0}$. Let $(\widetilde{X}_t, \widetilde{X}'_t)_{t \geqslant 0}$ have transition intensities from state $(x, x') \in \mathbb{Z}_+^{\boldsymbol{b}} \times \mathbb{Z}_+^{\boldsymbol{b}'}$:

to	with intensity
$(x + e_1, x' + e_1)$	$\beta_1(x) \wedge \beta_1(x')$
$(x + e_1, x')$	$\beta_1(x) - \beta_1(x) \wedge \beta_1(x')$
$(x, x' + e_1)$	$\beta_1(x') - \beta_1(x) \wedge \beta_1(x')$
$(x - e_k, x' - e_k)$	$\widehat{\gamma}_k(x, x')$
$(x - e_k, x')$	$\gamma_{kk+1}(x) - \widehat{\gamma}_k(x, x')$
$(x, x' - e_k)$	$\gamma_{kk+1}(x') - \widehat{\gamma}_k(x, x')$

(2.22)

where $k = 1, \dots, d$. Assume that $(\widetilde{X}_0, \widetilde{X}'_0) = (x, x)$ so that the coupled processes are started at the same initial state $x \in \mathbb{Z}_+^{\boldsymbol{b}}$. Let $D = (D_t)_{t \geqslant 0}$ and $D' = (D'_t)_{t \geqslant 0}$ be the counting processes associated with $\widetilde{X}$ and $\widetilde{X}'$, respectively.

Theorem 2.5.10. *If $D(0) \preceq D'(0)$ then $D(t) \preceq D'(t)$ for all $t \geqslant 0$.*

Proof. Fix $t > 0$ and suppose that $D(t) = y \preceq y' = D'(t)$. A transition of $(\widetilde{X}, \widetilde{X}')$ from (x, x') to $(x + e_1, x' + e_1)$ yields a transition of (D, D') from (y, y') to $(y + e_1, y' + e_1)$, and since $y + e_1 \preceq y' + e_1$, the order is preserved. Transitions of $(\widetilde{X}, \widetilde{X}')$ to $(x - e_k, x' - e_k)$ are analogously shown to preserve the order. Transitions to $(x, x' + e_1)$ and $(x, x' - e_k)$ gives transitions of (D, D') to $(y, y' + e_1)$ and $(y, y' + e_{k+1})$, respectively, hence, they also preserves the order.

A transition of $(\widetilde{X}, \widetilde{X}')$ to $(x - e_k, x')$ implies a transition of (D, D') to $(y + e_{k+1}, y')$, but such a transition has positive intensity only if

$$\gamma_{kk+1}(x) - \widehat{\gamma}_k(x, x') > 0$$

which imply

$$\mu_k(x_k) a_{kk+1}(x_m) > \mu_k(x'_k) a_{kk+1}(x'_m).$$

This means that $x_k > x'_k$ or $x_{k+1} < x'_{k+1}$. If $x_k > x'_k$ then, using (2.21),

$$y_{k-1} - y_k > y'_{k-1} - y'_k$$

which means

$$y'_k - y_k > y'_{k-1} - y_{k-1} \geqslant 0.$$

Hence, $y'_k > y_k$ and $y + e_{k+1} \preceq y'$ after the transition. Similarly, if $x_{k+1} < x'_{k+1}$ then

$$y'_k - y_k > y'_{k+1} - y_{k+1} \geqslant 0$$

and $y + e_{k+1} \preceq y'$.

Finally, a transition of $(\widetilde{X}, \widetilde{X}')$ to $(x + e_1)$ implies a transition of (D, D') to $(y + e_1, y')$. Such a transition has positive intensity only if

$$\beta_1(x) > \beta_1(x') \text{ which means } x_1 < x'_1.$$

Hence, $y'_0 - y_0 > 0$, that is $y + e_1 \preceq, y'$, and since all transitions of $(\widetilde{X}, \widetilde{X}')$ preserves the ordering of (D, D'), the proof is complete. $\square$

Theorem 2.5.10 states that in tandem SBE networks with persistent units, the number of departures at every node, up to time t, is stochastically increasing in the buffer sizes. Specifically, this applies to the last node, hence the throughput from the network,

$$\lim_{t\to\infty} \frac{1}{t} D_d(t)$$

(recall Definitions 2.4.1 and 2.4.2) is stochastically increasing in the buffer sizes. For $a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}}$ and $a'_{km}(x) = \mathbf{1}_{\{x_m < b'_m\}}$, a proof of Theorem 2.5.10 can also be found in Meester and Shanthikumar [17] and Anantharam and Tsoucas [2].

A tandem queueing network with d nodes is commonly used as a model of a production line. Many algorithms have been suggested for finding good buffer allocations, i.e., vectors $\boldsymbol{b}$, so that the throughput from the last node is maximised. Meester and Shanthikumar [17] and Anantharam and Tsoucas [2] have shown that the number of departures (and throughput) is not only stochastically increasing, but also concave in $\boldsymbol{b}$. The result in Meester and Shanthikumar [17] is established by a construction of uniform processes and a coupling of the departures at the nodes in the uniform chains.

The proof of Theorem 2.5.10 is almost identical to the proof of the next theorem. Let X and X' be two independent tandem SBE network with d nodes and persistent units. The networks X and X' are identical apart from the service rates for which

$$\mu_k(x_k) \leqslant \mu'_k(x_k) \quad \text{for all } x \in \mathbb{Z}_+^{\boldsymbol{b}}$$

is assumed. Thus, by letting

$$\gamma_{kk+1}(x) = \mu_k(x) a_{kk+1}(x_{k+1}) \quad \gamma'_{kk+1}(x) = \mu'_k(x) a_{kk+1}(x_{k+1}),$$

for $k = 1, \ldots, d$, where $a_{dd+1} \equiv 1$, and

$$\widehat{\gamma}_k(x, x') = \gamma_{kk+1}(x) \wedge \gamma'_{kk+1}(x')$$

a coupling of X and X', $(\widetilde{X}, \widetilde{X}')$, similar to the one before can be constructed; Use the newly defined $\gamma'_{kk+1}(x)$ and $\widehat{\gamma}_k(x, x')$ in (2.22) where the last line is replaced by

to	**with intensity**
$(x, x' - e_k)$	$\gamma'_{kk+1}(x') - \widehat{\gamma}_k(x, x').$

Theorem 2.5.11. *The throughput of a tandem SBE network with persistent units is stochastically increasing in the service rates at the nodes.*

Proof. The coupling $(\widetilde{X}, \widetilde{X}')$ of X and X' is such $\widetilde{X}_0 = \widetilde{X}'_0$, and $D = (D_t)_{t\geqslant 0}$ and $D' = (D'_t)_{t\geqslant 0}$ is the counting processes associated with $\widetilde{X}$ and $\widetilde{X}'$, respectively. The proof is identical to the proof of Theorem 2.5.10 if one notes that $\mu_k(x) > \mu'_k(x')$ implies $x_k > x'_k$. Thus, $D(0) \preceq D'(0)$ implies $D(t) \preceq D'(t)$ for all $t \geqslant 0$. Hence,

$$\lim_{t\to\infty} \frac{1}{t} D_d(t) \leqslant \lim_{t\to\infty} \frac{1}{t} D'_d(t)$$

shows the dominance in throughput. $\square$

Chapter 3

Perfect simulation

3.1 COUPLING IDENTICAL SBE NETWORKS

The coupling (2.19) for two identical SBE networks yields a domination of realisations, or trajectories, of the network processes. The coupling is now constructed by embedding the processes in a multi-dimensional Poisson process which not only has a nice interpretation but is also easily implemented as a computer programme.

Let $(Z_t)_{t\geqslant 0}$ and $(Z'_t)_{t\geqslant 0}$ be two copies of a Poisson environment of dimension $r \in \mathbb{N}$. First the coupling of Corollary 2.5.2, $(Z_t, Z'_t)_{t\geqslant 0}$, is constructed by the embedding technique, and after which, the coupling (2.19) is constructed in a similar manner. Let

$$\lambda_z = \sum_{k=1}^{r} \nu_k + \sigma_k \tag{3.1}$$

and $(N_t)_{t\geqslant 0}$ be a homogeneous Poisson process on $\mathbb{R}_+$ with intensity λ_z. Define $(A_k)_{k=1}^r$ to be a partition of $[0, \sum_k \nu_k]$, such that $l(A_k) = \nu_k$ for $k = 1, \dots, r$. Similarly, define $(B_k)_{k=1}^r$ so that $l(B_k) = \sigma_k$, and $(A_k \cup B_k)_{k=1}^r$ is a partition of $[0, \lambda_z]$. Let $\phi_z : \{0,1\}^r \times [0,1] \to \{0,1\}^r$ be the function

$$\phi_z(x, u) = \begin{cases} x + e_k \mathbf{1}_{\{x_k = 0\}} & \text{if } \lambda_z u \in B_k \\ x - e_k \mathbf{1}_{\{x_k = 1\}} & \text{if } \lambda_z u \in A_k. \end{cases} \tag{3.2}$$

Finally, let $(U_n)_{n\geqslant 0}$ be a sequence of independent random variables, uniform on $[0, 1]$.

Theorem 3.1.1. *Let $\preceq$ be the usual partial ordering of E and retain the notation in the previous paragraph. Fix $z, z' \in E$ such that $z \preceq z'$. Let $(\widehat{Z}_n, \widehat{Z}'_n)_{n\geqslant 0}$ have initial value $(\widehat{Z}_0, \widehat{Z}'_0) = (z, z')$, and define recursively,*

$$\widehat{Z}_{n+1} = \phi_z(\widehat{Z}_n, U_n) \qquad \widehat{Z}'_{n+1} = \phi_z(\widehat{Z}'_n, U_n)$$

for $n \geqslant 0$. Then $(Z_t, Z'_t)_{t\geqslant 0}$, given by

$$Z(t) = \widehat{Z}_{N(t)} \qquad Z'(t) = \widehat{Z}'_{N(t)}$$

for all $t \geqslant 0$, is a coupling of two identical Poisson environments such that $Z(t) \preceq Z'(t)$ for all $t \geqslant 0$ almost surely, whenever $\widehat{Z}_0 \preceq \widehat{Z}'_0$, or equivalently, $Z(0) \preceq Z'(0)$.

Proof. The process $(\widehat{Z}_n, \widehat{Z}'_n)_{n\geqslant 0}$ on $\{0,1\}^r \times \{0,1\}^r$ has transition probabilities from state (z, z') given by

to	**with probability**
$(z + e_k \mathbf{1}_{\{z_k=0\}}, z' + e_k \mathbf{1}_{\{z'_k=0\}})$	σ_k/λ_z
$(z - e_k \mathbf{1}_{\{z_k=1\}}, z' - e_k \mathbf{1}_{\{z'_k=1\}})$	ν_k/λ_z,

for $k = 1, \ldots, r$. Further, $(\widehat{Z}_n, \widehat{Z}'_n)_{n\geqslant 0}$ is a subordinate chain of the Markov chain $(Z_t, Z'_t)_{t\geqslant 0}$ on $\{0,1\}^r \times \{0,1\}^r$, with $(N_t)_{t\geqslant 0}$ as the clock. Evaluation of the transition intensities of $(Z_t, Z'_t)_{t\geqslant 0}$ from state (z, z'), where $z \preceq z'$, to different states, by (1.14), yields

to	**with intensity**
$(z + e_k, z' + e_k)$	$\sigma_k \mathbf{1}_{\{z'_k=0\}}$
$(z + e_k, z')$	$\sigma_k \mathbf{1}_{\{z_k=0\}} \mathbf{1}_{\{z'_k=1\}}$
$(z - e_k, z' - e_k)$	$\nu_k \mathbf{1}_{\{z_k=1\}}$
$(z, z' - e_k)$	$\nu_k \mathbf{1}_{\{z_k=0\}} \mathbf{1}_{\{z'_k=1\}}$.

Introducing $\sigma_k(z) = \sigma_k \mathbf{1}_{\{z_k=0\}}$, $\nu_k(z) = \nu_k(z) \mathbf{1}_{\{z_k=1\}}$ as before, then

$$\begin{aligned}
\sigma_k \mathbf{1}_{\{z_k=0\}} \mathbf{1}_{\{z'_k=1\}} &= \sigma_k \mathbf{1}_{\{z_k=0\}} (1 - \mathbf{1}_{\{z'_k=0\}}) = \sigma_k \mathbf{1}_{\{z_k=0\}} - \sigma_k \mathbf{1}_{\{z'_k=0\}} \\
&= \sigma_k(z) - \sigma_k(z'),
\end{aligned}$$

and $\nu_k \mathbf{1}_{\{z_k=0\}} \mathbf{1}_{\{z'_k=1\}} = \nu_k(z') - \nu_k(z)$. A comparison of this with (2.17) reveals that $(Z_t, Z'_t)_{t\geqslant 0}$ has the same transition rates as the coupling in Corollary 2.5.2, and the statement follows. $\square$

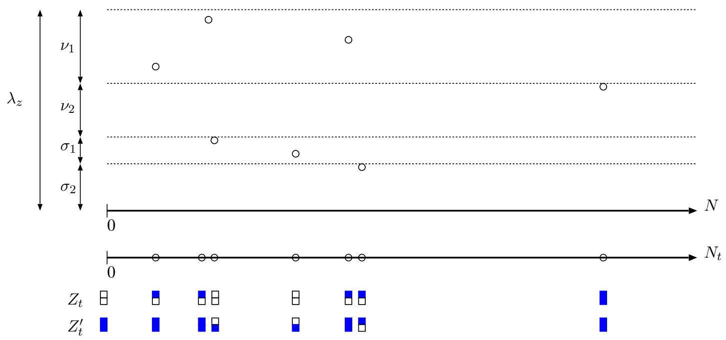

Figure 3.1. *The two-dimensional Poisson process N on $[0,\lambda_z]\times\mathbb{R}_+$ generates points that fall in different sets of the partition of the interval. Each partition corresponds to an event of the process $(Z_t, Z_t') = (\widehat{Z}_{N(t)}, \widehat{Z}'_{N(t)})$, here shown as two* ON/OFF*-processes of dimension $r = 2$. A sample trajectory is shown at the bottom where filled boxes indicate the 'on' states, and empty boxes the 'off' states.*

To conclude, the coupling is constructed by updating the subordinate processes $(\widehat{Z}_n)_{n\geqslant 0}$ and $(\widehat{Z}'_n)_{n\geqslant 0}$ with the same transition rule ϕ_z, the same sequence of uniform variables, $(U_n)_{n\geqslant 0}$, and, finally, the same clock $(N_t)_{t\geqslant 0}$. To use the same transition rule ϕ_z is equivalent to having the same partition $(A_k)_{k=1}^r$, $(B_k)_{k=1}^r$ of the interval $[0,\lambda_z]$.

If one enumerates the points, $(T_n)_{n\geqslant 0}$, of $(N_t)_{t\geqslant 0}$ such that $T_n < T_{n+1}$, $n \geqslant 0$, it can be seen that the sequence $(T_n, \lambda_z U_n)_{n\geqslant 0}$ is nothing but the enumeration of points of a two-dimensional Poisson process on $[0,\lambda_z]\times\mathbb{R}_+$ with the Lebesgue measure as parameter measure. See Section 1.4 and Figure 3.1 for aid in visualising the construction.

An SBE network process $(X_t, Z_t)_{t\geqslant 0}$ can be constructed from a subordinate chain $(X_n, Z_n)_{n\geqslant 0}$ in a similar manner. For $(x,z) \in \mathbb{Z}_+^{\boldsymbol{b}}\times\{0,1\}^r$ and a transition rule ϕ,

$$\phi : \mathbb{Z}_+^{\boldsymbol{b}}\times\{0,1\}^r\times[0,1] \to \mathbb{Z}_+^{\boldsymbol{b}}\times\{0,1\}^r,$$

the subordinate chain is given recursively by

$$(X_0, Z_0) = (x,z), \qquad (X_{n+1}, Z_{n+1}) = \phi((X_n, Z_n), U_n),\ n \geqslant 0,$$

where $(U_n)_{n\geqslant 0}$ is an i.i.d. sequence of random variables uniform on $[0,1]$. The clock, of rate $\lambda > \sup q_{(x,z),(x',z')}$, and the sequence $(U_n)_{n\geqslant 0}$ can be simultaneously generated from N, a two-dimensional Poisson process on $\mathbb{R}_+ \times [0, \lambda]$ with the Lebesgue measure as parameter measure. The particular transition rule to be devised is induced by a partition of $[0, \lambda]$ so that $((X_n, Z_n), (X'_n, Z_n))_{n \geqslant 0}$, given by

$$(X_{n+1}, Z_{n+1}) = \phi((X_n, Z_n), U_n) \quad (X'_{n+1}, Z_{n+1}) = \phi((X'_n, Z_n), U_n),$$

is the coupling (2.19) of the subordinate chains for two SBE network processes.

Let $(X_t, Z_t)_{t \geqslant 0}$ be a d-node SBE network. Recall the assumptions that $\beta_k(x)$ is non-increasing in x_k, and $\mu_k(x)$ is non-decreasing in x_k. Let

$$\beta_k = \beta_k(0) \qquad \mu_k = \mu_k(b_k)$$

for $k = 1, \ldots, d$. Further, define $\beta = \sum_{k=1}^{d} \beta_k$, $\mu = \sum_{k=1}^{d} \mu_k$ and $\lambda = \beta + \mu + \lambda_z$, where λ_z is given by (3.1). Then, λ is the total rate of events in an SBE network, and β_k, for example, is the maximum intensity of arrivals to node k of that network. The interval $[0, \lambda]$ is partitioned into sets

$$I(\lambda_z),\ I(\beta_k),\ I(\mu_k), \quad k = 1, \ldots, d.$$

Fix a state $(x, z) \in \mathbb{Z}_+^{\boldsymbol{b}} \times \{0,1\}^r$ and a sequence of values, $u = (u_n)_{n=1}^{5}$, on $[0, 1]$. To simplify notation, let $p_{k0} = p_k$ for $k = 1, \ldots, d$, and $I(p_{km})$, $m = 0, \ldots, d$, be a partition of the unit interval for every $k = 1, \ldots, d$. Let ϕ_z be given by (3.2) and define $\phi : \mathbb{Z}_+^{\boldsymbol{b}} \times \{0,1\}^r \times [0,1]^5 \to \mathbb{Z}_+^{\boldsymbol{b}} \times \{0,1\}^r$ as follows:

(3.3)

- If $\lambda u_1 \in I(\lambda_z)$ then $\phi((x, z), u) = (x, \phi_z(z, u_2))$.
- If $\lambda u_1 \in I(\beta_k)$ then there is a potential *arrival* to node k. If $\beta_k u_2 \leqslant f_k(z) \beta_k(x_k)$ then $\phi((x, z), u) = (x + e_k, z)$.
- If $\lambda u_1 \in I(\mu_k)$ then there is a potential *departure* from node k. If $\mu_k u_2 \leqslant g_k(z) \mu_k(x_k)$ and
  - $u_3 \in I(p_{k0})$ then $\phi((x, z), u) = (x - e_k, z)$;
  - $u_3 \in I(p_{km})$ there is a potential transfer of a unit to node m of the network. If $u_4 \leqslant a_{km}(x_m)$ then $\phi((x, z), u) = (x - e_k + e_m, z)$, and otherwise, if $u_4 > a_{km}(x_m)$ and $u_5 \leqslant c_{km}(x)$ then $\phi((x, z), u) = (x - e_k, z)$.
- Otherwise, $\phi((x, z), u) = (x, z)$.

The transition function ϕ, stated algorithmically above, induces a partition of the interval $[0, \lambda]$ into sets corresponding to arrivals, departures and transfers of units, for every state (x, z). The way the sets overlap for different state combinations (x, z) is what determines the coupling. The sets are given implicitly by ϕ; as an example of how the sets for different states (x, z) relate to each other, consider the following. Suppose $\lambda u_1 \in I(\beta_k)$. Since $\beta_k(x)$ is non-increasing in x_k, $\beta_k u_2 \leqslant f_k(z)\beta_k(x_k)$ implies $\beta_k u_2 \leqslant f_k(z)\beta_k(n)$ for all $n \leqslant x_k$. Hence, induced sets of arrivals at node k, $I(\beta_k(n))$, may be defined as

$$I(\beta_k(b_k)) \subseteq I(\beta_k(b_k - 1)) \subseteq \cdots \subseteq I(\beta_k(0)) \subseteq I(\beta_k),$$

a sequence of subsets of $I(\beta_k)$ such that $l(I(\beta_k(x_k))) = f_k(z)\beta_k(x_k)$. The conditions on u_1 and u_2 in the construction of ϕ, may then be expressed as $\lambda u_1 \in I(\beta_k(x_k))$.

Fix $z \in \{0,1\}^r$ and $x, x' \in \mathbb{Z}_+^{\boldsymbol{b}}$ such that $x \preceq x'$, and let $U = (U_n)_{n=1}^5$ be i.i.d. uniform on $[0, 1]$. The probabilities for transitions from $((x, z), (x', z))$ to $(\phi((x, z), U), \phi((x', z), U))$ shall now be established. Clearly, the transition probability from state $((x, z), (x', z))$ to $((x, z'), (x', z'))$, $z \neq z'$, is

$$\frac{\lambda_z}{\lambda} \mathsf{P}(\phi_z(z, U_2) = z'),$$

which means the continuous-time process $Z_{N(t)}$ has the transition intensities of a Poisson environment. Transitions from $((x, z), (x', z))$ to

- $((x + e_k, z), (x' + e_k, z))$
  occur if $\lambda U_1 \in I(\beta_k)$, $\beta_k U_2 \leqslant f_k(z)\beta_k(x)$ and $\beta_k U_2 \leqslant f_k(z)\beta_k(x')$. Using $\beta_k(x') \leqslant \beta_k(x)$ and the uniform distribution of U_1 and U_2, the transition probability is

$$\frac{1}{\lambda} f_k(z)\beta_k(x') = \frac{1}{\lambda}\beta_k(x', z).$$

- $((x + e_k, z), (x', z))$
  occur if $\beta_k U_2 \leqslant f_k(z)\beta_k(x)$ and $\beta_k U_2 > f_k(z)\beta_k(x')$. That is, the transition probability is

$$\begin{aligned}
&\mathsf{P}(\lambda U_1 \in I(\beta_k),\ f_k(z)\beta_k(x') < \beta_k U_2 \leqslant f_k(z)\beta_k(x)) \\
&\quad = \frac{\beta_k}{\lambda} \cdot \frac{1}{\beta_k}(f_k(z)\beta_k(x) - f_k(z)\beta_k(x')) \\
&\quad = \frac{1}{\lambda}(\beta_k(x, z) - \beta_k(x', z)).
\end{aligned}$$

- $((x,z),(x'-e_k,z))$
  happen only if $\lambda U_1 \in I(\mu_k)$ and $g_k(z)\mu_k(x_k) < \mu_k U_2 \leqslant g_k(z)\mu_k(x'_k)$. Furthermore, either $U_3 \in I(p_{k0})$ or $U_3 \in I(p_{km})$, $U_4 > a_{km}(x')$ and $U_5 \leqslant c_{km}(x')$. Hence, the transition probability is

$$\begin{aligned}
&\mathsf{P}\Bigg(\lambda U_1 \in I(\mu_k),\ g_k(z)\mu_k(x) < \mu_k U_2 \leqslant g_k(z)\mu_k(x'),\\
&\quad \{U_3 \in I(p_{k0})\} \bigcup_{m=1}^{d} \{U_3 \in I(p_{km}), U_4 > a_{km}(x'), U_5 > c_{km}(x')\}\Bigg)\\
&\quad = \frac{\mu_k}{\lambda} \cdot g_k(z)\frac{\mu_k(x') - \mu_k(x)}{\mu_k}\Bigg(p_k + \sum_{m=1}^{d} p_{km}(1 - a_{km}(x'))c_{km}(x')\Bigg).
\end{aligned}$$

- $((x-e_k,z),(x',z))$
  require that $\lambda U_1 \in I(\mu_k)$ and $\mu_k U_2 \leqslant g_k(z)\mu_k(x)$. Further, $U_3 \in I(p_{km})$ for some $m = 1,\dots,d$, for which $U_4 > a_{km}(x)$ and $U_5 \leqslant c_{km}(x)$. Since, x' is left unchanged, $U_4 > a_{km}(x')$ and $U_5 > c_{km}(x')$, which means that the transition probability is

$$\begin{aligned}
&\mathsf{P}\Bigg(\lambda U_1 \in I(\mu_k),\ \mu_k U_2 \leqslant g_k(z)\mu_k(x),\\
&\quad \bigcup_{m=1}^{d} \{U_3 \in I(p_{km}), U_4 > a_{km}(x), c_{km}(x') \leqslant U_5 \leqslant c_{km}(x)\}\Bigg)\\
&\quad = \frac{\mu_k}{\lambda} \cdot \frac{g_k(z)\mu_k(x_k)}{\mu_k}\sum_{m=1}^{d} p_{km}(1 - a_{km}(x))[c_{km}(x) - c_{km}(x')],
\end{aligned}$$

  remembering that $a_{km}(x)$ is non-increasing in x_m, and $c_{km}(x)$ is non-decreasing in x.

- $((x-e_k,z),(x'-e_k,z))$
  occur only if $\lambda U_1 \in I(\mu_k)$ and $\mu_k U_2 \leqslant g_k(z)\mu_k(x)$. Either $U_3 \in I(p_{k0})$ or $U_3 \in I(p_{km})$ for some m, for which $U_4 > a_{km}(x)$ and $U_5 > c_{km}(x')$. Hence, the transition probability is

$$\begin{aligned}
&\mathsf{P}\Bigg(\lambda U_1 \in I(\mu_k),\ \mu_k U_2 \leqslant g_k(z)\mu_k(x),\\
&\quad \{U_3 \in I(p_{k0})\} \bigcup_{m=1}^{d} \{U_3 \in I(p_{km}), U_4 > a_{km}(x), U_5 \leqslant c_{km}(x')\}\Bigg)\\
&\quad = \frac{\mu_k}{\lambda} \cdot \frac{g_k(z)\mu_k(x)}{\mu_k}\Bigg(p_k + \sum_{m=1}^{d} p_{km}(1 - a_{km}(x))c_{km}(x')\Bigg).
\end{aligned}$$

- $((x,z),(x'-e_k+e_m,z))$
  happen only if $\lambda U_1 \in I(\mu_k)$ and $g_k(z)\mu_k(x_k) < \mu_k U_2 \leqslant g_k(z)\mu_k(x'_k)$. Further, $U_3 \in I(p_{km})$ and $U_4 \leqslant a_{km}(x')$, which implies that the transition probability is

$$\begin{aligned}\mathsf{P}\Big(&\lambda U_1 \in I(\mu_k),\ g_k(z)\mu_k(x) < \mu_k U_2 \leqslant g_k(z)\mu_k(x'),\\ &U_3 \in I(p_{km}), U_4 \leqslant a_{km}(x')\Big)\\ &\qquad= \frac{\mu_k}{\lambda}\cdot g_k(z)\frac{\mu_k(x')-\mu_k(x)}{\mu_k}p_{km}a_{km}(x').\end{aligned}$$

- $((x-e_k+e_m,z),(x',z))$
  require $\lambda U_1 \in I(\mu_k)$, $\mu_k U_2 \leqslant g_k(z)\mu_k(x_k)$ and $U_3 \in I(p_{km})$. Further, $a_{km}(x') < U_4 \leqslant a_{km}(x)$ and $U_5 > c_{km}(x')$. Hence, the transition probability is

$$\begin{aligned}\mathsf{P}\Big(&\lambda U_1 \in I(\mu_k),\ \mu_k U_2 \leqslant g_k(z)\mu_k(x'),\\ &U_3 \in I(p_{km}), a_{km}(x) < U_4 \leqslant a_{km}(x'), U_5 > c_{km}(x')\Big)\\ &\qquad= \frac{\mu_k}{\lambda}\cdot \frac{g_k(z)\mu_k(x)}{\mu_k}p_{km}[a_{km}(x)-a_{km}(x')](1-c_{km}(x')).\end{aligned}$$

- $((x-e_k+e_m,z),(x'-e_k,z))$
  require the same as above, but the restriction on U_5 is replaced by $U_5 \leqslant c_{km}(x')$. Hence, the transition probability is

$$\begin{aligned}\mathsf{P}\Big(&\lambda U_1 \in I(\mu_k),\ \mu_k U_2 \leqslant g_k(z)\mu_k(x'),\\ &U_3 \in I(p_{km}), a_{km}(x) < U_4 \leqslant a_{km}(x'), U_5 \leqslant c_{km}(x')\Big)\\ &\qquad= \frac{\mu_k}{\lambda}\cdot \frac{g_k(z)\mu_k(x)}{\mu_k}p_{km}[a_{km}(x)-a_{km}(x')]c_{km}(x').\end{aligned}$$

- $((x-e_k+e_m,z),(x'-e_k+e_m,z))$
  happen if $\lambda U_1 \in I(\mu_k)$, $\mu_k U_2 \leqslant g_k(z)\mu_k(x_k)$, $U_3 \in I(p_{km})$ and $U_4 \leqslant a_{km}(x')$. Thus, the transition probability is

$$\begin{aligned}\mathsf{P}\Big(&\lambda U_1 \in I(\mu_k),\ \mu_k U_2 \leqslant g_k(z)\mu_k(x'), U_3 \in I(p_{km}), U_4 \leqslant a_{km}(x')\Big)\\ &\qquad= \frac{\mu_k}{\lambda}\cdot \frac{g_k(z)\mu_k(x)}{\mu_k}p_{km}a_{km}(x').\end{aligned}$$

Therefore, using ϕ as a transition rule, one may construct a subordinate chain, $(X_n, Z_n)_{n\geqslant 0}$,

$$(X_{n+1}, Z_{n+1}) = \phi((X_n, Z_n), V_n)),\ n \geqslant 0,$$

of an SBE network process $(X_t, Z_t)_{t\geqslant 0}$,

$$(X(t), Z(t)) = (X_{N(t)}, Z_{N(t)}),\ t \geqslant 0,$$

where $V_n \overset{\mathscr{D}}{=} (U_1, \ldots, U_5)$ are independent, and $(N(t))_{t\geqslant 0}$ is a homogeneous Poisson process with rate λ. Fix two states, $(x, z) \neq (x', z')$. The transition probability of (X_n, Z_n) from (x, z) to (x', z'), $p_{(x,z),(x',z')}$, induces a transition rate of $(X(t), Z(t))$: $q_{(x,z),(x',z')} = \lambda p_{(x,z),(x',z')}$. Thus, inspection of the transition probabilities of (X_n, Z_n) reveals that ϕ gives not only the correct intensities for an SBE network but also constructs the coupling (2.19). Thus, the reward of the lengthy work to obtain a transition rule for the subordinate chain $(X_n, Z_n)_{n\geqslant 0}$ is a proof of the following theorem.

Theorem 3.1.2. *Let $(\widehat{X}_n, \widehat{Z}_n)_{n\geqslant 0}$ and $(\widehat{X}'_n, \widehat{Z}'_n)_{n\geqslant 0}$ be discrete-time Markov chains on $E\times\{0,1\}^r$, and $(N_t)_{t\geqslant 0}$ a time homogeneous Poisson process with rate λ. If the chains satisfy $(\widehat{X}_0, \widehat{Z}_0) = (x, z)$ and $(\widehat{X}'_0, \widehat{Z}'_0) = (x', z)$, where $x \preceq x'$ and $z \in \{0,1\}^r$, then*

$$(\widehat{X}_{n+1}, \widehat{Z}_{n+1}) = \phi((\widehat{X}_n, \widehat{Z}_n), V_n)$$
$$(\widehat{X}'_{n+1}, \widehat{Z}'_{n+1}) = \phi((\widehat{X}'_n, \widehat{Z}'_n), V_n),$$

for $n \geqslant 0$, where $V_n \overset{\mathscr{D}}{=} (U_1, \ldots, U_5)$ is a vector of independent random variables uniform on $[0,1]$, and ϕ satisfies (3.3), then

$$(\widehat{X}_n, \widehat{Z}_n) \preceq (\widehat{X}'_n, \widehat{Z}'_n) \quad \text{[a.s.]}$$

for all $n \geqslant 0$. Specifically, with

$$(X_t, Z_t) = (\widehat{X}_{N(t)}, \widehat{Z}_{N(t)}) \qquad (X'_t, Z'_t) = (\widehat{X}'_{N(t)}, \widehat{Z}'_{N(t)}),$$

then $(X_t, Z_t) \preceq (X'_t, Z'_t)$ almost surely, for all $t \geqslant 0$.

Note that for a random variable, U, uniform on $[0,1]$,

$$Y = \frac{U-a}{b-a}\bigg| U\in[a,b],\ 0 \leqslant a < b \leqslant 1,$$

implies $Y \overset{\mathscr{D}}{=} U$. Using this result, the transition rule ϕ can be rewritten as a mapping,

$$\phi : \mathbb{Z}_+^{\boldsymbol{b}} \times \{0,1\}^r \times [0,1] \to \mathbb{Z}_+^{\boldsymbol{b}} \times \{0,1\}^r,$$

which is precisely the form of mapping needed for a two-dimensional Poisson embedding. An algorithmic representation of the transition rule ϕ is found in Figure 3.2.

Although no coupling such that $(X_t, Z_t) \preceq (X'_t, Z'_t)$ [a.s.], for $t \geqslant 0$, when $X_0 \preceq X'_0$ and $Z_0 \preceq Z'_0$, can be found unless $Z_0 = Z'_0$, the above construction is sufficient for formulating an algorithm which generates perfect samples from the stationary distribution.

3.2 THE REVISED PROPP–WILSON ALGORITHM

Let E be a state space and $U = (U_n)_{n \geqslant 1}$ a sequence of i.i.d. random variables uniform on $[0,1]$. Define

$$\Phi_n^n(x,U) = x, \quad \Phi_n^k(x,U) = \Phi_{n+1}^k(\phi(x,U_n),U), \quad 0 \leqslant n < k, \forall x \in E,$$

where ϕ is the transition rule for some discrete-time Markov chain X on E. Then for an initial state X_0, $X_n = \Phi_0^n(X_0, U)$ for $n \geqslant 0$.

Theorem 3.2.1. *If X is a discrete-time irreducible aperiodic Markov chain with transition rule ϕ, then Φ_0^n is a constant mapping almost surely, as $n \to \infty$, that is,*

$$\mathsf{P}\left(\lim_{n\to\infty} \Phi_0^n(x,U) = c(\omega),\ \forall x \in E\right) = 1.$$

Proof. Since X is irreducible and aperiodic, there exists an m such that $\mathsf{P}(X_n = y | X_0 = x) > 0$ for all $x, y \in E$ and $n \geqslant m$. Thus, $\Phi_0^m(\,\cdot\,, U)$ is a constant mapping with positive probability. The mappings $\Phi_{km}^{(k+1)m}(\,\cdot\,, U)$ are independent for $k \geqslant 0$, and the Borel–Cantelli Lemma gives us

$$\mathsf{P}\left(\bigcup_{k \geqslant 0} \{\Phi_{km}^{(k+1)m}(\,\cdot\,, U) \text{ constant}\}\right) = 1.$$

Thus, some m-step transition will eventually give a constant mapping, and

$$\Phi_0^{nm} = \Phi_{(n-1)m}^{nm} \circ \cdots \circ \Phi_0^m$$

yields that if $\Phi_{km}^{(k+1)m}$, for example, is a constant mapping, then Φ_0^n, $n \geqslant (k+1)m$ is a constant mapping. □

```
function φ(x, z, u)
u₁ ← λu
if u₁ < λ_z then
      u₂ ← u₁/λ_z
      return (x, φ_z(z, u₂))
u₁ ← u₁ − λ_z
k ← 1
while (u₁ > β_k + μ_k)
      u₁ ← u₁ − (β_k + μ_k)
      k ← k + 1
if u₁ ⩽ β_k then
      if u₁ ⩽ f_k(z)β_k(x) then return (x + e_k, z)
      else return (x, z)
u₂ ← u₁ − β_k
n ← 1
while (u₂ > μ_k(n))
      n ← n + 1
if n > x_k then return (x, z)
u₃ ← (u₂ − μ_k(n − 1))/(μ_k(n) − μ_k(n − 1))
if u₃ > g_k(z) then return (x, z)
u₃ ← u₃/g_k(z)
if u₃ ⩽ p_k then return (x − e_k, z)
m ← 1
while (u₃ > p_km)
      u₃ ← u₃ − p_km
      m ← m + 1
u₄ ← u₃/p_km
n ← b_m
while (u₄ > a_km(n))
      n ← n − 1
if x_m ⩽ n then return (x − e_k + e_m, z)
u₅ ← (u₄ − a_km(n + 1))/(a_km(n) − a_km(n + 1))
if u₅ ⩽ c_km(x) then return (x − e_k, z)
return (x, z)
```

Figure 3.2. *The transition rule ϕ written as pseudo-code. Note that the values $u_1, \ldots, u_5$, if used, must not depend on x.*

In the following Φ^k is used to denote Φ_0^k.

3.2.1 THE ALGORITHM

Suppose that $U = (U_n)_{n\geqslant 0}$ and $U' = (U'_n)_{n\geqslant 0}$ are independent sequences of i.i.d. random variables uniformly distributed on $[0,1]$. Let K and K' be the random indices given by

$$K = \inf\{k > 0 : \Phi^k(x,U) \text{ constant, for all } x \in E\}$$
$$K' = \inf\{k > 0 : \Phi^k(x,U') \text{ constant, for all } x \in E\},$$

for a k-step transition rule Φ^k of an ergodic Markov chain on E. By Theorem 3.2.1, K and K' are finite almost surely. Clearly K and K' are independent and $K \stackrel{\mathscr{D}}{=} K'$. Now,

$$\{\Phi^K(x,U') \text{ constant, for all } x \in E\} = \{K' \leqslant K\},$$

and by symmetry, $\mathsf{P}(K' \leqslant K) = \mathsf{P}(K \leqslant K')$, so

$$2\mathsf{P}(K' \leqslant K) = \mathsf{P}(K' \leqslant K) + \mathsf{P}(K' \geqslant K) = 1 + \mathsf{P}(K' = K) \geqslant 1. \quad (3.4)$$

The discrete-time subordinate chains $(\widetilde{X}'_n)_{n\geqslant 0}$ given by $\widetilde{X}'_n = \Phi^n(x,U')$, $n \in \mathbb{Z}_+$, $x \in E$, all take the same value at time K if the event $K' \leqslant K$ occurs; the chains, as functions of the initial states x, are said to have *coalesced* at time K. From (3.4) follows that the probability of coalescence, p, of $\Phi^K(x,U')$ is at least $\frac{1}{2}$.

To make $\mathsf{P}(K' \leqslant K) = \frac{1}{2}$, events $K = K'$ must be dealt with separately. In these situations, a fair independent coin is flipped to single out the 'smaller' of the two random variables. Let $\leqslant^*$ denote a stochastic binary relation with an implicit independent random 'coin' V, or formally, let $R \subset \mathbb{Z}_+ \times \mathbb{Z}_+ \times \{0,1\}$ be the trinary relation given by

$$\begin{aligned} x < y &\Rightarrow (x,y,0),(x,y,1) \in R \\ x = y &\Rightarrow (x,y,1) \in R, \end{aligned}$$

and $K \leqslant^* K' \equiv (K,K',V)\in R$, where V is a Bernoulli random variable independent of K and K'. Compare $\leqslant^*$ with $\leqslant$. Then

$$\begin{aligned} 2\mathsf{P}(K' \leqslant^* K) &= 2\mathsf{P}((K',K,V) \in R) \\ &= 2\mathsf{P}(K' < K) + 2\mathsf{P}(K' = K)\mathsf{P}(V = 1) \\ &= \mathsf{P}(K' < K) + \mathsf{P}(K' = K) + \mathsf{P}(K' > K) = 1. \end{aligned} \quad (3.5)$$

Hence, $\mathsf{P}(K' \leqslant^* K) = \frac{1}{2}$, and $\Phi^K(x, U')$ is a constant mapping if $K' \leqslant^* K$. This is done for two reasons: the first is to assure that $K \mid K \leqslant^* K'$ has the same distribution as $\min(K, K') \equiv K \wedge K'$:

$$\begin{aligned}
&\mathsf{P}(K \wedge K' \leqslant t) \\
&\qquad = \mathsf{P}(K \wedge K' \leqslant t, K < K') + \mathsf{P}(K \wedge K' \leqslant t, K = K') \\
&\qquad\quad + \mathsf{P}(K \wedge K' \leqslant t, K > K') \\
&\qquad = 2\mathsf{P}(K \leqslant t, K < K') + 2\mathsf{P}(K \leqslant t, K = K', V = 1) \qquad (3.6)\\
&\qquad = 2\mathsf{P}(K \leqslant t, K \leqslant^* K') \\
&\qquad = 2\mathsf{P}(K \leqslant t | K \leqslant^* K')\mathsf{P}(K \leqslant^* K') \\
&\qquad = \mathsf{P}(K \leqslant t | K \leqslant^* K').
\end{aligned}$$

Further, K and K' are symmetric with respect to the relation $\leqslant^*$, and in the same manner as above, one may show that

$$\max(K, K') \equiv K \vee K' \overset{\mathscr{D}}{=} K | K \nleqslant^* K' \overset{\mathscr{D}}{=} K | K' \leqslant^* K. \tag{3.7}$$

The second reason is that simulation of a geometrically distributed random variable with parameter p is simpler for $p = \mathsf{P}(K \leqslant^* K') = \frac{1}{2}$ than for $p = \mathsf{P}(K \leqslant K')$.

Let $(U_n^{(k)})_{n \geqslant 0}$, $({U'}_n^{(k)})_{n \geqslant 0}$, $k = 1, 2, \ldots$, be independent sequences of i.i.d. random variables uniformly distributed on $[0, 1]$. Let the sequences together with the transition mapping Φ generate independent random variables K_k, K'_k

$$\begin{aligned}
K_k &= \inf\{k > 0 : \Phi^k(x, U^{(k)}) \text{ constant, for all } x \in E\} \\
K'_k &= \inf\{k > 0 : \Phi^k(x, U'^{(k)}) \text{ constant, for all } x \in E\},
\end{aligned} \tag{3.8}$$

for $k \in \mathbb{N}$. Let N be given by

$$N = \inf\{k \geqslant 1 : K'_k \leqslant^* K_k\} \tag{3.9}$$

which, by independence and (3.5), is geometrically distributed with parameter $p = \frac{1}{2}$, and thus finite almost surely.

Theorem 3.2.2. *For mappings*

$$\Psi_n = \Phi^{K_1} \circ \cdots \circ \Phi^{K_n}, \; n \geqslant 1,$$

with implicit sequences $U'^{(1)}, \ldots, U'^{(n)}$*, then*

$$\Psi_N(\,\cdot\,) | N \overset{\mathscr{D}}{=} \pi. \tag{3.10}$$

Proof. By the almost sure finiteness of N and K_k, K'_k, $k \geqslant 1$,

$$K = K_1 + \cdots + + K_N | N < \infty$$

almost surely. Further, $\Phi^{K_N}|N$ is a constant mapping as is $\Psi_N|N$. If $X \stackrel{\mathscr{D}}{=} \pi$ and is independent of everything mentioned, then $\Phi^n(X) \stackrel{\mathscr{D}}{=} \pi$ for all $n \in \mathbb{Z}_+$, but $\Phi^n(X) = \Psi_N|N$ for all $n \geqslant K$ which proves the theorem. □

In fact, N is what is known as a *strong stationary* stopping time with respect to the sequence $(U^{(k)}, U'^{(k)})_{k\geqslant 1}$.

To achieve the revised Propp–Wilson algorithm, only calculation of the composition in Equation (3.10) remains. Since $(K_k)_{k\geqslant 1}$ is a sequence of i.i.d. random variables, the random mappings Φ^{K_k} are i.i.d. for $k \in \mathbb{N}$. Furthermore, conditioned on N, $K_1, \ldots, K_{N-1}$ is an i.i.d. sequence, conditionally independent of K_N and K'_N, and $K_i|N$, for $i = 1, \ldots, N-1$, have distribution from (3.6) and (3.9)

$$K_i|N \stackrel{\mathscr{D}}{=} K|K' \not\leqslant^* K \stackrel{\mathscr{D}}{=} K|K \leqslant^* K' \stackrel{\mathscr{D}}{=} K \wedge K'.$$

Hence,

$$\Phi^{K_i}(\,\cdot\,, U'^{(i)})\Big| N \stackrel{\mathscr{D}}{=} \Phi^{K \wedge K'}(\,\cdot\,, U \vee U'), \quad i = 1, \ldots, N-1, \tag{3.11}$$

where $U \vee U'$ denotes the sequence corresponding to the 'largest' of K and K', i.e.,

$$U \vee U' = \begin{cases} U' & \text{if } K \leqslant^* K' \\ U & \text{otherwise.} \end{cases}$$

Also conditioned on N,

$$K_N|N \stackrel{\mathscr{D}}{=} K|K' \leqslant^* K \stackrel{\mathscr{D}}{=} K \vee K',$$

and, using self-explanatory notation,

$$\Phi^{K_N}(\,\cdot\,, U'^{(N)})\Big| N \stackrel{\mathscr{D}}{=} \Phi^{K \vee K'}(\,\cdot\,, U \wedge U'). \tag{3.12}$$

Thus, conditioned on N, equations (3.11) and (3.12), gives a simple way calculate the composite mapping $\Psi_N|N$ in (3.10), where N is a geometric random variable with parameter $p = \frac{1}{2}$.

The result is formulated as an algorithm: the revised Propp–Wilson algorithm.

1. Generate N, geometrical distributed with parameter $p = \frac{1}{2}$.
2. For $k = 1, \dots, N$, construct K_k and K'_k from sequences $U^{(k)}, U'^{(k)}$ by (3.8).
3. Report the result of the constant mapping (3.13)

$$\Phi^{K_N \wedge K'_N} \circ \dots \circ \Phi^{K_2 \wedge K'_2} \circ \Phi^{K_1 \vee K'_1}$$

as a sample from the stationary distribution.

Suppressed in (3.13) is that the mappings $\Phi^{K_i \wedge K'_i}$, $i = 2, \dots, N$, are formed from $U^{(i)} \vee U'^{(i)}$, and $\Phi^{K_1 \vee K'_1}$ from $U^{(1)} \wedge U'^{(1)}$. Note that none of the random mappings $\Phi^{K_i \wedge K'_i} \circ \dots \circ \Phi^{K_2 \wedge K'_2} \circ \Phi^{K_1 \vee K'_1}$ have distribution π, i.e., the distribution of (3.10) conditioned on $N = i$.

Mappings satisfying (3.12) are called *winning mappings* and those satisfying (3.11) *losing mappings*. The composite mapping (3.10) is a composition of one winning mapping followed by a number of losing mappings, such that the number of mappings is a geometrically distributed random variable. The term ‘winning’ should be seen in light of the coalescence of the mapping induced by U',

$$\Phi^K(\,\cdot\,, U')\Big|\, K' \leqslant^* K,$$

which has coalesced *before* time K in some competition. Similarly for the losing mappings where the chains, $\Phi^K(\,\cdot\,, U')|K \leqslant^* K'$, have not coalesced by time K induced by U.

When evaluating the performance of the algorithm on the networks, measurements of the number of iterations are made. That is,

$$W = K|K' {\leqslant^*} K \stackrel{\mathscr{D}}{=} K {\wedge} K' \qquad L = K|K {\leqslant^*} K' \stackrel{\mathscr{D}}{=} K {\vee} K', \tag{3.14}$$

are the number of transitions in a winning and losing mapping, respectively, which leads to the total number of iterations,

$$S = W + \sum_{k=1}^{N-1} L_k,$$

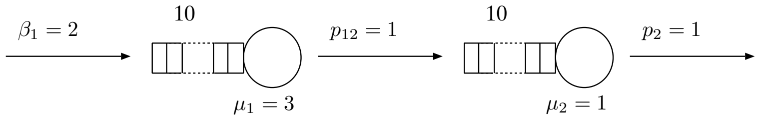

Figure 3.3. *An example of an SBE network with two nodes.*

where $(L_k)_{k=1}^{N-1}$ are i.i.d. , $L_k \stackrel{\mathscr{D}}{=} L$, and N is the total number of mappings needed before the sample was obtained.

Note that determining whether Φ^n is a constant mapping or not, can be significantly simplified by imposing some restrictions on E and ϕ. Suppose that the transition rule ϕ is monotone, i.e.,

$$x \preceq x' \;\;\Rightarrow\;\; \phi(x,u) \preceq \phi(x',u) \quad \text{for all } u \in [0,1].$$

This is the case with the coupling in Theorem 3.1.2 of an SBE network induced by the transition rule in Figure 3.2 (or Eq. (3.3)) without an environment. From the monotonicity assumption, and since $\mathbf{0} \preceq x \preceq \boldsymbol{b}, \forall x \in E$, induction gives

$$\Phi_0^n(\mathbf{0},U) \preceq \Phi_0^n(x,U) \preceq \Phi_0^n(\boldsymbol{b},U), \quad \forall x \in E,\; n \geqslant 0.$$

If for some $n > 0$, $\Phi^n(\mathbf{0},U) = \Phi_0^n(\boldsymbol{b},U)$, then $\Phi_0^n(\,\cdot\,,U)$ is a constant mapping. Thus, it suffices to follow the evolution of $\Phi^n(\mathbf{0},U)$ and $\Phi^n(\boldsymbol{b},U)$ until $\Phi^n(\mathbf{0},U) = \Phi^n(\boldsymbol{b},U)$, which implies a constant mapping at time n. An illustration of the algorithm applied on an SBE network can be found in Example 3.2.1.

Example 3.2.1. The SBE network $(X_t)_{t\geqslant 0}$ in Figure 3.3 has parameters

$$d = 2 \quad \boldsymbol{b} = (10,\,10) \quad p_{12} = 1 \quad c_{12} = 0.7 \quad p_2 = 1$$
$$\beta_1(x) = 2\cdot\mathbf{1}_{\{x_1<10\}} \quad \mu_1(x) = 3\cdot\mathbf{1}_{\{x_1>0\}} \quad \mu_2(x) = \mathbf{1}_{\{x_2>0\}}.$$

The successive mappings Φ^n made by the algorithm when it was applied to the network are shown in Figure 3.4. The perfect sample was obtained by a composition of one winning mapping and two losing mappings, $N = 3$:

$$\Phi^{K_1\vee K_1'} \circ \Phi^{K_2\wedge K_2'} \circ \Phi^{K_3\wedge K_3'} : x \mapsto (0,10),\; \forall x \in E,$$

where

$$K_1\vee K_1' = 40 \quad K_2\wedge K_2' = 57 \quad K_3\wedge K_3' = 40.$$

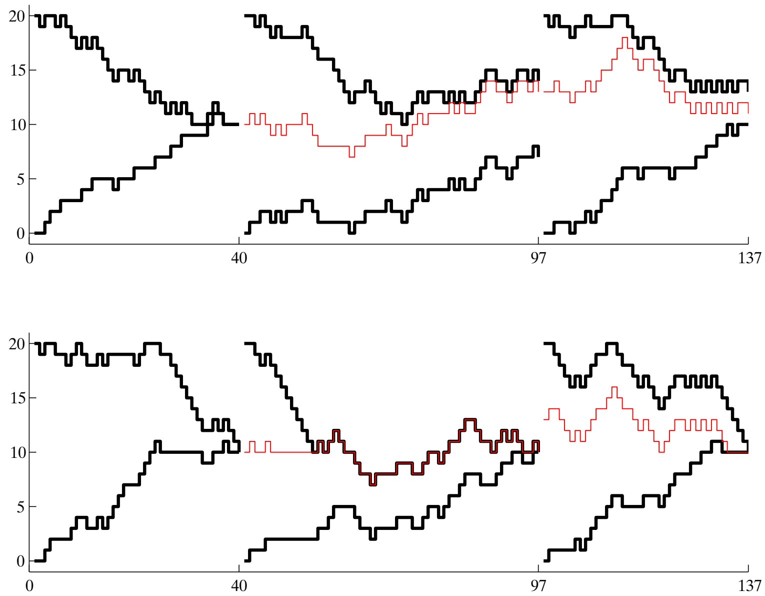

Figure 3.4. *The mappings obtained for a single simulation of the SBE network in Figure 3.3. Shown as thick lines are* $S(\Phi^n(\mathbf{0}, U))$ *and* $S(\Phi^n(\boldsymbol{b}, U))$ *(top), and* $S(\Phi^n(\mathbf{0}, U'))$ *and* $S(\Phi^n(\boldsymbol{b}, U'))$ *(bottom). The upper graph shows a winning mapping together with two successive losing mappings. The composition is indicated by the grey line. The sample obtained was* $(1, 10)$.

3.2.2 PERFECT SIMULATION WITH AN ENVIRONMENT

The arguments in previous sections need only a slight modification to efficiently handle simulations of processes with random environments.

Suppose that $(Y(t))_{t\geqslant 0} = (X(t), Z(t))_{t\geqslant 0}$ is an ergodic Markov process on $E = F \times G$, where $(Z(t))_{t\geqslant 0}$ is a random environment to $(X(t))_{t\geqslant 0}$. Let $(Y_n)_{n\geqslant 0} = (X_n, Z_n)_{n\geqslant 0}$ be the subordinate chain of $(Y(t))_{t\geqslant 0}$. Denote the transition rule of $(Y_n)_{n\geqslant 0}$ by ϕ, and that of $(Z_n)_{n\geqslant 0}$ by ϕ_z. Define recursively

$$\Phi_z^0(z,U) = z \quad \Phi_z^{n+1}(z,U) = \phi_z(\Phi_z^n(z,U), U_{n+1}),$$

$$\Phi_{xz}^n(z,U) = (x, \Phi_z^n(z,U)),$$

$n \geqslant 0, x \in F$, where $U = (U_n)_{n\geqslant 1}$ is a sequence of i.i.d. random variables uniform on $[0,1]$. Use the revised Propp–Wilson algorithm (3.13) and notation in (3.14) to obtain a mapping

$$\Phi_z^{L_{N-1}} \circ \cdots \circ \Phi_z^{L_1} \circ \Phi_z^{W} | N : z \mapsto z' \quad \forall z \in G,$$

that is, a constant mapping with distribution π_z, the stationary distribution of $(Z_n)_{n\geqslant 0}$ (and hence $(Z(t))_{t\geqslant 0}$ by Theorem 1.3.1) on G. Now, regard the whole chain $(Y_n)_{n\geqslant 0}$ and apply N' successive mappings according to the Propp–Wilson scheme, to obtain the mapping

$$\Phi^{L'_{N'-1}} \circ \cdots \circ \Phi^{L'_1} \circ \Phi^{W'} \circ \left(\Phi_{xz}^{L_{N-1}} \circ \cdots \circ \Phi_{xz}^{L_1} \circ \Phi_{xz}^{W}\right) | N', N$$

with distribution π on E. The parenthesis marks the mapping that coalesced the environment.

The reason for using two phases, coalescence of the environment followed by the whole process, is that a transition rule ϕ, monotone in x for fixed z, may be used. That is,

$$x \preceq x' \quad \text{implies} \quad \phi((x,z),u) \preceq \phi((x',z),u), \quad \forall u \in [0,1],\ z \in G. \tag{3.15}$$

This is exactly the setting in Theorem 3.1.2 with transition rule (3.3). The environment has a monotone transition rule ϕ_z from (3.2). The revised Propp–Wilson algorithm generates a sample

$$\widehat{Z} = \Phi_z^{L_{N-1}} \circ \cdots \circ \Phi_z^{L_1} \circ \Phi_z^{W},$$

$Z \stackrel{\mathscr{D}}{=} \pi_z$, from the stationary distribution of the environment. Then $\widehat{Z}$ is used as an initial value for the environments of the networks processes $(X_t, Z_t)_{t\geqslant 0}$, and

$$\phi((\mathbf{0}, \widehat{Z}), u) \preceq \phi((x, \widehat{Z}), u) \preceq \phi((\boldsymbol{b}, \widehat{Z}), u)$$

for all $x \in E$ and $u \in [0, 1]$.

Some vindicating remarks.

- This is really a monotone transition rule simulation. The environment is simulated with a monotone transition rule ϕ_z, and the simulation is started in $\mathbf{0}_r$ and $\mathbf{1}_r$, the minimal and maximal state on G respectively. When the two chains have coalesced, $\widehat{Z}$ is obtained. From that point, the whole process is tracked but instead of considering the evolution from an initial minimal and maximal state, $\mathbf{0}$ and $\boldsymbol{b}$, randomised such states, namely $(\mathbf{0}, \widehat{Z})$ and $(\boldsymbol{b}, \widehat{Z})$ are used, where $\widehat{Z} \stackrel{\mathscr{D}}{=} \pi_z$.

- It should be clear that the same technique can be used to simulate Markov chains where the random environment depends on another environment when the transition rule ϕ_z is monotone in the sense of (3.15). By recursively obtaining samples from the stationary distribution of a subenvironment by the revised Propp–Wilson algorithm, one may use these as randomised minimal and maximal initial states for the joint environment.

- The algorithm can of course be modified to be used on the continuous chains $(X(t))_{t\geqslant 0}$ and $(X'(t))_{t\geqslant 0}$ generated from the independent sequences U and U' and clocks N and N'. Defining

  $$K = \inf\{t > 0 : X^{(x)}(t) \text{ constant, for all } x \in E\}$$

  where $X^{(x)}$ is the chain with initial state $X(0) = x$, and similarly for K'. This would rid us from ties and thus, $K|K \leqslant^* K' \stackrel{\mathscr{D}}{=} K|K \leqslant K'$.

3.2.3 COMMENTS ON THE ALGORITHM

For an overview of pertinent algorithms and a survey of applications for which perfect simulations have been used advantageously, see Wilson [28] and Dimakos [7].

The revised Propp–Wilson algorithm was originally presented in Wilson [29] under the name *Coupling from the past-algorithm with a read-once source of randomness*. It is a refinement of the *Coupling from the past-algorithm* (CFTP) from Propp and Wilson [19]. The CFTP algorithm works by starting the Markov chains that describes the network at earlier and earlier time points, thus reusing the random numbers that generate the transitions; in this example, the algorithm uses the same two-dimensional Poisson embedding for all starting points. The construction in Theorem 3.2.2 is the CFTP algorithm with randomised starting points.

The algorithm in Wilson [29] is not exactly the same as the one presented here. There, U and U' induce a reference chain and a passive chain, respectively. Successive mappings of the passive chain are made, induced by the reference chain, until the first winning mapping is obtained. That map is then composed with the losing mappings before the next winning mapping. In our setting, U and U' induce chains that are used interchangeably to produce winning and losing mappings. This construction, which is based on Häggström [10] enables environments to be efficiently handled.

Perfect or *exact* simulation are terms for algorithms that generate samples from an unknown stationary distribution of a Markov chain. Using terminology from the theory of randomised algorithms, the Propp–Wilson is said to be a *Las Vegas* algorithm for random sampling. Its running time is stochastic but almost surely it gives an unbiased sample. The number of iterations for the algorithm before it terminates depends on the sample obtained. 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. Other 'interruptible' algorithms for perfect simulation have been developed; see for example Fill [8].

3.3 SIMULATION PERFORMANCE OF SBE NETWORKS

To study the effectiveness of the revised Propp–Wilson algorithm when applied to SBE networks is the main topic of this section. The performance of the sampling algorithm determines how large a network can be simulated in a reasonable amount of time. Networks of different sizes and parameter settings, but of given structures, are simulated, and the number of iterations is measured and plotted for each network. This should give an indication of the expected amount of work needed for obtaining samples from a stationary distribution.

Each simulation in this section is started with the same random seed, thus giving the same sequence of pseudo-random numbers. This is of no practical importance since the simulations use different amounts of random numbers, but that is how the simulations were carried out.

For each parameter setting, $X_1, \ldots, X_n$ are the n independent samples obtained by simulation. Associated with each sample is a sequence of mappings, and to each mapping a number of iterations, i.e., the number of transitions, including pseudo-transitions, necessary to produce the mapping. The transitions (iterations) in a winning mapping are referred to as w-transitions, and those in losing mappings, as l-transitions. The notation of the previous section is used. In (3.14) the w-transitions were denoted $W_1, \ldots, W_n$, and L_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, N_i - 1$, were the number of l-transitions in the j^{th} losing mapping in the i^{th} sample.

The number of mappings, N_i, $i = 1, \ldots, n$, is a geometrically distributed random variable with parameter $\frac{1}{2}$, so $\mathsf{E}[N_i] = 2$. The total number of iterations before a sample is obtained, is denoted

$$S_i = W_i + \sum_{j=1}^{N_i-1} L_{ij}. \tag{3.16}$$

The expected value $\mathsf{E}[W_i]$ is estimated by the mean, denoted

$$m_w = \frac{1}{n} \sum_{i=1}^{n} W_i,$$

and similarly m_l is the estimate of $\mathsf{E}[L_{ij}]$.

3.3.1 PERFORMANCE BY NETWORK SIZE

Consider the d-node SBE network X with the set of parameters

$$b_k = 100 \quad \beta_k(x) = \frac{2}{d}\mathbf{1}_{\{x_k \leqslant b_k\}} \quad \mu_k(x) = 3 \cdot \mathbf{1}_{\{x_k > 0\}} \quad p_{km} = \frac{\mathbf{1}_{\{k \neq m\}}}{d}$$

for $k, m \in \{1, \ldots, d\}$. That is, the network structure is a complete graph with d vertices. The routing probabilities are chosen so that units departing node k, chose the destination uniformly over the remaining $d-1$ nodes and the outside. There is no direct rerouting of units, $p_{kk} = 0$, $k = 1, \ldots, d$. The blocking policy is given by

$$a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}}, \qquad c_{km}(x) = 0,$$

meaning that a unit is accepted at a node if the buffer is not full, and rejected units remain at the present node in the network.

The choice of arrival rates $\beta_k(x)$ is motivated by the total intensity of units entering node k, here denoted Λ_k. They satisfy the traffic equations and, by symmetry,

$$\Lambda_k(x) = \beta_k(x) + \sum_{m \neq k} \Lambda_m(x) p_{km} = \frac{2}{d}\mathbf{1}_{\{x_k \leqslant b_k\}} + (d-1)\Lambda_k(x)\frac{1}{d},$$

or, solved for $\Lambda_k(x)$,

$$\Lambda_k(x) = 2 \cdot \mathbf{1}_{\{x_k \leqslant b_k\}},$$

for $k = 1, \ldots, d$. That is, the rate at which units enter a node, does not depend on d, and thus, the traffic load should virtually be the same for every node and network size. The performance of the algorithm, as measured in the number of iterations, should therefore reflect the dependency of network size d.

Networks with $d = 1, \ldots, 100$ nodes were simulated, and for each d, $n = 100$ perfect samples, $X_1^{(d)}, \ldots, X_n^{(d)}$, were taken, yielding a total of 10 000 samples. The number of iterations, $W_i^{(d)}$ and $L_{ij}^{(d)}$, associated with each sample are shown in Figure 3.5.

The expectations and variances of $W_i^{(d)}$ and $L_{ij}^{(d)}$ increase with the size of the network, d. The total number of mappings, $\sum_{i=1}^{n} N_i$, was 19 870 as compared with the expected number, $2n = 20\,000$. The samples of the total number of iterations, $S_i^{(d)}$, are shown in Figure 3.6. They are visibly clustered into groups determined by the number of mappings. Note that $S_1^{(d)}, \ldots, S_n^{(d)}$ are the cumulative sums of the observations in Figure 3.5, given by (3.16).

The n observations on $W^{(d)}$ are used to estimate $\mathsf{E}[W^{(d)}]$. The empirical mean $m_w(d)$ is compared to a least square fitted quadratic polynomial, $\widehat{m}_w(d)$, estimated from data using the model $\mathsf{E}[W^{(d)}] = a_w \cdot d^2$. Graphs of $m_w(d)$ and $\widehat{m}_w(d)$ as functions of d, can be found in Figure 3.7.

The average number of w-iterations, $m_w(d)$, coincides with the estimated trend $\widehat{m}_w(d)$. The observed estimated quadratic trend is given by

$$\widehat{m}_w(d) = \widehat{a}_w \cdot d^2 = 367.9 \cdot d^2$$

for $1 \leqslant d \leqslant 100$. The residuals $W_i^{(d)} - \widehat{m}_w(d)$ and $m_w(d) - \widehat{m}_w(d)$ are shown in Figure 3.8. Even though the variance increases, the behaviour

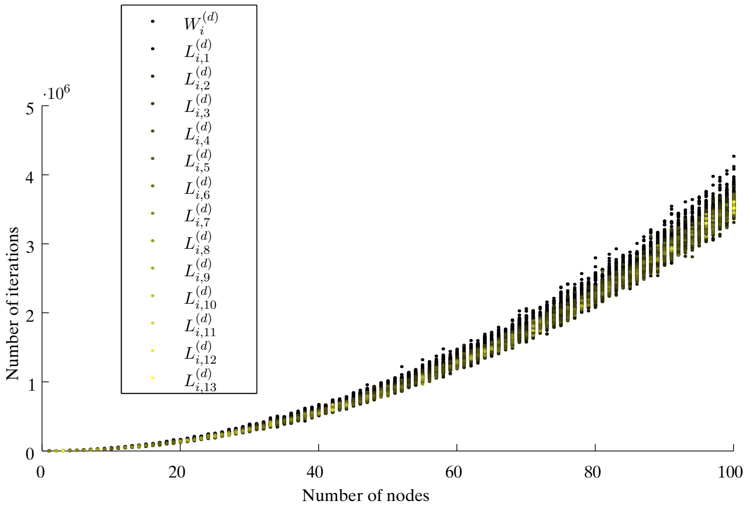

Figure 3.5. *The colours indicate at which phase, j, the value L_{ij} was obtained. No sample required more than 14 consecutive mappings, $N_i^{(d)} \leqslant 14$. Larger networks require more iterations, and the variance of the number of iterations increases with the network size. The number of observations on w-transitions and l-transitions was 19 870.*

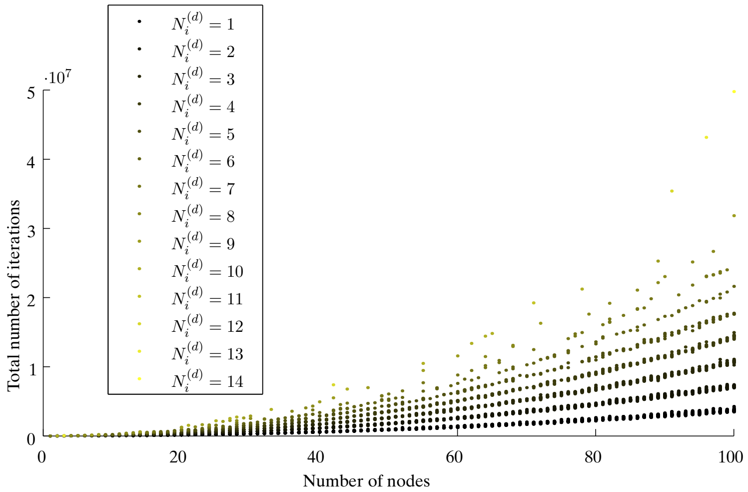

Figure 3.6. *The total number of iterations, $S_i^{(d)}$, for different network sizes, d. For a fixed network size, the samples on the total number of iterations are clustered into groups consisting of samples obtained on the first, second, third, and so on, mapping.*

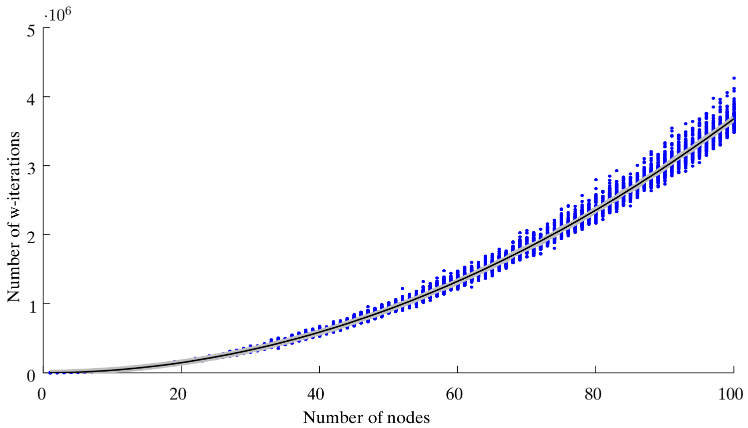

Figure 3.7. *The number of w-iterations for different network sizes. The thick grey line is the mean of the $n = 100$ samples, $m_w(d)$, and the thin black line is a least square fitted quadratic trend, $\widehat{m}_w(d)$.*

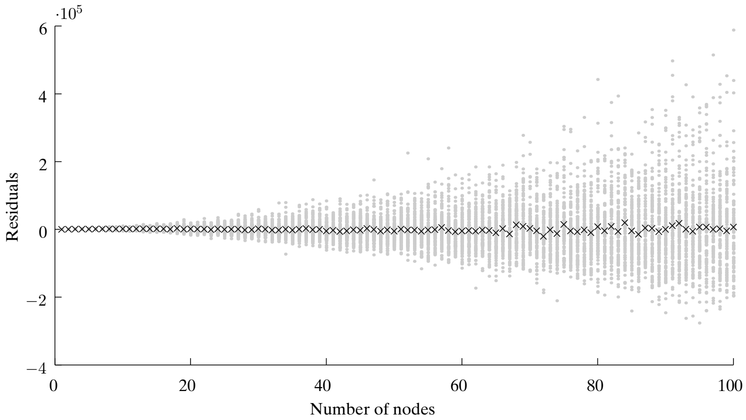

Figure 3.8. *The residuals between the observations $W_i^{(d)}$ and the fitted quadratic trend $\widehat{m}_w(d)$ are shown as dots. The crosses are the residuals between the mean $m_w(d)$ and the trend $\widehat{m}_w(d)$.*

of the average number of iterations is captured by the estimated quadratic polynomial.

Similarly, let $m_l(d)$ be the mean over i, j of $L_{ij}^{(d)}$, and $\widehat{m}_l(d)$ a least square fitted quadratic trend. Figure 3.9 shows $\widehat{m}_w(d)$ and $\widehat{m}_l(d)$. The difference $\widehat{m}_w(d) - \widehat{m}_l(d)$ increases with the network size d, but is comparatively small with respect to the mean number of l-iterations, $\widehat{m}_l(d)$. The number of w-iterations is approximately 20% larger than the l-iterations for small networks, and 4% larger for huge networks. The estimated relative differences are

$$\frac{\widehat{m}_w(d) - \widehat{m}_l(d)}{\widehat{m}_l(d)} = \widehat{a}_w / \widehat{a}_l - 1,$$

regardless of d, with the observed value 4.01%. Table 3.10 contains the observed relative differences in means, $m_w(d)/m_l(d) - 1$.

3.3.2 PERFORMANCE BY NETWORK STRUCTURE

To examine how the connectivity of a network affects the number of iterations, randomly connected d-node networks are constructed; the arrival rates are chosen so that each node has the same traffic load, regardless of the network topology. Let X be a d-node SBE network with buffer sizes $b_k = 100$ and service rates $\mu_k(x) = 3 \cdot \mathbf{1}_{\{x_k > 0\}}$ for $k = 1, \dots, d$. Fix a

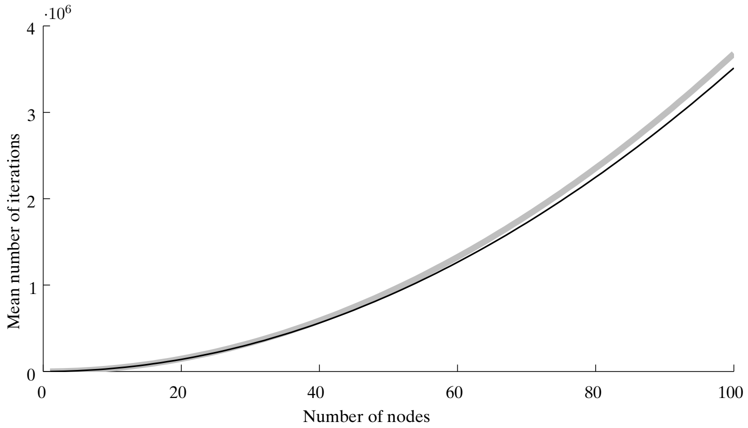

Figure 3.9. *The fitted quadratic trends for the number of iterations before a winning mapping is obtained, $\widehat{m}_w(d)$ (thick line), and the corresponding for a losing mapping, $\widehat{m}_l(d)$ (thin line).*

Table 3.10. The observed differences in means $m_w(d)/m_l(d) - 1$ for networks of different sizes d.

d	$\frac{m_w(d)}{m_l(d)} - 1$	d	$\frac{m_w(d)}{m_l(d)} - 1$
1	0.2095	60	0.0422
10	0.1140	70	0.0533
20	0.0744	80	0.0571
30	0.0715	90	0.0413
40	0.0469	100	0.0434
50	0.0550		

probability p, and for $k, m \in \{1, \dots, d\}$, $k \neq m$, let

$$I_{km} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{else} \end{cases}$$

be independent random variables. The variable I_{km} indicates whether or not there is a direct route from node k to m. Let

$$M_m = \sum_{\substack{k=1,\dots,d \\ k \neq m}} I_{km}$$

be the number of nodes with a direct route to m, which, by construction, is binomially distributed. For $k, m \in \{1, \dots, d\}$, $k \neq m$, let

$$p_{km} = I_{km} \frac{1}{d}$$

and $p_{kk} = 0$. The arrival intensities

$$\beta_k(x) = 2 \frac{d - M_k}{d} \mathbf{1}_{\{x_k < b_k\}},$$

make $\Lambda_k(x)$, the intensity of units entering node k, satisfy $\Lambda_k(x) = 2 \cdot \mathbf{1}_{\{x_k < b_k\}}$. This, since the traffic equations

$$\Lambda_k(x) = \beta_k(x) + \sum_{m=1}^{d} p_{mk} \Lambda_m(x),$$

imply

$$2 \cdot \mathbf{1}_{\{x_k < b_k\}} = \beta_k(x) + \frac{M_k}{d} 2 \cdot \mathbf{1}_{\{x_k < b_k\}}$$

which, after rearrangement, give the desired arrival rates.

The blocking policy is given by

$$a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}} \quad \text{and} \quad c_{km}(x) = 0.$$

Thus, in each network, every node sends serviced units to a binomially distributed number of nodes. Each node has the same probability of being included in the selection, and the routing is uniform over the nodes in the selection. The probability that a unit departs the network from node k is $1 - M_k/d$, which is at least $1/d$. Blocked units remain at their current nodes.

For a network size $d = 20$, the connection probability p is varied, and for each choice of p, $n = 100$ independent perfect samples $X_1^{(p)}, \ldots, X_{100}^{(p)}$ are taken, each on different randomly connected networks. As in the previous section, $W_i^{(p)}$, $L_{ij}^{(p)}$, and $S_i^{(p)}$, $i = 1, \ldots, n$, $j = 1, \ldots, K_i^{(p)}$, are compared. The average number of w-transitions is denoted $m_w(p)$, and $m_l(p)$ is the average number of l-transitions.

Figure 3.11 shows the observed values of $W_i^{(p)}$ and $L_{ij}^{(p)}$ for different values of p. There is nothing special with small values of the connection probability p, although the setting $p = 0$ is rather artificial; the networks consist of d identical isolated $M/M/1/b$ nodes, each of them could have been simulated separately. For values of p around 0.8, the number of w- and l-iterations, rather surprisingly, increases. The impact of this, with respect to the total number of iterations, is shown in Figure 3.12. In the region around 0.8 more values of p were examined to get a better view of the behaviour of $W_i^{(p)}$ and $L_{ij}^{(p)}$:

$$p \in \{0, 0.1, \ldots, 0.7, 0.71, 0.72, \ldots, 0.80, 0.805, 0.810, \ldots, 0.995\}.$$

Figure 3.13 shows the number of w-iterations together with the averages $m_w(p)$.

When the network size was changed to $d = 5$, Figures 3.14–3.16 were obtained. The networks possess a pattern of a large number of iterations for p around 0.8, similar to the randomly connected networks of size $d = 20$.

3.3.3 PERFORMANCE BY NETWORK LOAD

The arrival intensities are varied to examine how the network load affects the simulation time. For fixed $\mu_k(x)$, $\beta_k(x)$ is varied to obtain predetermined traffic loads ρ. The networks are fully connected $d = 20$ node networks with uniform routing. That is, the d-node SBE network X has the set of parameters

$$b_k = 100 \quad \beta_k(x) = \frac{3}{d}\rho \mathbf{1}_{\{x_k \leqslant b_k\}} \quad \mu_k(x) = 3.0 \cdot \mathbf{1}_{\{x_k > 0\}} \quad p_{km} = \frac{\mathbf{1}_{\{k \neq m\}}}{d}$$

for $k, m \in \{1, \ldots, d\}$ and some constant $\rho > 0$. That is, the network structure is a complete graph with d vertices. The routing probabilities are chosen so that units, which depart node k, choose their destination uniformly between the remaining $d - 1$ nodes and the outside. There is no direct

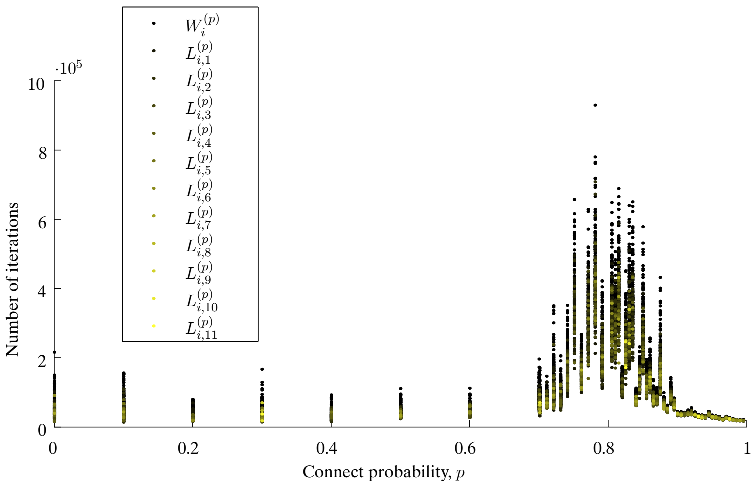

Figure 3.11. *The colours indicate at which mapping the particular observation of the number of iterations was obtained. No sample required more than 12 consecutive mappings,* $N_i^{(p)} \leqslant 12$.

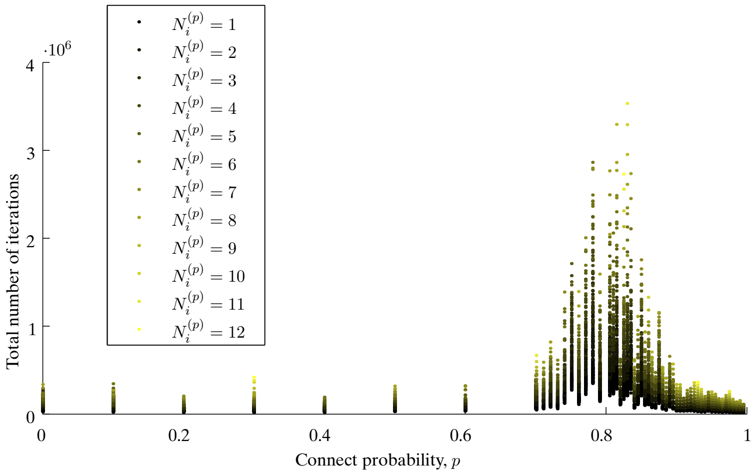

Figure 3.12. *The total number of iterations* $S_i^{(p)}$ *for the different connection probabilities. For fixed* p*, the samples of the total number of iterations form groups of those obtained at the first, second, third, and so on, mapping.*

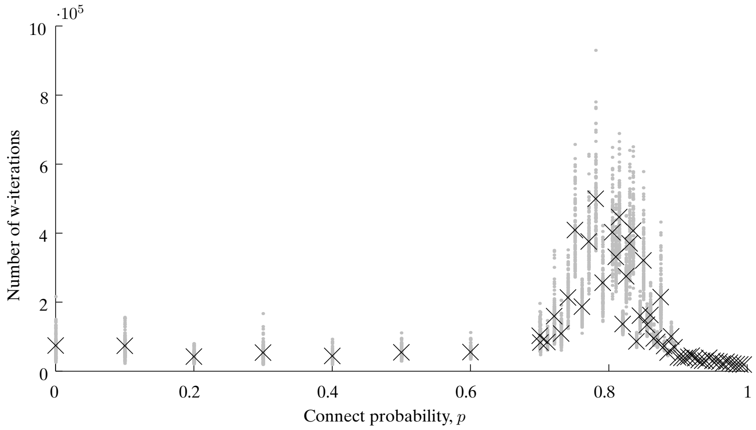

Figure 3.13. *The number of w-iterations $W_i^{(p)}$ for the different connection probabilities (light dots). The average value $m_w(p)$ is shown as thin crosses.*

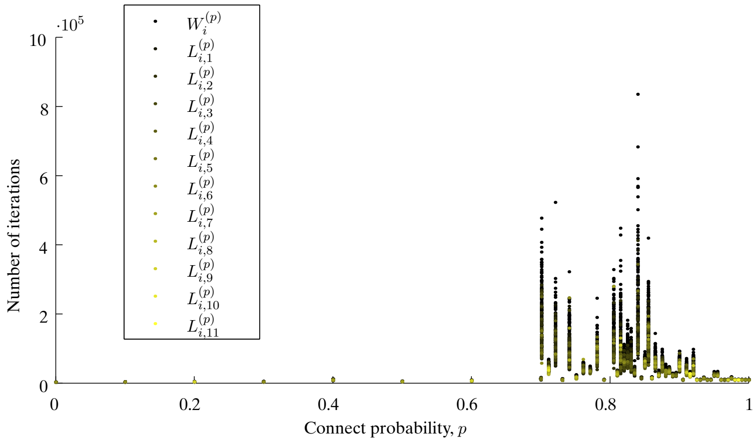

Figure 3.14. *The colours indicate at which mapping the particular observation of the number of iterations was obtained. Again, no sample required more than 12 consecutive mappings, $Ni^{(p)} \leqslant 12$. The network size was $d = 5$.*

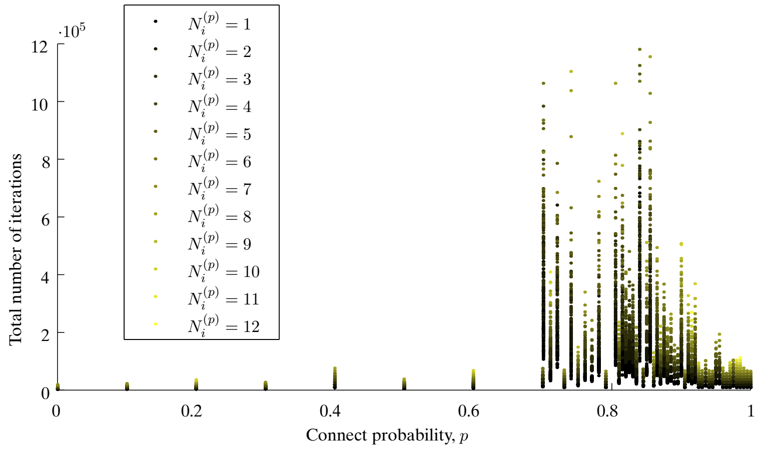

Figure 3.15. *The total number of iterations $S_i^{(p)}$ for the different connection probabilities and $d = 5$. For fixed p the samples of the total number of iterations form groups of those obtained at the first, second, third, and so on, mapping.*

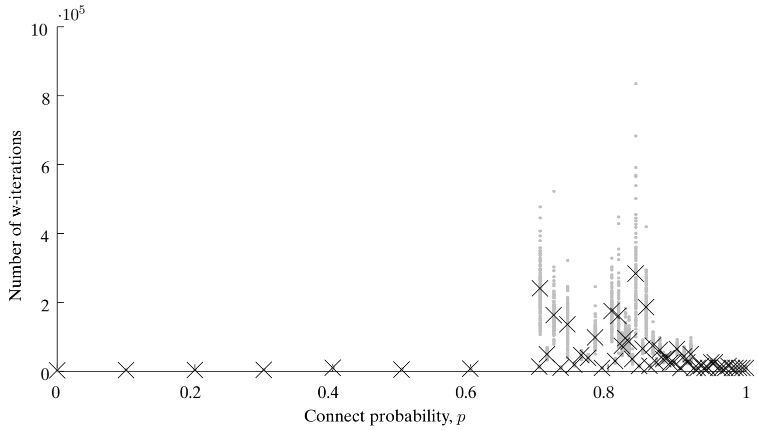

Figure 3.16. *The number of w-iterations $W_i^{(p)}$ for the different connection probabilities (light dots) when $d = 5$. The average value $m_w(p)$ is shown as thin crosses.*

rerouting of units, $p_{kk} = 0$, $k = 1, \ldots, d$. The blocking policy is given by

$$a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}}, \qquad c_{km}(x) = 0,$$

meaning that, units are persistent and accepted at a node if they fit in the buffer. The traffic load ρ, common to all nodes, is varied, and measurements on the number of iterations are made. The rate of events in the network, λ, increases with ρ, but the number of transitions do not depend on the actual time-scale.

An interesting scenario is the load-balanced setting, $\rho = 1$. If $0 < x_k < b_k$, $k = 1, \ldots, d$, the intensity at which units enter node k, Λ_k, satisfies

$$\Lambda_k = \beta_k(x) + \sum_{\substack{m=1,\ldots,d \\ m \neq k}} \Lambda_m p_{mk} = \rho \frac{3}{d} + (d-1)\Lambda_k \frac{1}{d} \quad \Rightarrow \quad \Lambda_k = 3.0\rho,$$

and thus, when $\rho = 1$, Λ_k balances the service rate, μ_k.

Let $d = 20$ and, for each value of ρ, let $X_1^{(\rho)}, \ldots, X_n^{(\rho)}$ be $n = 100$ independent perfect samples on the number of units in the network. Proceeding as in previous sections, samples on the number of iterations are taken as the traffic load ρ is varied from 0.1 (low intensity traffic) to 10.0 (heavily overloaded nodes). The samples $W_i^{(\rho)}$ and $L_{ij}^{(\rho)}$ can be found in Figure 3.17. The number of w- and l-iterations decreases for unbalanced loads, and the largest number of iterations is obtained when $\rho = 1$. When $\rho < 1$ more iterations are needed to obtain a coalescence of a mapping than when having a high traffic load, $\rho > 1$. The averages $m_w(\rho)$ can be found in Table 3.19 and Figure 3.18 shows the total number of iterations, $S_i^{(\rho)}$.

When the desistance probability is changed to $c_{km}(x) = 1$, a repetition of the previous measurements gave the samples of w- and l-iterations shown in Figure 3.20. The estimates of the expected number of w-iterations, $m_w(\rho)$, are also collected in Table 3.19. There is still a peak in the number of iterations for load-balanced nodes, $\rho = 1$, but for ρ close to 1 the number of iterations is higher for $\rho > 1$ than $\rho < 1$. Also, the variance of $W^{(\rho)}$ is higher for $\rho > 1$ than for $\rho < 1$. A comparison of m_w for $c = 0$ and $c = 1$ reveals that m_w is more sensitive to the traffic load when $c = 0$; the span $m_w(1/10)$ to $m_w(1)$ is greater when $c = 0$ than when $c = 1$.

It is not true that the traffic load ρ, by itself, can be used to explain the number of iterations needed for coalescence of the mappings. Let $d = 20$ and the networks be randomly connected (see Section 3.3.2). When node

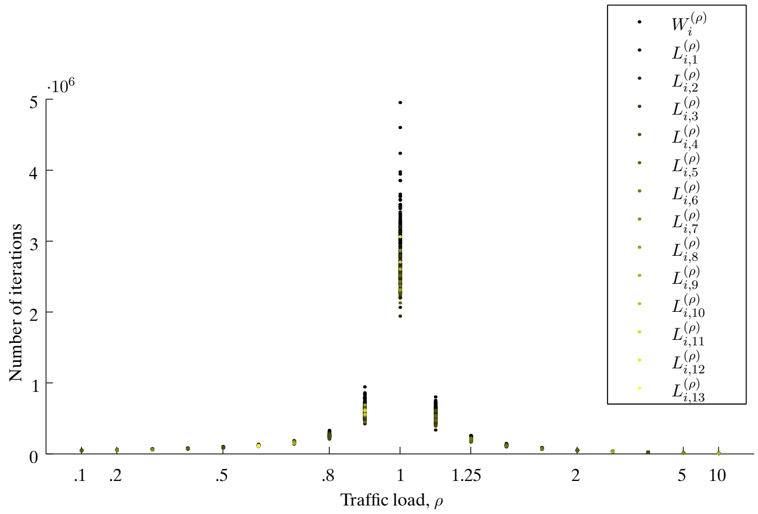

Figure 3.17. *The colours indicate at which mapping the particular observation of the number of iterations was obtained. No sample required more than 14 consecutive mappings,* $N_i^{(\rho)} \leqslant 14$.

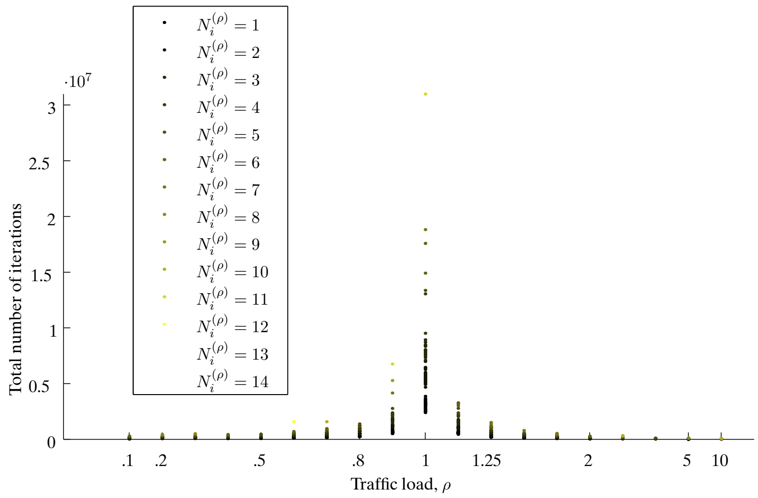

Figure 3.18. *The total number of iterations* $S_i^{(\rho)}$ *for values of the traffic load* ρ.

Table 3.19. Summary of the average number of w-iterations for different traffic loads ρ. There is a slight asymmetry between underloaded systems $\rho < 1$ and overloaded ditto, $\rho > 1$.

		$c_{km}(x) = 0$		$c_{km}(x) = 1$	
ρ	ρ^{-1}	$m_w(\rho)$	$m_w(\rho^{-1})$	$m_w(\rho)$	$m_w(\rho^{-1})$
.10	10/1	49162	7931.1	41380	10766
.20	10/2	56021	14619	46978	23070
.30	10/3	65344	23265	53024	42331
.40	10/4	76845	34958	61880	67850
.50	10/5	93859	51316	74562	$1.0594{\cdot}10^5$
.60	10/6	$1.2085{\cdot}10^5$	77519	93343	$1.5359{\cdot}10^5$
.70	10/7	$1.6704{\cdot}10^5$	$1.2278{\cdot}10^5$	$1.2633{\cdot}10^5$	$2.0446{\cdot}10^5$
.80	10/8	$2.6959{\cdot}10^5$	$2.2158{\cdot}10^5$	$1.9816{\cdot}10^5$	$2.8367{\cdot}10^5$
.90	10/9	$6.4857{\cdot}10^5$	$5.8963{\cdot}10^5$	$4.5238{\cdot}10^5$	$4.0781{\cdot}10^5$
1	1	$3.0604 \cdot 10^6$		$6.6519 \cdot 10^5$	

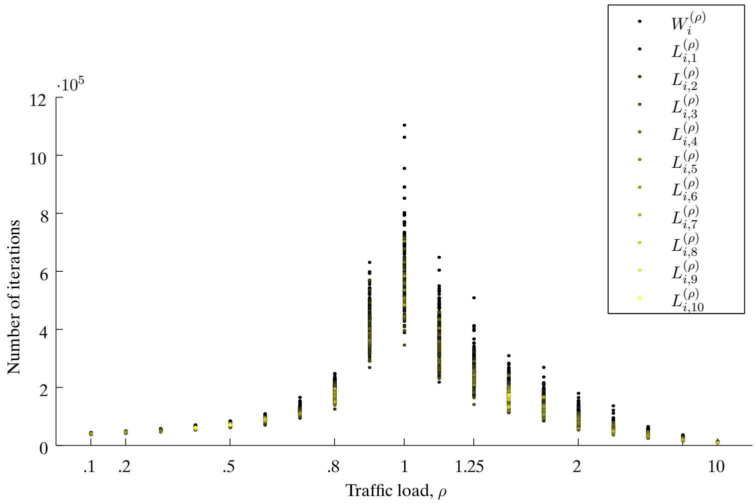

Figure 3.20. *The colours indicate at which phase the particular observation of the number of iterations was obtained. No sample required more than 11 consecutive mappings,* $N_i^{(\rho)} \leqslant 11$*. Desistance probability* $c_{km}(x) = 1$.

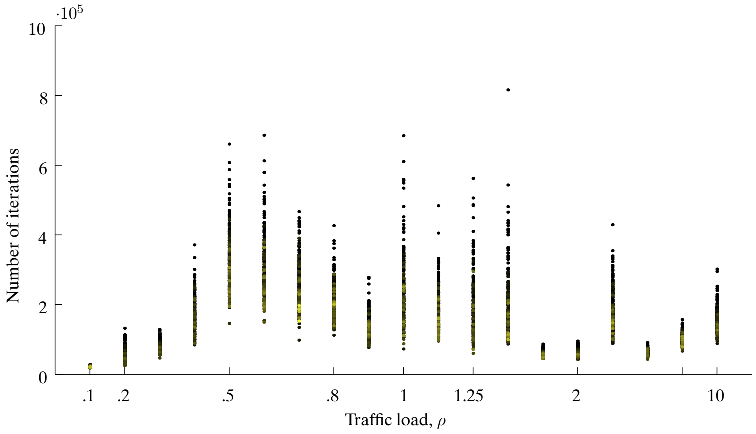

Figure 3.21. *The number of w- and l-transitions for a randomly connected network with* $d = 20$ *nodes. The colours indicate at which mapping the particular observation of the number of iterations was obtained. No sample required more than 13 consecutive mappings,* $N_i^{(\rho)} \leqslant 13$. *Desistance probability* $c_{km}(x) = 0$ *and connectivity probability* $p = 0.75$.

m has a binomial number of nodes, M_m, with a direct route to it, and the arrival rate to node m is $\beta_m(x) = 3\rho\mathbf{1}_{\{x_m<b_m\}}(d - M_m)/d$, the intensity of arrivals to node m, when $0 < x_m < b_m$, is $\Lambda_m = 3\rho$, irrespective of M_m. Samples on $W_1^{(\rho)}, \ldots, W_n^{(\rho)}$, and corresponding for l-iterations, each from an independent random network with connectivity probability $p = 0.75$, are shown in Figure 3.21 for sample size $n = 100$. Here, no sharp increase in the number of iterations for evenly balanced nodes, $\rho = 1$, can be seen, but a rather erratic appearance. Different loads give different convergence times in a non-predictable way.

3.3.4 PERFORMANCE BY BUFFER SIZES

How the restrictions on the queue lengths affect the simulation performance is investigated next. Let $d = 20$ and the SBE network X have the set of parameters

$$b_k = b \quad \beta_k(x) = \frac{2.0}{d}\mathbf{1}_{\{x_k \leqslant b_k\}} \quad \mu_k(x) = 3.0 \cdot \mathbf{1}_{\{x_k>0\}} \quad p_{km} = \frac{\mathbf{1}_{\{k\neq m\}}}{d}$$

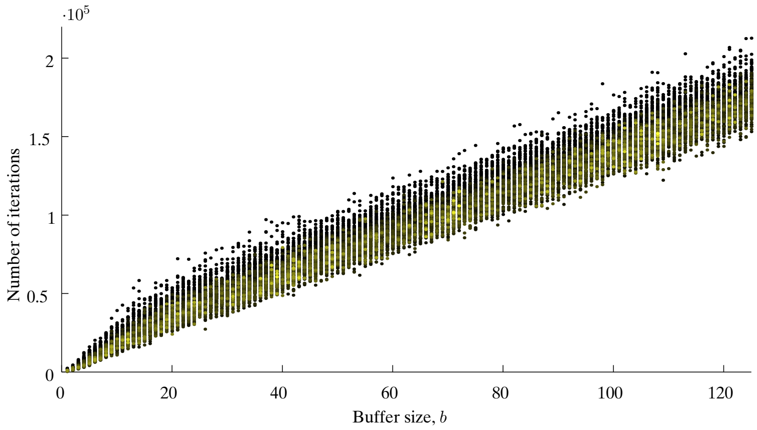

Figure 3.22. *The colours indicate at which mapping the particular observation of the number of iterations was obtained.*

for $k, m \in \{1, \dots, d\}$ and some constant $b \in \mathbb{N}$. Units are persistent and blocked when they cannot fit into a buffer. Here, all buffers have size b. For each value of b between 1 and 125, $X_1^{(b)}, \dots, X_n^{(b)}$ is $n = 100$ samples from the stationary distribution. The number of w- and l-iterations associated with the perfect samples are shown in Figure 3.22. The number of iterations increase with the buffer sizes, and for sufficiently large buffers, in a seemingly linear way. In Figure 3.23 the samples of total number of iterations $S_i^{(b)}$ are shown.

The linear increase in iterations for large buffer sizes is clearly visible in Figure 3.24 where the observations $W_1^{(b)}, \dots, W_{100}^{(b)}$ are shown together with the averages $m_w(b)$ and $m_l(b)$.

To analyse how restrictions of the queue lengths at the nodes affect the simulation performance, networks are constructed in the following way. Let $B = (B_1, \dots, B_d)$ be a vector of independent random variables, uniformly distributed on $\{1, \dots, 100\}$. Let the $d = 20$-node SBE network have buffer size $b_k = B_k$ at node $k = 1, \dots, d$. The $n = 1000$ samples of $W^{(B)}$, with new values of B for each sample, are shown in Figure 3.25 as a function of m_B;

$$m_B = \frac{1}{d}\sum_{k=1}^{d} B_k \qquad \text{and} \qquad s_B^2 = \frac{1}{d-1}\sum_{k=1}^{d}(B_k - m_B)^2.$$

The number of w-iterations increases with the average buffer size in the net-

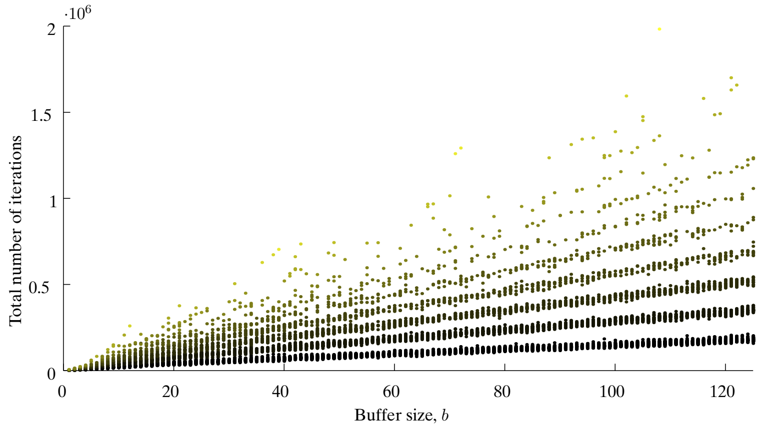

Figure 3.23. *The total number of iterations $S_i^{(b)}$ for the buffer sizes b. Again, the samples of total number of iterations are clustered in groups of those obtained at the first, second, third, and so on, mapping.*

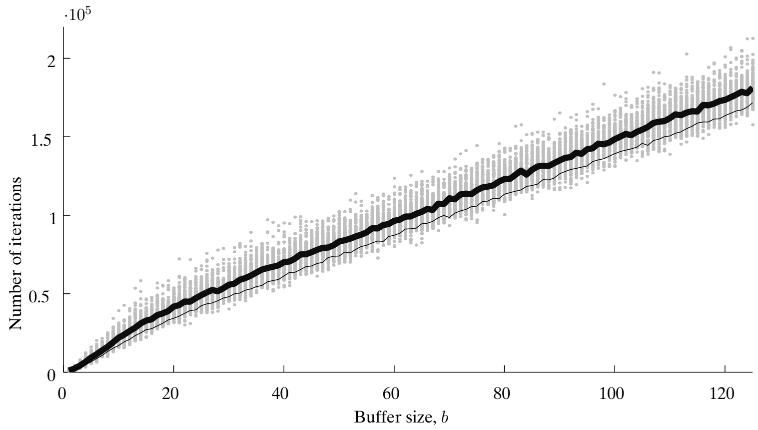

Figure 3.24. *The number of w-iterations $W_i^{(b)}$ shown as dots. The thick line is the average $m_w(b)$ belonging to the observations. The average $m_l(b)$, the average number of l-iterations, is shown as a thin line.*

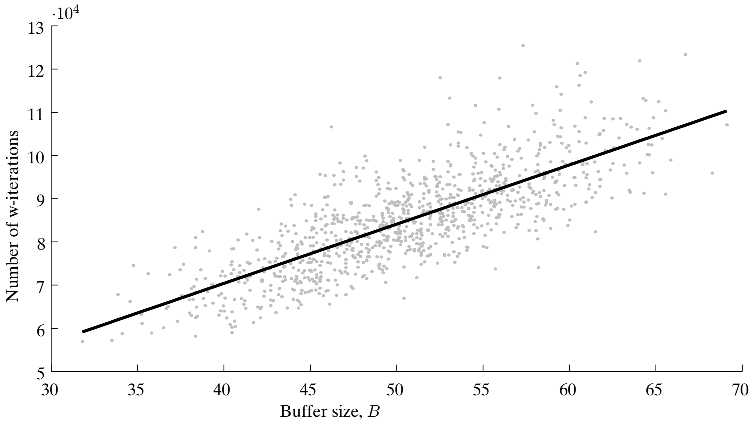

Figure 3.25. *The number of w-iterations* $W_i^{(B)}$ *shown as dots. The line is a least square fitted line* $m_w(m_B)$*. Here* $n = 1000$.

work, and in Figure 3.25 a least square fitted line, $m_w(m_B)$, is also shown. The observations $W_1, \ldots, W_{1000}$ versus s_B are plotted in Figure 3.26 and no correlation can be seen. The conclusion drawn is that discrepancies in buffer sizes do not contribute significantly to the number of iterations. What matters is the average buffer size of the network, which affect the number of iterations in a linear fashion.

3.3.5 CONCLUSIONS ON PERFORMANCE

Based on simulation, the following growth rates in the number of iterations before a sample on the stationary distribution of a network is obtained, are conjectured.

For fully connected networks with uniform routing the expected number of iterations in a winning and losing mappings grows quadratically in d, the dimension of the network. The variance of the number of iterations increased with the network size as well. The simulation time was larger for unbalanced traffic, and was larger for underloaded settings when units were persistent. When $c = 1$, overloaded settings may need more iterations than underloaded. The average buffer size of the network affect the number of iterations in a linear fashion, but discrepancies between buffer sizes at different nodes do not significantly affect the number of iterations.

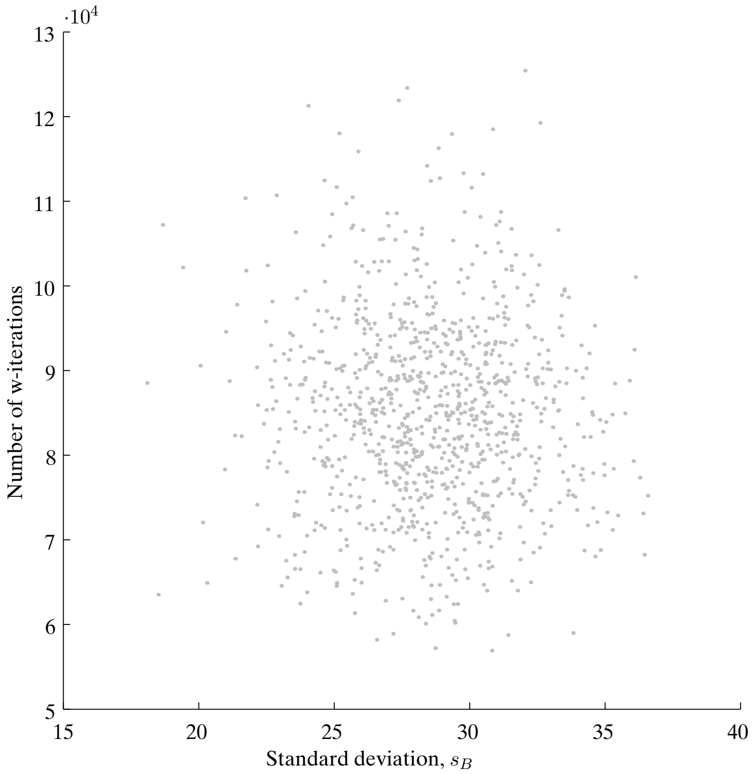

Figure 3.26. *The number of w-iterations* $W_i^{(s_B)}$ *shown as dots. Here* $n = 1000$*.*

Leaving the fully connected networks and considering randomly connected networks, some connection probabilities gave a significant increase in the number of iterations for the sizes of networks under study. Varying the traffic load gave different simulation times in a non-predictable way.

Although the number of iterations grew in a non-predictable way, the number of iterations were moderate, so even large randomly connected networks could be simulated in a reasonable amount of time.

3.4 COMMENTS ON THE IMPLEMENTATION

Some remarks should be made about the implementation of the revised Propp–Wilson algorithm.

In stochastic simulation studies there is always a note on generation of pseudo-random numbers, and now it is time for ours. The SPRNG (Scalable Parallel Pseudo Random Number Generator) library version 2.0 was used for generating pseudo-random numbers. The streams from the generator have passed all of a series of standard statistical tests for randomness. Further information and documents on the performance are available at the www-site SPRNG [25].

The simulation package was written in Objective C compiled with the GNU C compiler 3.01 and the programme was run on various platforms and computer architectures. The programme package and libraries will be available at URL `www.math.chalmers.se/~danm/research`.

Section 3.3 analyses the performance of the revised Propp–Wilson algorithm measured in the number of iterations. The actual CPU time is highly dependent on the computer platform the simulations are run on. As an indication of the computation time, we mention that the 64 scenarios in Section 4.1 each with 10 000 samples, thus a total of 640 000 samples, took nearly 6 hours on a three year old PC-equipment(450MHz PII), already outdated for some of the more demanding applications. Some parameter settings take significantly longer time than others. The $11 \times 10\,000$ samples in Section 4.1.1 took 15 minutes on a 800MHz PIII PC-equipment, the ones in Section 4.1.2 12 minutes.

Chapter 4

Simulation experiments of SBE networks

By using the revised Propp–Wilson algorithm to obtain perfect samples of the number of units, the properties of a network can be studied. To illustrate this, a network model is devised and analysed. It is not the method of analysis that is of particular interest. The network model, which is not to be considered as a model of a real, existing network, is used simply to indicate what can be achieved by simulation.

The chapter is descriptive in its nature, and the mathematical level is kept relatively basic. Simulations carried out are divided into two sections, in the first of which the particular network model, used throughout the chapter, is established. The parameters of the model are varied, and the effect on the stationary number of units is analysed. The different configurations of parameter settings are called scenarios; following the main subsection, selected scenarios are investigated in a bit more detail to discern some of the properties of the network. The performance of the network is also compared with that of an open Jackson network, and it is shown that, at some parameter settings, the stationary distribution may be approximated by the distributions for an open Jackson network.

The second part deals with other queueing-theoretical quantities, namely the waiting and sojourn times discussed in Section 2.4.2. Samples of these are constructed from trajectories of the network evolution started in equilibrium; they are analysed for a collection of parameter settings.

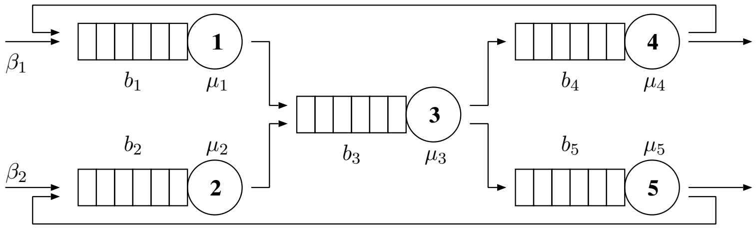

Figure 4.1. *A five node network where units arrive at two entrance nodes, one and two, and after service are sent to a common node, three. When departing this central node, units are independently routed to the last two nodes, four and five, from which the units leave the network or are routed back to the entrance nodes.*

4.1 ANALYSIS BASED ON SAMPLES

Consider the network depicted in Figure 4.1, which is a specific network analysed for different values of its parameters. In each configuration, $n = 10\,000$ samples are taken from the stationary distribution of the number of units in the network. From samples, $X_1, \dots, X_n$, the expected numbers of units at node i, $\mathsf{E}[<X, e_i>]$, for $i = 1, \dots, 5$, are estimated from averages, with observed values denoted m_i. The estimate of the expected total number of units in the network is denoted $S(m)$.

The acceptance probabilities are given by $a_{km}(x) = \mathbf{1}_{\{x_m < b_m\}}$, and the desistance probabilities satisfy $c_{km}(x) = c$. The other parameters are

$$
\begin{array}{rlrlrl}
b_1 &= 3 \text{ or } 15 & & & b_4 &= 10 \\
\beta_1(x) &= 2(1-p)\mathbf{1}_{\{x_1 < b_1\}} & & & \mu_4(x) &= 2.5 \cdot \mathbf{1}_{\{x_4 > 0\}} \\
\mu_1(x) &= \mu_1 \cdot \mathbf{1}_{\{x_1 > 0\}} & & & p_{41} &= p \\
p_{13} &= 1 & b_3 &= 5 \text{ or } 30 & p_4 &= 1-p \\
& & \mu_3(x) &= \mu_3 \cdot \mathbf{1}_{\{x_3 > 0\}} & & \\
b_2 &= b_1 & p_{34} = p_{35} &= 0.5 & b_5 &= 10 \\
\beta_2(x) &= \beta_1(x) & & & \mu_5(x) &= 2.5 \cdot \mathbf{1}_{\{x_5 > 0\}} \\
\mu_2(x) &= \mu_1(x) & & & p_{52} &= p \\
p_{23} &= 1 & & & p_5 &= 1-p
\end{array}
$$

The network is symmetric: the entrance nodes, one and two, have the same service rate, buffer size, and intensity of the arrival process. The same is true

Table 4.2. The theoretical values for the Jackson network which exist when $\rho = 0.80$.

	b_1	μ_1	b_3	μ_3	m_1	m_2	m_3	m_4	m_5	$S(m)$
J	∞	$\uparrow$	∞	$\uparrow$	5.000	5.000	5.000	5.000	5.000	25.000

for the last two nodes. The buffer sizes b_1 (b_2) and b_3, and the service rates μ_1 (μ_2) and μ_3, are varied for different values of the desistance probability c and routing probability p.

The arrival rates $\beta_1(x)$ and $\beta_2(x)$ are chosen so that, in the corresponding Jackson network, the arrival rate is constant for all values of p. To see this, the linear system of equations (2.3) is solved,

$$\begin{aligned}
\Lambda_1 &= 2(1-p) + \Lambda_4 p \\
\Lambda_2 &= 2(1-p) + \Lambda_5 p \\
\Lambda_3 &= \Lambda_1 + \Lambda_2 \\
\Lambda_4 &= \Lambda_3/2 \\
\Lambda_5 &= \Lambda_3/2,
\end{aligned}$$

to obtain $\Lambda_1 = \Lambda_2 = \Lambda_4 = \Lambda_5 = 2$, and $\Lambda_3 = 4$. With $\rho = 0.80$ and $\mu_i(x) = \mathbf{1}_{\{x_i>0\}}\Lambda_i/\rho$, the stationary distribution of node i in the corresponding Jackson network is, using (2.6),

$$\pi_i^J(x) = (1-\rho)\rho^x, \quad x \in \mathbb{Z}_+,$$

and the expected number of units is equal to 5 at all nodes.

There are two settings for each of the six parameters under study, which yield 64 scenarios. In Table 4.3, split in four parts (on pages 104–105), the results from simulation experiments are summarised. The buffer sizes $b_1 = 3$ and $b_1 = 15$ are abbreviated $\downarrow$ and $\uparrow$, indicating a low and a high setting, respectively; the same applies to b_3. The two values of μ_1 and μ_3 satisfy $\mu_i = \Lambda_i/\rho$ for $\rho = 0.80$ (high service rate) and $\rho = 1.2$ (low service rate), also indicated by $\uparrow$ and $\downarrow$, respectively. The theoretical results for the Jackson network are collected in Table 4.2.

Denote $S(m)$ in scenario i by S_i. To analyse the impact of the parameters, a performance measure is devised. For this the relative difference in total number of units in the network is measured and used, i.e., quotients S_i/S_j are examined for scenarios i and j. Performance (impact) measures may be chosen in many ways, e.g. absolute difference $|S_i - S_j|$, but the quotient is reasonably illustrative for this simple analysis.

Table 4.3.a. $p = 0.90,\ c = 1.00$

	b_1	μ_1	b_3	μ_3	m_1	m_2	m_3	m_4	m_5	$S(m)$
1	↑	↓	↑	↓	6.318	6.320	7.625	1.480	1.485	23.229
2	↓	↓	↑	↓	1.022	1.038	1.207	0.613	0.616	4.495
3	↑	↓	↓	↓	2.726	2.746	1.773	0.850	0.855	8.950
4	↓	↓	↓	↓	1.010	1.000	1.151	0.589	0.590	4.340
5	↑	↑	↑	↓	2.014	2.054	18.802	1.857	1.844	26.571
6	↓	↑	↑	↓	0.765	0.781	2.050	0.873	0.874	5.343
7	↑	↑	↓	↓	1.047	1.075	1.923	0.922	0.923	5.890
8	↓	↑	↓	↓	0.736	0.725	1.509	0.780	0.753	4.503
9	↑	↓	↑	↑	6.872	6.786	1.606	1.542	1.539	18.345
10	↓	↓	↑	↑	1.058	1.063	0.633	0.634	0.615	4.004
11	↑	↓	↓	↑	5.125	5.166	1.162	1.275	1.265	13.993
12	↓	↓	↓	↑	1.052	1.046	0.620	0.624	0.614	3.956
13	↑	↑	↑	↑	2.573	2.558	2.622	2.382	2.389	12.525
14	↓	↑	↑	↑	0.785	0.798	0.893	0.882	0.890	4.249
15	↑	↑	↓	↑	1.564	1.580	1.275	1.409	1.412	7.239
16	↓	↑	↓	↑	0.789	0.782	0.848	0.872	0.863	4.153

Table 4.3.b. $p = 0.00,\ c = 1.00$

	b_1	μ_1	b_3	μ_3	m_1	m_2	m_3	m_4	m_5	$S(m)$
17	↑	↓	↑	↓	10.942	10.922	14.163	1.706	1.690	39.423
18	↓	↓	↑	↓	1.736	1.720	3.753	1.134	1.110	9.453
19	↑	↓	↓	↓	10.884	10.936	2.461	1.164	1.153	26.598
20	↓	↓	↓	↓	1.741	1.721	1.924	0.949	0.938	7.273
21	↑	↑	↑	↓	3.527	3.507	24.995	1.880	1.845	35.755
22	↓	↑	↑	↓	1.237	1.222	14.277	1.729	1.706	20.171
23	↑	↑	↓	↓	3.555	3.511	2.975	1.397	1.404	12.842
24	↓	↑	↓	↓	1.208	1.223	2.555	1.183	1.200	7.369
25	↑	↓	↑	↑	10.952	10.976	1.877	1.803	1.820	27.427
26	↓	↓	↑	↑	1.739	1.743	1.062	1.100	1.121	6.765
27	↑	↓	↓	↑	10.883	10.908	1.395	1.577	1.569	26.332
28	↓	↓	↓	↑	1.731	1.740	0.997	1.041	1.079	6.588
29	↑	↑	↑	↑	3.453	3.523	3.789	2.892	2.903	16.559
30	↓	↑	↑	↑	1.220	1.216	1.734	1.790	1.751	7.711
31	↑	↑	↓	↑	3.502	3.565	1.819	2.262	2.206	13.353
32	↓	↑	↓	↑	1.213	1.218	1.385	1.586	1.606	7.007

Table 4.3.c. $p = 0.90,\ c = 0.00$

	b_1	μ_1	b_3	μ_3	m_1	m_2	m_3	m_4	m_5	$S(m)$
33	↑	↓	↑	↓	8.420	8.490	12.225	2.498	2.474	34.108
34	↓	↓	↑	↓	1.387	1.361	2.042	1.576	1.587	7.953
35	↑	↓	↓	↓	9.992	9.987	2.457	2.106	2.100	26.640
36	↓	↓	↓	↓	1.417	1.436	1.627	1.475	1.518	7.473
37	↑	↑	↑	↓	8.211	8.173	27.636	2.561	2.520	49.102
38	↓	↑	↑	↓	1.060	1.049	6.231	2.115	2.109	12.563
39	↑	↑	↓	↓	8.976	8.913	3.517	2.394	2.324	26.123
40	↓	↑	↓	↓	1.223	1.210	2.394	1.831	1.807	8.466
41	↑	↓	↑	↑	8.600	8.550	1.801	2.439	2.423	23.812
42	↓	↓	↑	↑	1.396	1.396	0.847	1.548	1.546	6.733
43	↑	↓	↓	↑	9.011	8.989	1.371	2.287	2.298	23.957
44	↓	↓	↓	↑	1.411	1.396	0.807	1.528	1.552	6.694
45	↑	↑	↑	↑	3.357	3.374	4.659	3.196	3.233	17.818
46	↓	↑	↑	↑	1.052	1.062	1.340	2.076	2.060	7.590
47	↑	↑	↓	↑	5.349	5.276	2.143	2.894	2.894	18.557
48	↓	↑	↓	↑	1.087	1.090	1.177	2.042	1.981	7.378

Table 4.3.d. $p = 0.00,\ c = 0.00$

	b_1	μ_1	b_3	μ_3	m_1	m_2	m_3	m_4	m_5	$S(m)$
49	↑	↓	↑	↓	11.367	11.358	14.655	1.762	1.716	40.857
50	↓	↓	↑	↓	1.745	1.724	3.691	1.123	1.131	9.414
51	↑	↓	↓	↓	12.891	12.936	2.498	1.182	1.194	30.701
52	↓	↓	↓	↓	1.843	1.840	2.029	0.973	0.964	7.649
53	↑	↑	↑	↓	10.691	10.716	27.906	1.885	1.848	53.045
54	↓	↑	↑	↓	1.271	1.281	15.240	1.752	1.714	21.259
55	↑	↑	↓	↓	11.650	11.667	3.558	1.616	1.629	30.120
56	↓	↑	↓	↓	1.574	1.552	2.806	1.281	1.301	8.514
57	↑	↓	↑	↑	11.015	10.897	1.967	1.804	1.844	27.526
58	↓	↓	↑	↑	1.720	1.725	1.058	1.140	1.127	6.769
59	↑	↓	↓	↑	11.719	11.653	1.429	1.604	1.603	28.008
60	↓	↓	↓	↑	1.736	1.754	1.022	1.099	1.104	6.715
61	↑	↑	↑	↑	3.592	3.595	4.695	3.191	3.167	18.239
62	↓	↑	↑	↑	1.230	1.233	1.782	1.778	1.803	7.826
63	↑	↑	↓	↑	6.083	6.086	2.237	2.782	2.825	20.013
64	↓	↑	↓	↑	1.290	1.256	1.486	1.667	1.692	7.391

The actual values of S_i are not very illuminating in themselves, since they are highly dependent on the whole collection of parameters under consideration; in three out of four scenarios, the parameters correspond to overloaded nodes. In these scenarios one may, presumably, increase the number of units, $S(m)$, by increasing the buffer sizes. 'Presumably' since this, as will be seen, is not true, not even in an underloaded setting. However tempting, it is not justified to order the parameters in general with respect to the impact measure. By the same argument, claiming one of the parameters to be of utmost importance is valid only with respect to the other possible configurations of parameters and the network topology under study.

First, the network is compared with a Jackson network. The stationary distribution of the corresponding open Jackson network exists only if $\rho < 1$; hence, only when both μ_1 and μ_3 are large, can $m_1, \ldots, m_5$ in the SBE network be compared with the ones in the Jackson network so that the effect of finite buffers can be examined. The total number of units in the SBE network for $\rho = 0.80$, (scenarios 13–16, 29–32, 45–48, 61–64) is always below the total number in the comparable Jackson network. Scenarios 13, 29, 45 and 61 have large buffer sizes, which means the SBE networks are as close to a Jackson network as possible with the current choice of parameters. Note, however, that in scenarios 47 and 63, there is a larger number of units at the entrance nodes, m_1 and m_2, than in the corresponding Jackson network. In these scenarios the desistance probability $c = 0$, so that units are persistent and remain at the entrance nodes when blocked at the middle node. Even though the service rates are high at all nodes, the centre node has a small buffer and is repeatedly saturated; the large buffers at the entrance nodes are filled with units waiting to be transferred. In scenarios 15 and 31, where $c = 1$, this phenomenon cannot be seen.

When b_3 is large, or both b_1 and b_3 are small, the difference between $c = 0$ and $c = 1$ for the feed-forward network is small. When $p = 0.90$, a change of c from 0 to 1 gives a substantially lower number of units at all the nodes. Actually, it is the combination of parameters $p = 0.90$ and $c = 1$ that produces an exceptionally low number of units. Heuristically, the effect of the desistance probability is significant only when blocking occurs on a large scale, which happens in the rerouting scenarios, $p = 0.9$.

One can see that the buffer sizes at the entrance nodes, b_1, heavily affect the number of units at every node of the network, in particular the total number of units in the network; larger buffers at the entrance nodes, $b_1 = 15$ ($\uparrow$), give, not surprisingly, a larger number of units. When μ_1 is small,

Table 4.4. The impact of b_1. For scenarios i and $j = i+1$, S_i/S_j measures the relative total number of units in the network between settings of $b_1 = 15$ and $b_1 = 3$.

i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j
1	**2**	5.1681	**25**	**26**	4.0546	**11**	**12**	3.5368	**61**	**62**	2.3305
5	**6**	4.9728	**51**	**52**	4.0135	**41**	**42**	3.5365	**29**	**30**	2.1476
9	**10**	4.5820	**27**	**28**	3.9973	**39**	**40**	3.0855	**3**	**4**	2.0622
49	**50**	4.3402	**37**	**38**	3.9084	**13**	**14**	2.9480	**31**	**32**	1.9057
33	**34**	4.2885	**19**	**20**	3.6571	**63**	**64**	2.7075	**21**	**22**	1.7726
59	**60**	4.1713	**43**	**44**	3.5790	**47**	**48**	2.5152	**15**	**16**	1.7432
17	**18**	4.1703	**35**	**36**	3.5647	**53**	**54**	2.4952	**23**	**24**	1.7428
57	**58**	4.0662	**55**	**56**	3.5374	**45**	**46**	2.3476	**7**	**8**	1.3081

the entrance nodes are overloaded and a change of b_1 alters the number of units in the network quite drastically. A curious result is that, even though μ_1 is large in scenario 5 (hence the entrance nodes are underloaded, and units leave the network when blocked), it has not only the greatest total number of units, $S(m)$, of all the networks with rerouting and $c = 1$, but also the second largest impact of b_1 as measured by S_5/S_6.

When forming quotients S_i/S_{i+1}, for i odd, and measuring the proportions of the total number of units in networks that differ only in b_1, it can be seen that the impact of b_1 is largest in scenario pairs 1/2, 5/6 and 9/10. These are overloaded scenarios, where at least one of μ_1 or μ_3 are small, b_3 is large, rerouting occurs and units failing to transfer leave the network. Again, note that only relative differences are measured: there are more units in the network when $c = 0$, but the difference in $S(m)$ between $b_1 = 15$ and $b_1 = 3$ is not as large. The quotients S_i/S_{i+1} are given in Table 4.4.

The nine scenarios with the largest differences in S_i/S_{i+1} have, with one exception, a large buffer at the centre node. The exception, scenario pair 59/60, has a small service rate at the entrance nodes, a high value of μ_3, no rerouting and persistent units, $c = 0$. Corollary 2.5.9 states that when $c = 1$ the number of units is stochastically increasing in the buffer sizes, but here, an increase of b_1 gave rise to the total number of units in all scenarios. The smallest impact S_i/S_j is obtained in scenario pairs 7/8 and 23/24. The greatest number of units at the entrance nodes is obtained in scenario 51 where units are persistent, all service rates are small, the buffer size b_3 is small and, naturally, b_1 is large.

When the buffer size at the middle node is examined, scenario pairs 5/7 and 21/23 have the greatest impact S_i/S_j on the total number of units in the

Table 4.5. The impact of b_3. For scenarios i and j, $S_i(m)/S_j(m)$ measures the relative total number of units in the network between settings of $b_3 = 30$ and $b_3 = 5$.

i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j
5	**7**	4.5111	**38**	**40**	1.4839	**6**	**8**	1.1866	**14**	**16**	1.0231
21	**23**	2.7842	**17**	**19**	1.4822	**30**	**32**	1.1004	**10**	**12**	1.0120
22	**24**	2.7374	**49**	**51**	1.3308	**34**	**36**	1.0642	**58**	**60**	1.0082
1	**3**	2.5953	**9**	**11**	1.3111	**62**	**64**	1.0588	**42**	**44**	1.0059
54	**56**	2.4968	**18**	**20**	1.2998	**25**	**27**	1.0416	**41**	**43**	0.9940
37	**39**	1.8796	**33**	**35**	1.2803	**2**	**4**	1.0356	**57**	**59**	0.9828
53	**55**	1.7611	**29**	**31**	1.2401	**46**	**48**	1.0288	**45**	**47**	0.9602
13	**15**	1.7301	**50**	**52**	1.2306	**26**	**28**	1.0269	**61**	**63**	0.9114

network. Here, the centre node is overloaded, μ_3 small, and the entrance nodes have large buffers and high service rates. Hence, there is a high probability that the centre node is saturated and the influence of b_3 is great. The difference between scenario pairs 5/7 and 21/23, is that routing occurs in the former. For scenario pair 21/23, there is only a minor difference in S_i/S_j in comparison with pair 22/24 for which b_1 is small, although the differences in the total number of units S_i are substantial.

The four scenario pairs with the greatest impact of b_3 have $c = 1$, but when μ_3 is small even scenarios with persistent units, $c = 0$, are substantially affected by b_3. When $c = 1$, $S(m)$ is sensitive to blocking, and blocking of units at the centre node is more frequent for a small buffer size b_3. The quotients S_i/S_j, for which i is a scenario where $b_3 = 30$ and j is the corresponding scenario with $b_3 = 5$, are given in Table 4.5.

When the service rate at the middle node is high, S_i/S_j is greatest in scenario pair 13/15, again with $c = 1$. In pairs 41/43, 57/59, 45/47 and 61/63, which are scenarios with desistance probability $c = 0$, the total number of units in the network is larger when b_3 is small, hence, $S_i/S_j < 1$. Scenario pairs 41/43 and 57/59 correspond to scenarios with overloaded entrance nodes. The average number of units at the centre node, m_3, is greatest in scenario 53 where the middle node is overloaded, units are persistent, all buffer sizes are large and the service rate at the entrance nodes is large; this means the entrance nodes are saturated with units waiting to be transferred.

When the effect of the service rates at the entrance nodes, as measured by S_i/S_j, is examined, one sees that it is greatest for scenario pair 19/23, where the the centre node has a small buffer and a low service rate. Furthermore, the desistance probability $c = 1$ forces blocked units to leave

Table 4.6. The impact of μ_1. It is the quotient between scenarios with low and high service rate, respectively.

i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j
19	**23**	2.0711	**41**	**45**	1.3364	**10**	**14**	0.9424	**1**	**5**	0.8742
27	**31**	1.9720	**43**	**47**	1.2910	**28**	**32**	0.9401	**58**	**62**	0.8649
11	**15**	1.9328	**17**	**21**	1.1026	**60**	**64**	0.9084	**2**	**6**	0.8412
25	**29**	1.6563	**35**	**39**	1.0198	**44**	**48**	0.9073	**49**	**53**	0.7702
3	**7**	1.5195	**51**	**55**	1.0193	**52**	**56**	0.8984	**33**	**37**	0.6946
57	**61**	1.5092	**20**	**24**	0.9870	**42**	**46**	0.8871	**34**	**38**	0.6331
9	**13**	1.4647	**4**	**8**	0.9638	**36**	**40**	0.8827	**18**	**22**	0.4687
59	**63**	1.3995	**12**	**16**	0.9527	**26**	**30**	0.8773	**50**	**54**	0.4428

the network. In scenario 19, the service rate μ_1 is small and the number of units present at the entrance nodes is large; changing μ_1 to 2.5 (scenario 23) reduces the number of units at these nodes. In these scenarios, the middle node is overloaded even for small values of μ_1, and thus, even though an increase in the service rate μ_1 gives a larger average number of units m_3, this increase does not compensate for the decrease in the number of units at the entrance nodes. However, the choice of μ_3 does not substantially affect S_i/S_j; compare S_{19}/S_{23} with S_{27}/S_{31} for which the centre node is underloaded. Actually, increasing μ_3 does indeed let more units flow through the centre node but, due to the small buffer size b_3, many units from the entrance nodes are still blocked for high service rates μ_1, thus reducing the total number of units as in the scenarios with a small μ_3. In Table 4.6, the quotients S_i/S_j are collected, where i belongs to a scenario with a small μ_1 and j to one with a large μ_1.

In the Jackson networks, the total number of units decreases with the service rates at the nodes of the network. For the particular SBE network under consideration, using Table 4.6, a majority of the scenarios have $S_i/S_j < 1$, which corresponds to settings for which the total number of units rose with the service rate. The most extreme examples correspond to scenarios in which the centre node has a large buffer and a low service rate. The only exception is scenario pair 17/21 where the network is a feed-forward network, $c = 1$ and the buffers at the entrance nodes are large.

The effect on the total number of units, measured for the two values of the service rate at the middle node, μ_3, is greatest for scenario pair 53/61. Scenario 53 has the largest average number of units at the centre node, m_3, as well as the highest total number of units in the network, $S(m)$. Here, the units are persistent in a feed-forward network with large buffers and a high service rate at the entrance nodes, i.e., when the service rate μ_3 is

Table 4.7. The impact of μ_3. The quotients between scenarios i and j with low and high service rate, respectively.

i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j
53	**61**	2.9083	**49**	**57**	1.4843	**34**	**42**	1.1812	**4**	**12**	1.0970
37	**45**	2.7557	**17**	**25**	1.4374	**56**	**64**	1.1519	**51**	**59**	1.0961
54	**62**	2.7163	**33**	**41**	1.4324	**40**	**48**	1.1476	**8**	**16**	1.0843
22	**30**	2.6160	**39**	**47**	1.4077	**52**	**60**	1.1392	**24**	**32**	1.0516
21	**29**	2.1592	**18**	**26**	1.3975	**2**	**10**	1.1226	**19**	**27**	1.0101
5	**13**	2.1214	**50**	**58**	1.3906	**36**	**44**	1.1165	**23**	**31**	0.9617
38	**46**	1.6552	**1**	**9**	1.2662	**35**	**43**	1.1120	**7**	**15**	0.8136
55	**63**	1.5050	**6**	**14**	1.2576	**20**	**28**	1.1040	**3**	**11**	0.6396

raised to obtain scenario 61, Scenario 53 becomes one of the Jackson-like ones. When μ_3 is small, the centre node is saturated as are the entrance nodes since units are persistent. When μ_3 is increased, the average numbers of units at all of these nodes are reduced. Table 4.7 contains a list of the quotients S_i/S_j, for which i is a scenario with a small μ_3 and j is one with a high μ_3.

There are scenarios with a total number of units that increases with the service rate: units are persistent and buffer size b_3 is small, but b_1 large. The exception is scenario pair 19/27, a feed-forward network with a small service rate μ_1, for which $S_{19}/S_{27} = 1.0101$ was obtained.

The total number of units, $S(m)$, is larger for the feed-forward networks, $p = 0$, than for the corresponding scenarios with $p = 0.90$. Hence, $S_i/S_j > 1$ when i belongs to a scenario with $p = 0$ and j to one with positive rerouting probability. The greatest contrast is obtained by comparing scenarios 6 and 22, for which the estimate $S(m)$ changed from $S_6 = 5.343$ to $S_{22} = 20.171$. These scenarios have small buffer sizes at the entrance nodes and low service rates at the middle node. The observed difference in the total number of units is due mainly to the increased number of units at the middle node, m_3, for the feed-forward network. The eight scenario pairs with the greatest impact measures, S_i/S_j, for the rerouting probability, are these with persistent units. Actually, of the sixteen largest, only one scenario pair has $c = 0$; consequently, only one of the lowest sixteen has $c = 1$. The conclusion drawn is that the number of units in the network is mainly affected by the routing probability, p, when blocked units leave the network. Another implication is that blocking occurs on a larger scale in the rerouting networks. For example, if scenarios 38 and 54 are compared, m_4 and m_5 are larger when $p = 0.90$, but the total $S(m)$ is still larger for the

Table 4.8. The impact measures S_i/S_j for scenarios differing by the value of p.

i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j
22	**6**	3.7751	**54**	**38**	1.6921	**29**	**13**	1.3221	**63**	**47**	1.0785
19	**3**	2.9718	**26**	**10**	1.6895	**49**	**33**	1.1979	**62**	**46**	1.0311
23	**7**	2.1803	**32**	**16**	1.6873	**50**	**34**	1.1836	**61**	**45**	1.0236
18	**2**	2.1033	**20**	**4**	1.6758	**59**	**43**	1.1691	**52**	**36**	1.0236
27	**11**	1.8818	**28**	**12**	1.6650	**57**	**41**	1.1559	**56**	**40**	1.0057
31	**15**	1.8445	**24**	**8**	1.6365	**55**	**39**	1.1530	**58**	**42**	1.0054
30	**14**	1.8148	**25**	**9**	1.4951	**51**	**35**	1.1524	**60**	**44**	1.0031
17	**1**	1.6972	**21**	**5**	1.3457	**53**	**37**	1.0803	**64**	**48**	1.0019

feed-forward network. The larger number of units at the last two nodes indicates that many of the units present fail to transfer to the entrance nodes, even though the averages m_1 and m_2 are larger in the feed-forward network.

For $c = 0$ compared with $c = 1$, the 13 scenario pairs with the lowest impact correspond to feed-forward networks, again justifying the hypothesis of the feed-forward network being quite insensitive to the desistance probability, i.e., units are seldom blocked. The pair 55/23 is the pair of scenarios which is the most affected by the value of c among all feed-forward networks. Here the entrance nodes have large buffers b_1 and high service rates, whereas the middle buffer is small and the service rate is low. When changing the networks in scenario pair 55/22 to rerouting networks (scenario pair 39/7), one sees that S_{39}/S_7 is the greatest of all scenario pairs.

In scenario pair 50/18, the total number of units decreases when $c = 1$ is changed to $c = 0$, which seems to contradict Theorem 2.5.4. This effect must be attributed to randomness in the samples. The simulation of the scenarios is not a simulation of the coupling used in the proof of Theorem 2.5.4. The processes that generate the samples continue for different length of time, which causes the streams of random numbers to be unsynchronised, hence, only stochastic domination is assured.

To summarise the results of the impact measures for the network, the buffer sizes at the entrance nodes are the most dominant of all parameters under study. Some scenarios reduced the total number of units in the network when the buffer size at the middle node was increased, while most scenarios increased the number of units when the service rate μ_1 was increased, in contrast to Jackson networks. An increase in μ_3 also yielded a larger value of $S(m)$ in some scenarios. The feed-forward networks had more

Table 4.9. The impact measures S_i/S_j ford scenarios differing by the value of c.

i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j	i	j	S_i/S_j
39	**7**	4.4351	**48**	**16**	1.7765	**33**	**1**	1.4684	**54**	**22**	1.0539
35	**3**	2.9765	**34**	**2**	1.7695	**45**	**13**	1.4226	**52**	**20**	1.0518
47	**15**	2.5633	**36**	**4**	1.7220	**41**	**9**	1.2980	**49**	**17**	1.0364
38	**6**	2.3513	**43**	**11**	1.7121	**56**	**24**	1.1555	**60**	**28**	1.0193
55	**23**	2.3454	**44**	**12**	1.6919	**51**	**19**	1.1543	**62**	**30**	1.0150
40	**8**	1.8803	**42**	**10**	1.6817	**61**	**29**	1.1014	**57**	**25**	1.0036
37	**5**	1.8480	**63**	**31**	1.4987	**59**	**27**	1.0636	**58**	**26**	1.0007
46	**14**	1.7865	**53**	**21**	1.4836	**64**	**32**	1.0549	**50**	**18**	0.9958

units than the rerouting ones; they were more insensitive to the values of the desistance probabilities.

In subsequent subsections some of the scenarios are examined in a more detail. The last subsection The end of this section covers a scenario containing an environment.

4.1.1 THE SCENARIO PAIR 21/29

The difference between scenarios 21 and 29 is the service rate at the middle node μ_3. Letting μ_3 range from 3.0 to 5.0 in steps of 0.2, the eleven settings of the service rate are simulated, each generating 10 000 samples on the stationary distribution, π. The setting $\mu_3 = 5.0$ is scenario 29, and when $\mu_3 = 3.2$ it is close to the one in scenario 21. The buffer sizes at the entrance nodes are large, as is the buffer at the centre node. It is a feed-forward network, $p = 0$, and the desistance probability is $c = 1$. The empirical distributions $\widehat{\pi}_1^{(\mu_3)}$, $\widehat{\pi}_3^{(\mu_3)}$ and $\pi_5^{(\mu_3)}$ are shown in Figures 4.10, 4.11 and 4.12, respectively.

Due to symmetry and the desistance probabilities $c_{13}(x) = c_{23}(x) = c = 1$, the marginal distributions of the number of units at the entrance nodes, $\pi_1^{(\mu_3)}$ and $\pi_2^{(\mu_3)}$, are equal and do not depend on μ_3, that is, $\pi_1 = \pi_2$. The marginal distribution of the number of units at the middle node, $\pi_3^{(\mu_3)}$, has an inflection for μ_3 close to 4.0, i.e., a parameter setting for which the number of units at the middle node is uniform over $\{0, \dots, 10\}$. The number of units at the last two nodes when μ_3 is small is stochastically dominated by the corresponding number when μ_3 is large, and by symmetry, $\pi_4^{(\mu_3)} = \pi_5^{(\mu_3)}$. If the entrance nodes had unlimited capacity, $\lim_{\mu_3 \to \infty} \pi_5^{(\mu_3)}$ would be, due

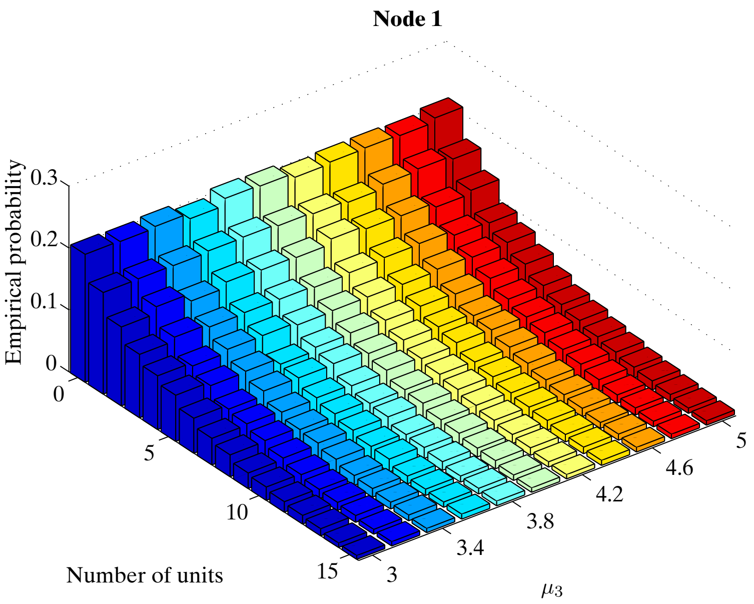

Figure 4.10. *Estimates, $\widehat{\pi}_1^{(\mu_3)}$, of the stationary distribution, $\pi_1^{(\mu_3)}$, of the number of units at node 1 as a function of the service rate $\mu_3(x)$. Since $c = 1$, the distribution does not depend on $\mu_3(x)$; for symmetry reasons, $\pi_1 = \pi_2$.*

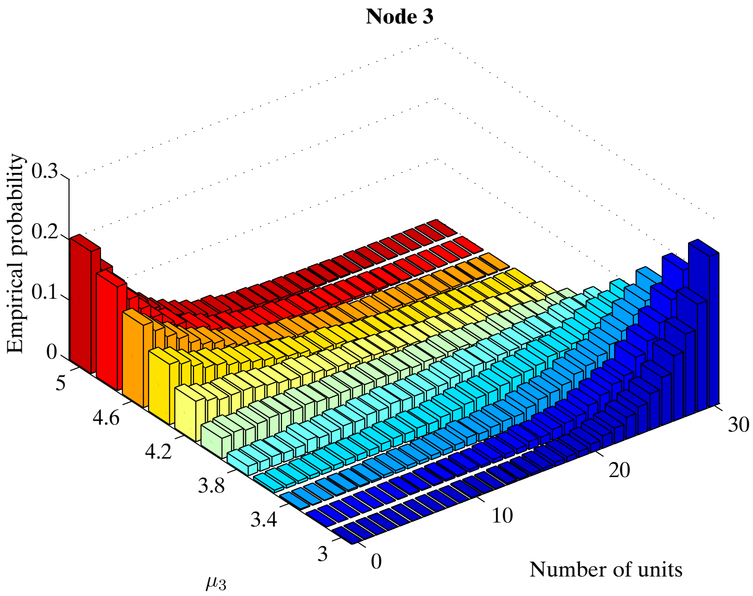

Figure 4.11. *Estimates of the stationary distribution, $\pi_3^{(\mu_3)}$, of the number of units at the centre node 3 as a function of the service rate $\mu_3(x)$.*

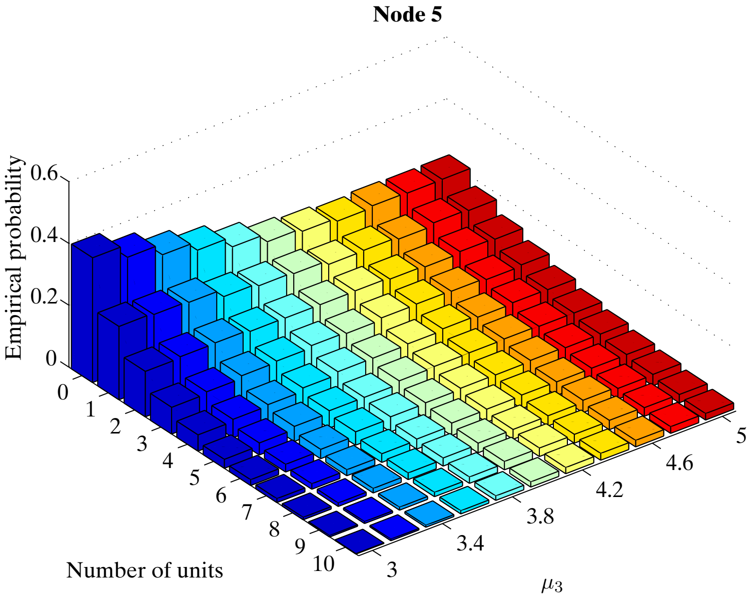

Figure 4.12. *Estimates of the stationary distribution, $\pi_5^{(\mu_3)}$, of the number of units as a function of the service rate $\mu_3(x)$. For symmetry reasons, $\pi_5 = \pi_4$.*

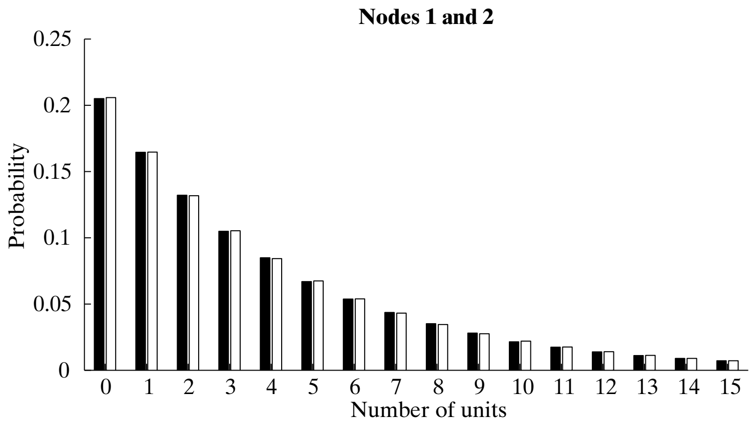

Figure 4.13. *Estimates $\widehat{\pi}_1$, of π_1 based on 220 000 independent samples (black bars), together with the theoretical distribution; a truncated geometric distribution with parameter $\rho = 0.80$ (white bars).*

to the network structure, a truncated geometric distribution on $\{0, \ldots, b_5\}$ with parameter $\rho = 0.80$.

Let us consider the two entrance nodes. Due to the desistance probabilities and the feed-forward structure, the two nodes behave as two independent M/M/1/15-nodes in isolation. Using Theorem 2.3.1, the marginal distribution π_1 is given by

$$\pi_1(x) = \frac{(1-\rho_1)\rho_1^x}{1-\rho_1^{16}}, \quad x = 0, 1, \ldots, 15, \tag{4.1}$$

a truncated geometric distribution with the parameter $\rho_1 = 0.80$. Using the symmetry argument, the samples $X_{1,1}^{(\mu_3)}, \ldots, X_{n,1}^{(\mu_3)}$ and $X_{1,2}^{(\mu_3)}, \ldots, X_{n,2}^{(\mu_3)}$ for all values of μ_3 are combined to estimate the common distribution π_1. This yields 220 000 independent samples on π_1, and the new empirical distribution is shown in Figure 4.13 together with π_1 from (4.1). The setting is rather artificial but suffices as an example of a situation for which the stationary distribution is readily calculated. The curious reader might be interested to know that the observed p-value of a χ^2 goodness-of-fit measure was 0.43.

We do not have an explicit expression for π_3. If the entrance nodes had unlimited capacity, the feed-forward structure would make node 3 behave as an M/M/1/30 node in isolation with Poisson arrivals at the rate $\Lambda_1 + \Lambda_2 = 2\beta_1 = 4$, that is, π_3 would be a truncated geometric distribution with parameter $\rho_3 = 4/\mu_3$. According to (4.1), the fraction of time the entrance buffers are full is $\pi_1(b_1) = (1-\rho_1)\rho_1^{15}/(1-\rho_1^{16}) = 0.0072$; due to the PASTA property, $\pi_1(b_1)$ is also the fraction of blocked units, which means that more than 99% of the arriving units are allowed into the network. The throughput from each of the entrance nodes is, using (2.13) or (2.14),

$$\Lambda_1(1-\pi_1(b_1)) = (1-\pi_1(0))\mu_1 = 1 - \frac{1-\rho_1}{1-\rho_1^{16}} \cdot 2.5 = 1.9855.$$

That is, the rate of arrivals and the traffic intensity for the centre node are

$$\Lambda_3 = 2\Lambda_1(1-\pi_1(b_1)) = 3.9710 \quad \text{and} \quad \rho_3 = \frac{\Lambda_3}{\mu_3},$$

respectively. The empirical distribution of the number of units at node 3, obtained by simulation, is compared with a truncated geometric distribution using parameter ρ_3 for some values of μ_3, see Figure 4.14

From this figure, it is clear that the qualitative behaviour of the number of units at node 3 is captured by the approximation of a Poisson arrival process. Before any p-values of a χ^2 goodness-of-fit measure are calculated, the parameter ρ_3 is estimated from the samples by the maximum likelihood method, under the assumption of a truncated geometric distribution. These estimates, denoted $\widehat{\rho}_3$, assume a Poisson arrival process to the centre node, but since $\widehat{\rho}_3$ is estimated from data, it corresponds to the best traffic intensity, in a maximum likelihood sense, for a node with a Poisson arrival process. The results of χ^2 goodness-of-fit tests for the empirical distribution and a truncated geometric distribution, with parameters ρ_3 and $\widehat{\rho}_3$, respectively, are summarised in Table 4.15. The estimates $\widehat{\rho}_3$ are close to ρ_3, and the p-values indicate a good correspondence between the empirical distributions and the truncated geometric distributions. Hence, the small blocking probability at the entrance nodes makes a Poisson approximation work for the values of μ_3 studied.

By symmetry, $\pi_4^{(\mu_3)} = \pi_5^{(\mu_3)}$, and as for the entrance nodes, the samples $X_{1,4}^{(\mu_3)}, \dots, X_{n,4}^{(\mu_3)}$ and $X_{1,5}^{(\mu_3)}, \dots, X_{n,5}^{(\mu_3)}$, are combined to estimate the common distribution $\pi_4^{(\mu_3)}$. Encouraged by the satisfactory results of approximating the distribution of units at node 3 with a truncated geometric distribution, the arrival intensity to nodes 4 and 5 is calculated for different values of μ_3:

$$\Lambda_4 = \Lambda_5 = (1-\pi_3(0))\mu_3 p_{34} = \left(1 - \frac{1-\rho_3}{1-\rho_3^{b_3+1}}\right)\mu_3 \frac{1}{2},$$

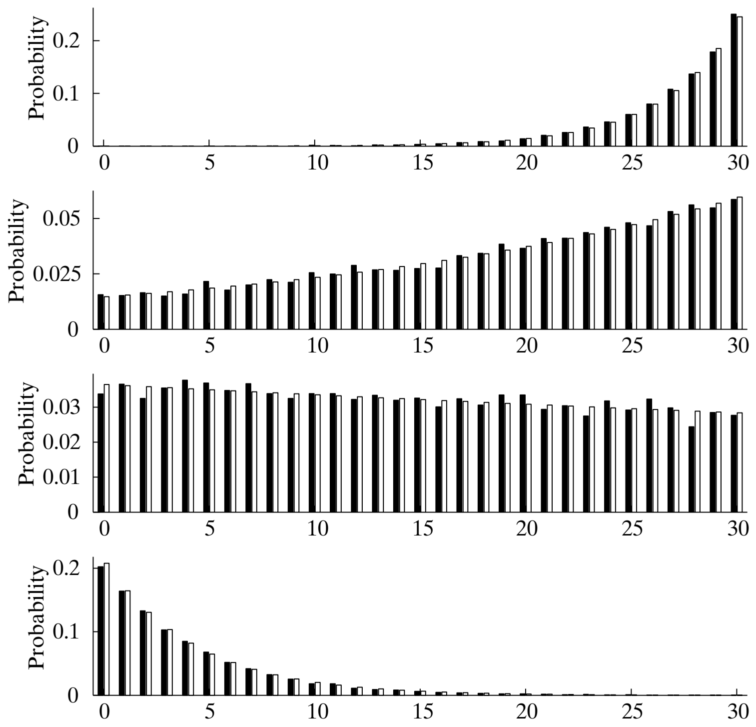

Figure 4.14. *Estimates* $\widehat{\pi}_3(x)$ *based on 10 000 independent samples for values of* μ_3 *(black bars). From top these correspond to* μ_3 *equal to* 3, 3.8, 4.0 *and* 5.0*. The white bars show the distributions of M/M/1/30 nodes with Poisson arrivals of intensity* Λ_3*: a truncated geometric distribution with parameter* ρ_3*. Near* $\mu_3 = 4.0$ *there is an inflection in the truncated geometric distribution, where the number of units is uniformly distributed on* $0, \ldots, 30$.

Table 4.15. The p-values of a goodness-of-fit test of equal distribution of $\widehat{\pi}_3$ and the distributions (4.1) with $\rho_3 = \Lambda_3^b/\mu_3$ and $\widehat{\rho}_3$ estimated from data.

μ_3	ρ_3	$\widehat{\rho}_3$	p-value(ρ_3)	p-value$(\widehat{\rho}_3)$
3	1.3237	1.3245	0.6122	0.5514
3.2	1.2409	1.2484	0.4211	0.6482
3.4	1.1680	1.1749	0.2246	0.6215
3.6	1.1031	1.1068	0.7468	0.9179
3.8	1.0450	1.0477	0.2187	0.3507
4	0.9928	0.9917	0.4038	0.3982
4.2	0.9455	0.9450	0.6768	0.6342
4.4	0.9025	0.9009	0.5567	0.5665
4.6	0.8633	0.8603	0.4284	0.5368
4.8	0.8273	0.8259	0.5611	0.5208
5	0.7942	0.7925	0.8638	0.8458

using ρ_3 and the approximation of π_3. The 20 000 samples give the empirical distribution $\widehat{\pi}_4$ which is compared with the truncated geometric distribution with parameter $\rho_4 = \Lambda_4/\mu_4$ in Figure 4.16. As before, the parameter ρ_4 is estimated from data using the maximum likelihood method, and the values of ρ_4 and $\widehat{\rho}_4$ are summarised in Table 4.17. The traffic intensities, ρ_4 and $\widehat{\rho}_4$, seem to be increasing towards 0.80, the traffic intensity in the corresponding open Jackson network. The difference between the distributions is small for the values of μ_3 under examination, and $\pi_4^{(\mu_3)}$ is so close to a truncated geometric distribution that an approximation can be used.

To conclude the approximations in this subsection, the distribution of the number of units at a node in scenario 22 can be approximated with a truncated geometric distribution. A network optimiser would be interested in selecting a service rate so that a desired throughput rate is obtained. The throughput, or stationary flowrate, of units through the network,

$$2\Lambda_4(1 - \pi_4(b_4)) = (1 - \pi_4(0))\mu_4 + (1 - \pi_5(0))\mu_5,$$

may be approximated with the above procedure, or estimated from data: either by using $\widehat{\pi}_4(0)$ and $\widehat{\pi}_5(0)$ directly, or by using $\widehat{\pi}_4(b_4)$ and $\widehat{\pi}_5(b_5)$ and estimating the throughput from the centre node, $\Lambda_4 + \Lambda_5$. A comparison of the approximated and the estimated throughputs, based on empirical distributions $\widehat{\pi}_4$ and $\widehat{\pi}_5$, is given in Figure 4.18.

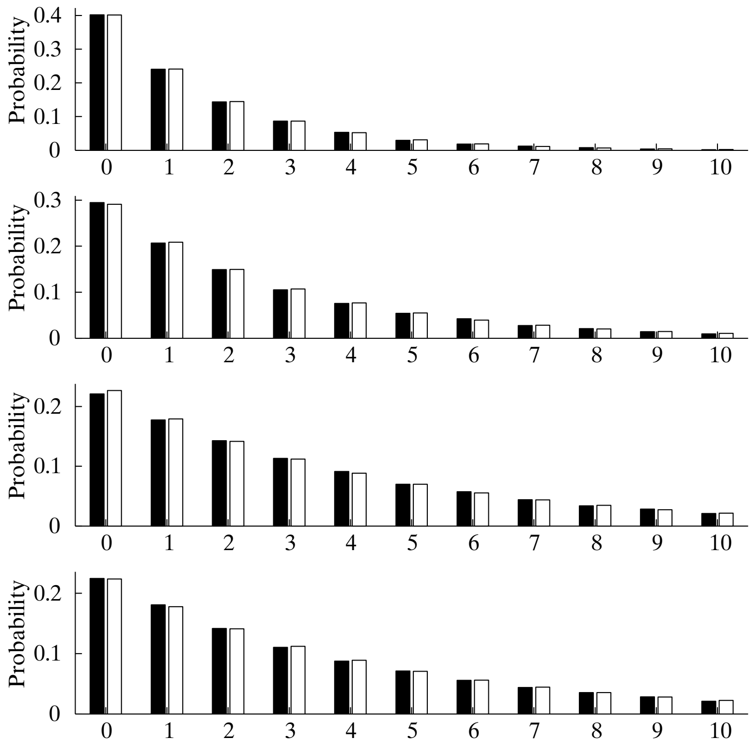

Figure 4.16. *Estimates* $\widehat{\pi}_4(x)$ *based on 20 000 independent samples for different values of* μ_3 *(black bars). From above they correspond to* μ_3 *equal to* 2.0, 3.0, 4.0 *and* 5.0. *The white bars represent a truncated geometric distribution with parameter* $\rho_4 = \Lambda_4/\mu_4$.

Table 4.17. The estimated parameters $\widehat{\rho}_4$ and calculated ρ_4 for a truncated geometric distribution belonging to a M/M/1/10 node.

μ_3	ρ_4	$\widehat{\rho}_4$	μ_3	ρ_4	$\widehat{\rho}_4$
3.0	0.6000	0.6005	4.2	0.7844	0.7862
3.2	0.6398	0.6403	4.4	0.7905	0.7934
3.4	0.6791	0.6770	4.6	0.7929	0.7960
3.6	0.7163	0.7156	4.8	0.7937	0.7916
3.8	0.7483	0.7466	5.0	0.7940	0.7917
4.0	0.7713	0.7713			

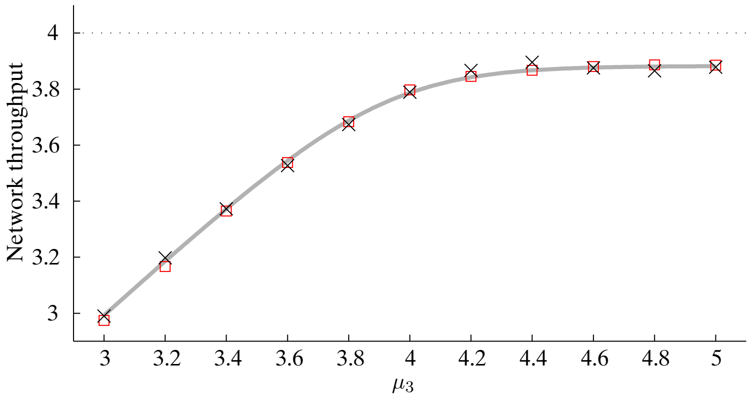

Figure 4.18. *Estimates of the network throughput, based on* $\widehat{\pi}_4(0)$ *and* $\widehat{\pi}_5(0)$*, shown as crosses. The grey line is the approximate throughput, calculated under the assumption of Poisson arrivals to every node in the network. The dotted line is the intensity of the arrival processes to the entrance nodes,* $\Lambda_1 + \Lambda_2 = 4$*. The squares are the throughput estimated from* $\widehat{\pi}_4(b_4)$*,* $\widehat{\pi}_5(b_5)$ *and estimated throughput from the centre node.*

Before considering a scenario where the approximation utterly fails, other network quantities are studied. A graph of the average number of units at nodes 1, 3 and 4 is shown in Figure 4.19. Also shown is the theoretical average for nodes 1 and 2, and the average obtained from a truncated geometric distribution with parameters ρ_3, $b = 30$ and ρ_4, $b = 10$, respectively. The difference between the real distribution π_3 and the approximation does not produce a significant difference in means.

In Figure 4.20 the expected total number of units in the network, $\mathsf{E}[S(X)]$, $X \overset{\mathscr{D}}{=} \pi$ as a function of μ_3, is estimated.

4.1.2 THE SCENARIO PAIR 22/30

The analysis of this scenario pair is the same as in the preceding section. The difference between the pairs 22/30 and 21/29 is the buffer size at the entrance nodes b_1. Here $b_1 = 3$ and, as a consequence, this highly limited buffer capacity gives a stuttering Poisson process as a departure process from the entrance nodes. Using a Poisson process to approximate the arrival process to node 3, which was accurate in scenario 21, is now definitely

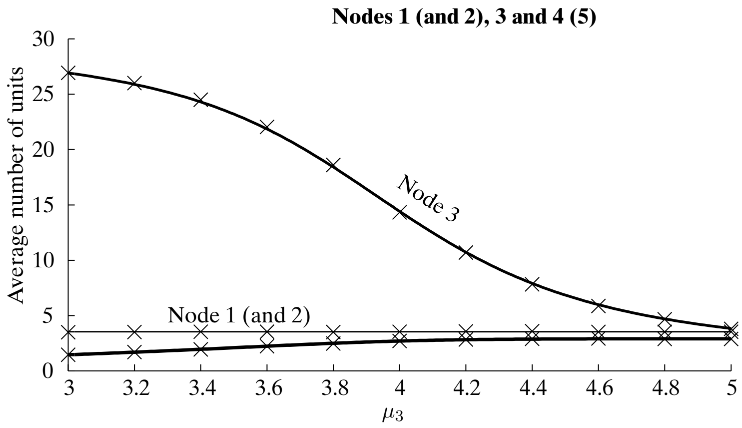

Figure 4.19. *Estimates of the average number of units, at the nodes of the network, shown as crosses. The estimates for nodes 1 (and 2) are based on 220 000 samples on π_1, as in Figure 4.13, and do not depend on μ_3. The estimates for node 3 are based on 10 000 samples, and for node 4, on 20 000 samples. Lines correspond to the theoretical expectations (for nodes 1 and 2) and the expectations of a truncated geometric distribution as a function of μ_3 (nodes 3–5).*

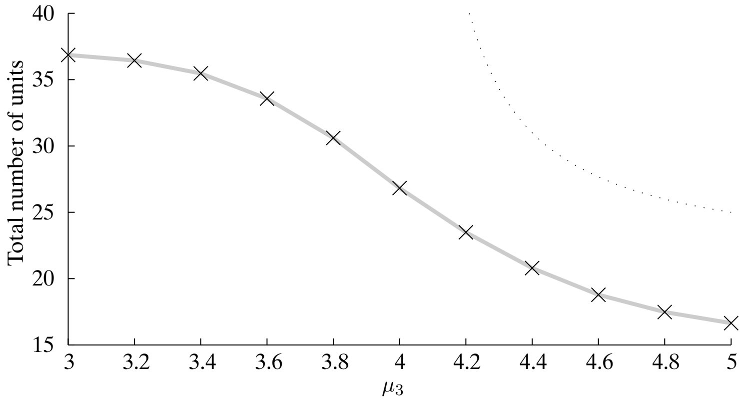

Figure 4.20. *The estimates of* $\mathsf{E}[S(X)]$, $X \stackrel{\mathscr{D}}{=} \pi$, *shown as crosses together with a linear interpolation. The dotted line shows the expected value of $S(X)$ for the Jackson network with the same set of parameters. Note that it is defined only for $\mu_3 > 4$.*

inappropriate.

The traffic load at the entrance nodes is $\rho_1 = 0.80$ as before. The fraction of time the entrance buffers are full is $\pi_1(b_1) = (1-\rho)\rho^3/(1-\rho^4) = 0.1734$. Hence, the fraction of allowed units is less than 83% of the arriving units, and the throughput from one of the entrance nodes is

$$\Lambda_1(1-\pi_1(b_1)) = (1-\pi_1(0))\mu_1 = 1 - \frac{1-\rho}{1-\rho^4} \cdot 2.5 = 1.6531.$$

Thus, the arrival intensity and traffic load to the middle node in the network are

$$\Lambda_3 = 2\Lambda_1(1-\pi_1(b_1)) = 3.3062 \quad \text{and} \quad \rho_3 = \frac{\Lambda_3}{\mu_3},$$

respectively. Note that Λ_3 is the arrival rate, but the small buffer size b_1 makes the departure process thin in comparison to a Poisson process; it is thin in the sense that not too many units may be close to each other. A comparison of the empirical distribution $\widehat{\pi}_3$ to a truncated geometric distribution with parameter ρ_3, as if the arrival process were (or at least close to) Poisson, is given in Table 4.21. Also shown are the maximum likelihood estimates $\widehat{\rho}_3$ obtained from data and the p-values of a χ^2 goodness of fit test between the two truncated geometric distributions and the empirical distribution. The estimates $\widehat{\rho}$ are quite close to ρ, but the p-values are very small, indicating a lack of fit between the distributions.

The conclusion drawn is that the approximation of Poisson flows is not suitable for bounded networks, in particular when some of the buffer sizes are small. With parameters $\widehat{\rho}$, estimated from data, the approximation is better, but when one has many samples, the empirical distributions are, naturally, preferable. In Section 4.1.4 another example of a peculiar situation, for which the truncated geometric distribution is highly inappropriate, is shown.

4.1.3 THE SCENARIO PAIRS 17/21 AND 49/53

The scenario pairs 17/21 and 49/53 have large buffers, b_1 and b_3, but slow service rates at the entrance and middle nodes. Let $\mu_3 = 4.0/\rho$, where $\rho = 1.2$, and vary the service rate at the entrance node μ_1 between 0.2 and 3.0. In both scenario pairs, there are feed-forward networks, and in 17/21 the desistance probability is $c = 1$, while scenario pair 49/53 has persistent units. Table 4.6 revealed that an increase of the service rate $\mu_1 = 2.0/\rho$,

Table 4.21. The p-values of a goodness-of-fit test of equal distribution between $\widehat{\pi}_3$ and the distribution given by (4.1) for $\rho_3 = \Lambda_3^b/\mu_3$, as well as $\widehat{\rho}_3$ estimated from data. Omitted are small values which were beyond the computer precision.

μ_3	ρ_3	$\widehat{\rho}_3$	p-value(ρ_3)	p-value$(\widehat{\rho}_3)$
3	1.1021	1.1249		0.1355
3.2	1.0332	1.0421	$1.5630{\cdot}10^{-10}$	0.0017
3.4	0.9724	0.9694	0.1343	0.3508
3.6	0.9184	0.9047		$4.9154{\cdot}10^{-4}$
3.8	0.8701	0.8503		$3.3875{\cdot}10^{-10}$
4	0.8266	0.8026		$7.0303{\cdot}10^{-11}$
4.2	0.7872	0.7599		$2.7094{\cdot}10^{-10}$
4.4	0.7514	0.7267		$5.5299{\cdot}10^{-10}$
4.6	0.7188	0.6916		$6.7052{\cdot}10^{-4}$
4.8	0.6888	0.6626		$1.6088{\cdot}10^{-6}$
5	0.6613	0.6389	$2.0531{\cdot}10^{-12}$	$1.3109{\cdot}10^{-5}$

from $\rho = 1.2$ to $\rho = 0.8$ in scenario 49/53, yielded a decrease in the total number of units in the network. Introducing a greater span of service rates, μ_1, shows that the total number of units in the network increasing even for small service rates μ_1 when $c = 1$, i.e., scenario 17/21.

Heuristically, for small values of μ_1, the entrance nodes are often saturated and many units are blocked. For sufficiently small values of μ_1, the arrival process to node 3 is close to a Poisson process with intensity $2\mu_1$. As μ_1 increases, the entrance nodes become less overloaded and more units are allowed into the network. However, the centre node eventually becomes overloaded; the arrival intensities $\beta_1 + \beta_2 = 4 > \mu_3$ and units are expunged when blocked at node 3. Further increase of μ_1 reduces the number of units at the entrance nodes, and increases the number of dropped units, which together lowers the total number of units in the network. This phenomenon is shown in Figure 4.22. When the desistance probabilities are $c = 0$ (scenario pair 49/53) units that fail to transfer from the entrance nodes remain there; Theorem 2.5.4 states that there is a stochastically larger number of units, at all nodes of the network, than seen when $c = 1$. In fact, here the number of units in the network rises for all values of μ_1 under examination, as can be seen in Figure 4.22.

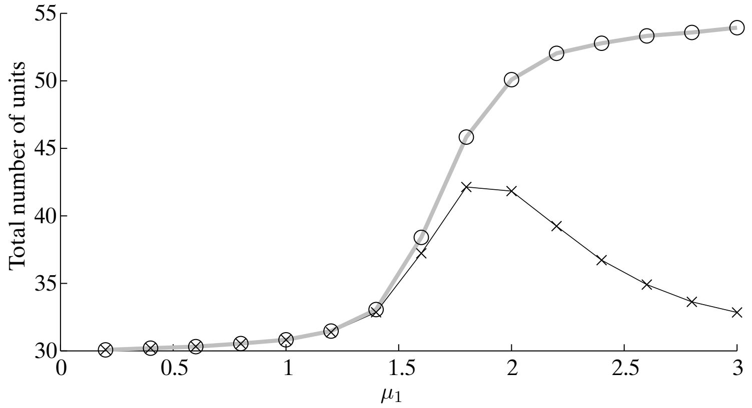

Figure 4.22. *Estimates of the total number of units in the network as a function of* μ_1, *the service intensity at the entrance nodes. The crosses are the means of 10 000 stationary samples each for the scenario with* $c = 1$, *and the line is a linear interpolation of the observations. The circles correspond to a simulation of the same network with desistance probabilities* $c = 0$.

4.1.4 SCENARIO 5

Among the 64 scenarios examined, scenario 5 has a peculiar distribution of the number of units at node 3, π_3. The empirical distribution is shown in Figure 4.23. The distribution of the number of units appears to be concave and, since every truncated geometric distribution is convex, this distribution at some parameters settings is highly inappropriate for approximating the number of units at a node.

4.1.5 SCENARIO PAIR 7/39

The scenario pair 7/39 differs in the value of the desistance probability c. Otherwise, the networks are the same rerouting ones with large buffers at the entrance nodes that have high service rates. The middle buffers are small and overloaded, that is, μ_3 is small. For the different values of the desistance probability c, as it changes from 0 to 1, 100 000 independent perfect samples on the network were taken. A large sample size is used, since conditional distributions of the number of units at different nodes of the network are estimated, and some of these are conditional distributions of

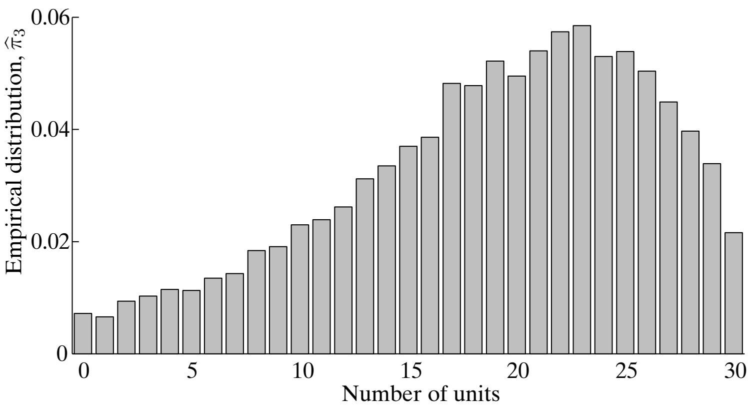

Figure 4.23. *The empirical distribution $\widehat{\pi}_3$ in scenario 5. Of the 64 scenarios generated, this is the only one having a bowl shape with low probabilities for small and large numbers of units at node 3.*

rare events. In open Jackson networks, where the stationary distribution satisfies (2.5), the number of units at different nodes at a fixed time t are independent random variables. Thus, the marginal distribution π_3, for example, can be estimated from stationary samples conditioned on X_1, and similarly, π_1 estimated from samples conditioned on X_3.

For each value of c, the same 100 000 samples are used to estimate the marginals $\pi_1, \ldots, \pi_5$, as well as the conditional distributions. Hence, the estimates of π_1 and $\mathsf{P}(X_1 = \cdot | X_3 = 0)$, say, are dependent in such way that the difference between them is as small as possible. Moreover, when comparing the two distributions, the discrepancies are smaller than if the samples had been independent, and the χ^2 goodness-of-fit measures underestimate the true differences. In Figure 4.24, the empirical marginal distributions $\widehat{\pi}_1$ are compared with the empirical distributions estimated from samples for which the middle node is empty, i.e., samples where $x_3 = 0$.

The middle node is overloaded due to the low service rate. Samples with $x_3 = 0$ are rare, especially for small values of c, but as c increases, the number of samples increases and gives better estimates of the conditional distributions $\mathsf{P}(X_1 = \cdot | X_3 = 0)$. When $c = 0$, units stay at the entrance nodes until they are allowed at the middle node when they may fit into the buffer. Of the 100 000 samples, only 5 454 samples with $x_3 = 0$ were obtained, which shows that the middle node is empty only for a small proportion of

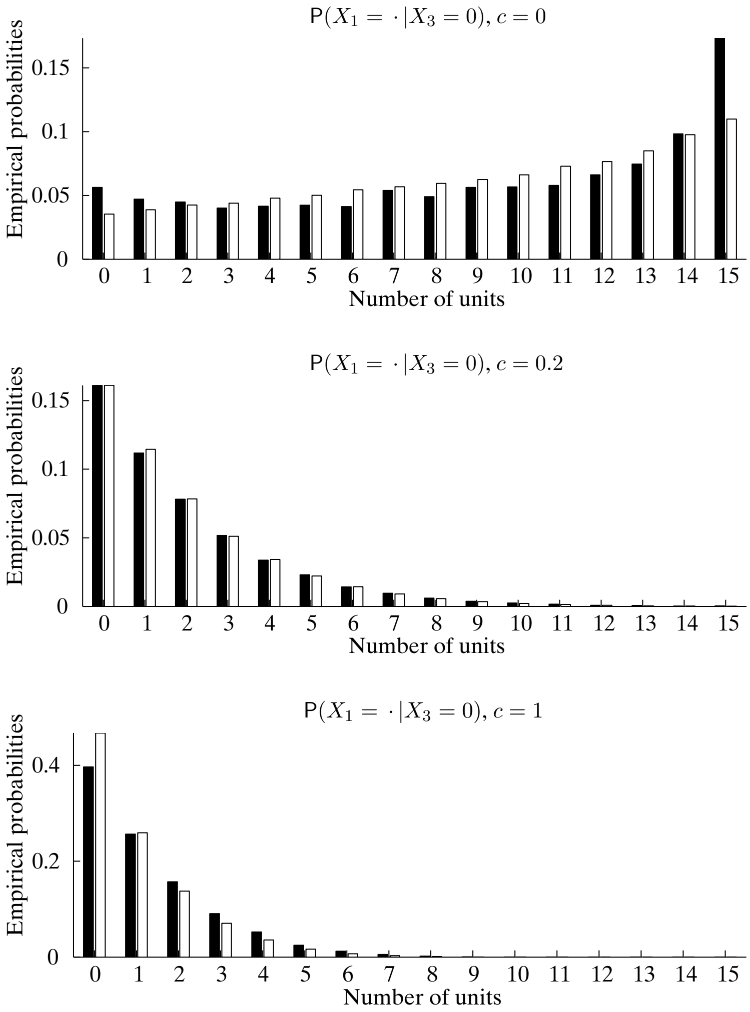

Figure 4.24. *Estimates of the distribution of the number of units at node 1 in the networks with desistance probability $c = 0$ (top), $c = 0.2$ (middle) and $c = 1$ (bottom). Black bars are the estimated conditional probabilities obtained from samples where $x_3 = 0$, i.e., when the middle node is empty. For $c = 0$, 5 454 samples were obtained, for $c = 0.2$, 16 432 samples, and for $c = 1.0$, 24 262 samples. Shown as white bars is the marginal distribution π_1 estimated from all 100 000 samples.*

the time. Estimating the conditional distribution of the number of units at node 1 based on these samples, one sees that the fraction of time node 1 is empty, $\pi_1(0)$, underestimates the conditional probability of an empty buffer at node 1. The same is true for the marginal distribution, $\pi_1(b_1)$, and $\mathsf{P}(X_1 = b_1 | X_3 = 0)$. Thus, conditioned on $X_3 = 0$, the probability of the entrance nodes being empty or full, is greater than the corresponding fraction of the total time.

When the desistance probability is changed to $c = 0.20$, the estimates indicate a close correspondence between the marginal distribution π_1 and the conditional distribution $\mathsf{P}(X_1 = \cdot | X_3 = 0)$. A goodness-of-fit test cannot be rejected on any reasonable significance level; the number of samples was 16 432. For $c = 1$, $\pi_1(0)$ overestimates the conditional probability, $\mathsf{P}(X_1 = 0 | X_3 = 0)$; in other words, if the middle node is empty, the probability of units at node 1 is higher than the fraction of time there are units there. The conditional distribution was estimated from 24 262 samples.

The conditional distributions when conditioning on $X_3 = b_3$ are shown in Figure 4.25. As the middle node is overloaded, samples with $X_3 = b_3$ are frequent, especially for small values of c. When units are persistent, $c = 0$, 35 266 samples with $x_3 = b_3$ were obtained from the total of 100 000, showing that the buffer at the middle node is full more than a third of the time. Here, $\pi_1(0)$ overestimates the conditional probability of an empty buffer at node 1, consequently, when the middle node is full, there is a higher probability of units being present at node 1. When $c = 0.10$, the estimates reveal a curiously shaped conditional distribution of the number of units at node 1. The probability of having a single unit at the entrance node is greater than that of having no units at all, however it is also greater than the probability of the buffer being full. The number of samples used for estimating the conditional distribution was 23 763. For $c = 1$, the proportion of time node 1 is empty, $\pi_1(0)$, underestimates the corresponding conditional time; units are less likely to be at the entrance nodes when the centre node is saturated. The conditional distribution was estimated from 7 730 samples.

Estimates of the conditional distributions $\mathsf{P}(X_3 = \cdot | X_1 = 0)$ for $c = 0$ and $c = 1$ are shown in Figure 4.26. When $c = 0$, the number of units at node 1 is large, and from 100 000 samples, only 3 535 had $x_1 = 0$. When $c = 1$, X_1 is stochastically smaller and, as a consequence, 46,726 samples were obtained. Since μ_3 is small, the centre node is overloaded, but visually, the distribution in Figure 4.26, when $c = 1$, could be mistaken for an underloaded setting. When $c = 0$, the marginal distribution seems to over-

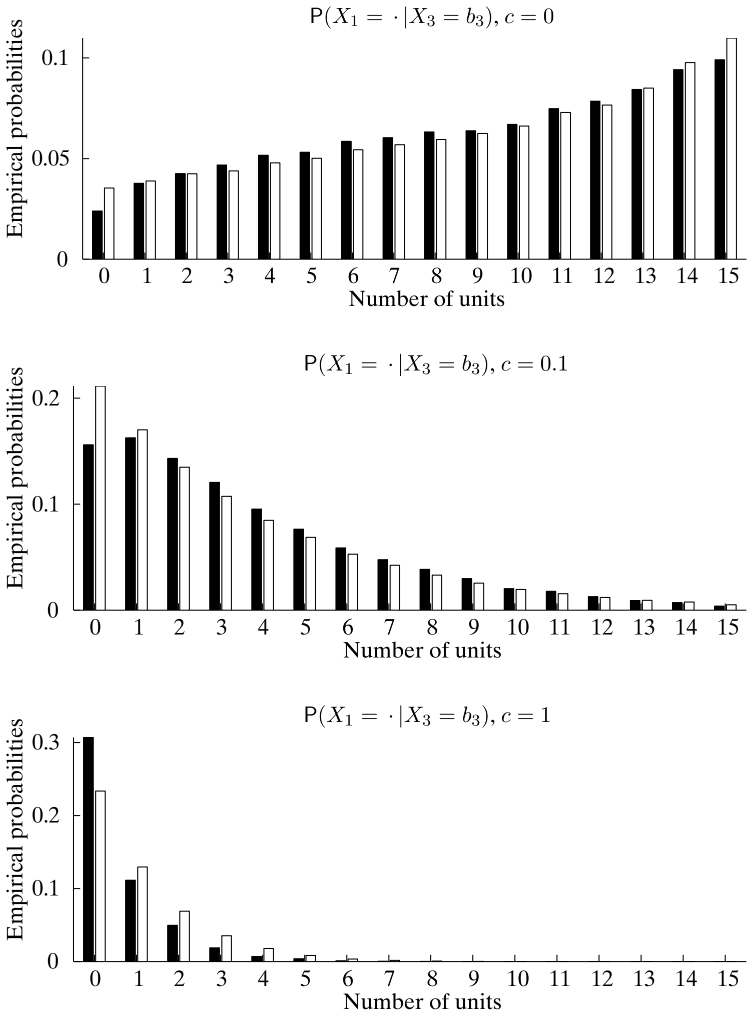

Figure 4.25. *Estimates of the distribution of the number of units at node 1 in the networks with desistance probabilities:* $c = 0$ *(top),* $c = 0.1$ *(middle) and* $c = 1$ *(bottom). Black bars are the estimated conditional probabilities obtained from samples where* $x_3 = b_3$*, i.e., when the middle node is full. Shown as white bars is the marginal distribution obtained from all 100 000 samples.*

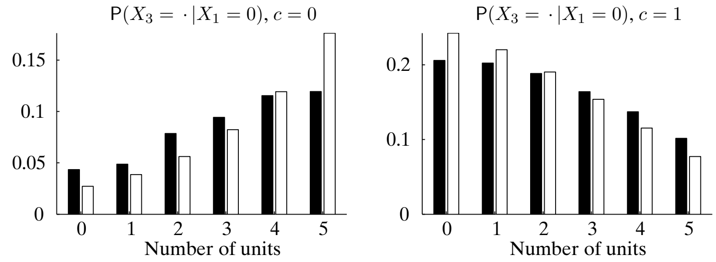

Figure 4.26. *Estimates of the distribution of the number of units at node 3 in the network with desistance probability $c = 0$ and $c = 1$. Black bars are the estimated conditional probabilities obtained from samples where $x_1 = 0$, i.e., when one of the entrance nodes is empty. Shown as white bars is the marginal distribution obtained from all 100 000 samples.*

estimate the probability of node 3 being full, when conditioned on $X_1 = 0$, and overestimate it being empty, when $c = 1$.

The distributions of the number of units at the last two nodes, where $X_1 = 0$ and $X_1 = b_1$ for $c = 0$, are shown in Figure 4.27. The number of samples where $x_1 = 0$ was 3 535, and for $x_1 = b_1$ it was 10 988. Due to the structure of the network, the distributions of the number of units at nodes 1 and 4 are highly dependent. The number of units at node 5 also depends on the number of units at node 1; the difference between nodes 4 and 5 is great considering their conditional distributions for different values of X_1.

The number of units at node 4, given that node 1 is full, is stochastically larger than given that node 1 is empty. This is intuitively clear, since units which are rerouted back to entrance node 1, may not leave when blocked. When $X_1 = 0$, the conditional distribution and the marginal distribution coincide approximately. For node 5, given that node 1 is empty, gives a stochastically larger number of units than conditioning on node 1 being full. A comparison with the marginal distribution, when $X_1 = 0$, reveals that the probability of node 5 not being empty, is greater than the estimate of the marginal distribution suggests.

The conditional distributions of the number of units at the other entrance node, node 2, when $X_1 = 0$ and $X_1 = b_1$ for $c = 0$ are shown in Figure 4.28. Again, the number of samples for which $x_1 = 0$ was 3 535, and the corresponding number for $x_1 = b_1$ was 10 988. For node 2, the condition that

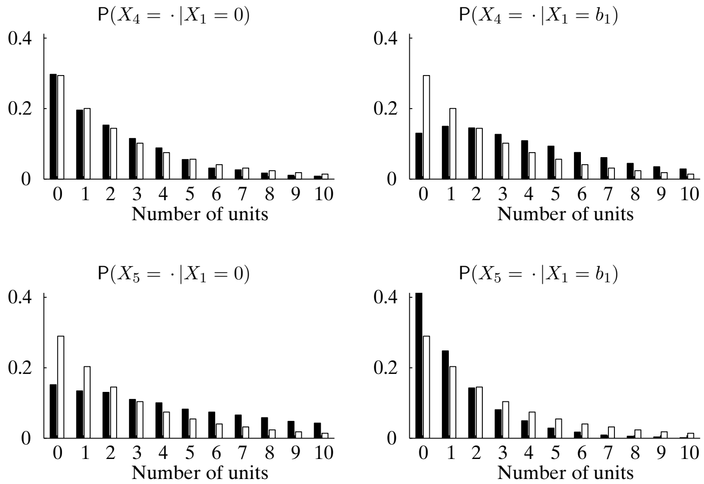

Figure 4.27. *Estimates of the distribution of the number of units at node 4 (top) and node 5 (bottom) conditioned on $X_1 = 0$ (left) and $X_1 = b_1$ (right) in the network with desistance probability $c = 0$. Black bars are the estimated conditional probabilities, and white bars are the marginal distribution obtained from all 100 000 samples. The marginal distributions, for symmetry reasons, are the same in all graphs.*

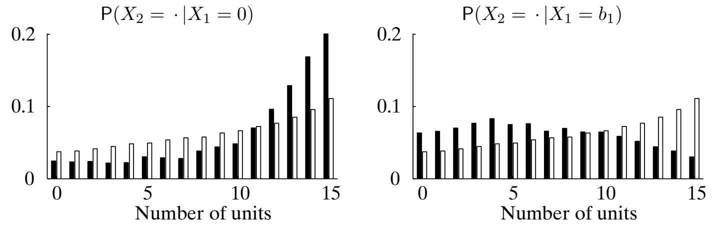

Figure 4.28. *Estimates of the distribution of the number of units at node 2 in the network with desistance probability $c = 0$. Black bars are the estimated conditional probabilities obtained from samples where $x_1 = 0$ (left) and $x_1 = b_1$ (right). Shown as white bars is the marginal distribution obtained from all 100 000 samples.*

node 1 is empty, gives a stochastically larger number of units at node 2 than the condition that node 1 is full. The probability of node 2 not being empty, given that $X_1 = 0$, is greater than the estimate of the marginal distribution suggests. The qualitative behaviour of node 2, as conditioned on node 1, is thus much the same as that of node 5.

4.1.6 SCENARIO 23 WITH COMPETING ENTRANCE NODES

Scenario 23 is a feed-forward network with desistance probability $c = 0$. The entrance nodes have large buffers and high service rates, while the middle node has a small buffer and a low service rate. Here, the entrance nodes have different service rates: $\mu_2(x) = 2.5 \cdot \mathbf{1}_{\{x_2>0\}}$, that is, the usual high setting, and $\mu_1(x) = 2.5\nu \cdot \mathbf{1}_{\{x_1>0\}}$ for some constant $\nu \geqslant 1$.

Due to the limited buffer capacity at the overloaded middle node, units leaving node 1 compete for admittance at the middle node with units from node 2. For $\nu > 1$, the faster service rate at node 1 enables units at that node to make transfer attempts more often than units at node 2. Hence, for a given time period, units from node 1 are more likely to be allowed at the middle, while the number of units at node 2 increases, since the middle buffer is likely to be full when a unit finishes receiving service.

Taking 10 000 perfect samples from the network for selected values of $\nu \geqslant 1$, Figure 4.29 is obtained. As ν rises the service rate at node 1 increases and the traffic load decreases. The number of units present at node 1 is stochastically decreasing in ν. For node 2, where units compete for admit-

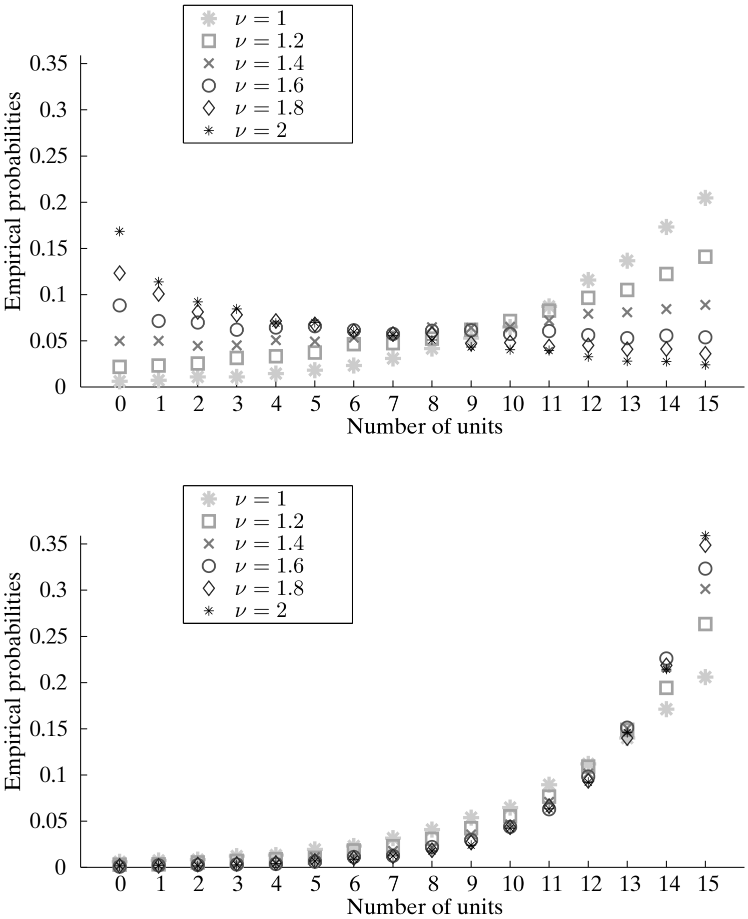

Figure 4.29. *Estimates of the distribution of the number of units at node 1 (left) and node 2 (right) for service rates* μ_1. *The service rate* μ_2 *is constant, but units at node 2 compete for admittance to node 3 with units from node 1; for large values of* μ_1, *the latter have an advantage. Shown is* $\mu_1 = \nu \cdot \mu_2$ *for some values of* $\nu \geqslant 1$.

tance at the middle node with units from node 1, the number of units is stochastically increasing in ν. This is the reason why only tandem networks are considered in Theorem 2.5.11.

4.1.7 SERVER FAILURE AT NODE 3

Consider scenario 5 which has large buffers in a rerouting network with desistance probabilities $c = 1$. The entrance nodes have high service rates μ_1, but the middle node has a low service intensity μ_3. The distribution of the number of units at the middle node was shown in Figure 4.23 on page 126.

Let $Z = (Z_t)_{t \geqslant 0}$ be an ON/OFF process with parameters ν and σ. Assume further that Z is an one-dimensional Poisson environment to the network in scenario 5, but

$$\mu_3(x) = z \cdot \mu_3 \cdot \mathbf{1}_{\{x_3 > 0\}},$$

i.e., the function $g_3(z) = z$ in Definition 2.3.1. The ON/OFF process Z determines the operating periods of the server at the middle node. When $Z_t = 0$ the server is turned 'off', and no units can leave the node until the server is again turned on and units may finish receiving service.

The distribution of the number of units at the middle node, as a function of the fraction of time the server at node 3 is operating, is investigated briefly. Since the expected 'on' length for the ON/OFF process is $1/\sigma$, and the expected cycle length is $1/\nu + 1/\sigma$,

$$\frac{1/\sigma}{1/\sigma + 1/\nu} = \frac{\nu}{\nu + \sigma}$$

is the average fraction of time the server is operating. The service rate, μ_3, is adjusted to

$$\mu_3 = \mu \frac{\nu + \sigma}{\nu} = \mu \left(1 + \frac{\sigma}{\nu} \right)$$

for some constant $\mu > 0$, that is, the service intensity, μ_3, times the proportion of time the server is working is constantly μ. The constant μ is chosen in such way that when $\nu \to \infty$, i.e., the server is constantly on, scenario 5, $\mu = \mu_3 = 10/3$, is obtained.

Let $\sigma = 5$, i.e., the expected length of an 'on' period is $1/5$ of a time unit, and vary ν between 1 and 9. This means the fraction of time the server is 'on' goes from $1/6$ to $9/14$. A graph of the distribution of the number of units at the middle node is shown in Figure 4.30. For each setting of ν, 100 000 samples are taken, and the distribution is a mixture of the distributions conditioned on the state of the environment. For $\nu = 1$, the estimated conditional distributions of the number of units are also shown in the figure.

In the previous setting, ν was varied to obtain different time proportions of server operation. When both ν and σ are varied so that $\nu/(\nu + \sigma)$ is constant, the fraction of time the server is operating does not change, but the distribution of the number of units at the middle node does.

To see this, let the rate of the server, when active, be huge compared with the arrival rate, so that the server, if working, quickly empties the buffer and expunges the units present. If ν and σ are both small, the environment does not often change its state and the server is allowed to work undisturbed for long time periods, which generates long periods with few units present. When the server is turned 'off', a queue builds up in the buffer. If, on the other hand, ν and σ both are large compared with the service rate, but $\nu/(\nu + \sigma)$ is the same as before, the fraction of time the server operates is preserved. As the environment rapidly changes its state, the distributions of the number of units present conditioned on the state of the environment are more or less the same.

As $\nu, \sigma \to \infty$, one would think that the number of units present increases, since the server may never finish serving one unit before it is switched off, and when again turned on, the remaining service time of the unit is still the same exponential distribution; the 'on' and 'off' switching of the server does not prevent units from entering the node. This argument, however, is not necessarily true. As ν and σ increase, the state of the environment at the times of successive departure attempts from node k, that is Poisson events that fall in the strip $I(\mu_k)$ in the embedding, becomes less correlated; the limit probability of an active server is $p = \nu/(\nu + \sigma)$, independently of previous attempt. Thus, the time until a departure from node k is a sum of exponentially distributed random variables with a geometrically distributed number of terms. Hence, the sum is exponentially distributed with intensity $\mu_k p = \mu$, the unboosted intensity of the server as if it were constantly on. Two conclusions can be drawn. First, in the limit as ν and σ increase but $p = \nu/(\nu + \sigma)$ is constant, the distributions of the number of units conditioned on the state of the environment coincide. Second, the common conditional distribution is that of a server with rate $\mu_k p$ and constantly operating. If $\mu_k p$ is much smaller than the arrival rate, the number of units at the node is, of course, larger.

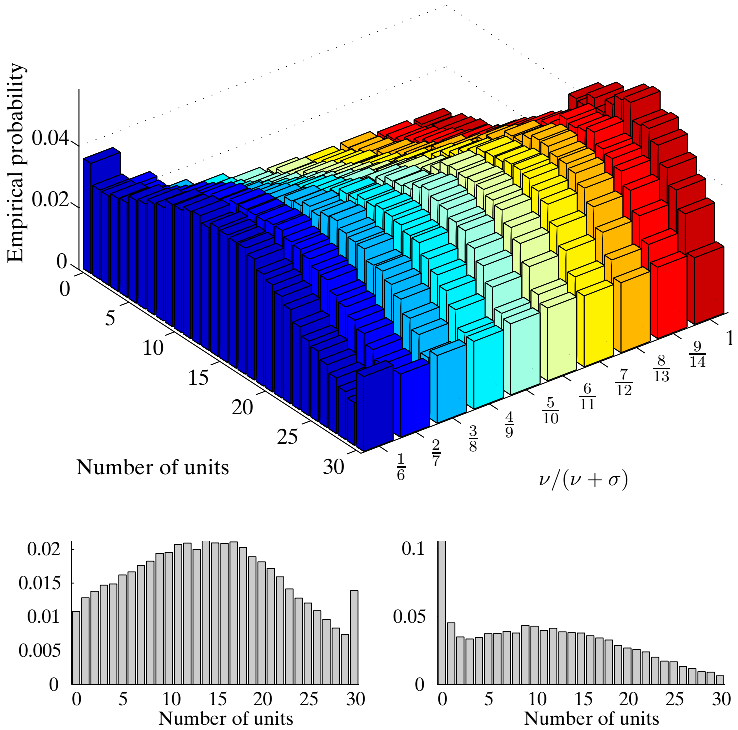

Figure 4.30.
Top: The estimated distribution of the number of units at the middle node for values of ν, where $1/\nu$ is the expected length of an operating period of the server. Also shown is the distribution from scenario 5, when the server is constantly on, which is estimated from 10 000 samples.
Bottom: The estimated distributions of the number of units at the middle node conditioned on the state of the environment for $\nu = 1$. To the left is the distribution given that the server is 'off', and to the right, that it is 'on'. The number of samples is 83 252 and 16 744, respectively.

A simulation of the network in scenario 5 with different on and off rates but constant fractions $\nu/(\nu+\sigma)$, can be found in Figure 4.31. If $\nu = \theta/5$ and $\sigma = \theta$ for some constant $\theta > 0$, then

$$\frac{\nu}{\nu+\sigma} = \frac{\theta/5}{\theta/5+\theta} = \frac{1}{6}$$

for all $\theta > 0$.

The factor θ is changed from small to large values to see how the number of units at the middle node is affected by the change in transition rate of the environment. Taking 10 000 stationary samples for each value of θ, the empirical distribution $\widehat{\pi}_3$ is shown in Figure 4.31. Note that in the limit $\nu \to \infty$, that is scenario 5, the centre node is overloaded and μ_3 is small. The service rate is boosted by a factor of 6, since it is operating only a fraction 1/6 of the time, i.e., when 'on', the server has enough capacity to serve the arrival process. For small values of θ, transitions occur more seldom in the environment, and the probability distribution π_3 becomes more like a distribution with support only on large and small values. As θ increases and transitions occur more frequently in the environment, the estimated distributions $\widehat{\pi}_3$, conditioned on the different states of the environment, seem to coincide. Also shown in Figure 4.31 are the conditional probabilities of an empty buffer and a full one at the middle node for a working server and an inoperational one, respectively. Here, as θ increases these conditional probabilities decrease.

In Figure 4.32, the estimated distributions of the number of units at node 3 given the state of the environment is shown for the case $\theta = 50$. When the server rapidly switches between its 'on' state and 'off' state the difference between the distributions decreases.

4.1.8 BATCH ARRIVALS TO THE ENTRANCE NODES

Consider scenario 7 for which the entrance nodes have large buffers and high service rates, while the middle node has a small buffer and low service rate. It is a network with rerouting, $p = 0.90$, and persistent units, $c = 1$.

With a constant arrival rate to the entrance nodes and a blocking mechanism that allows as many units as will fit into the buffer, Theorem 2.5.5 states that increasing the batch sizes from 1 gives stochastically more units at every node of the network. Let $\beta_1^{(k)} = \beta_1$ be the arrival rate of batches of size k. The number of units at every node is stochastically larger for

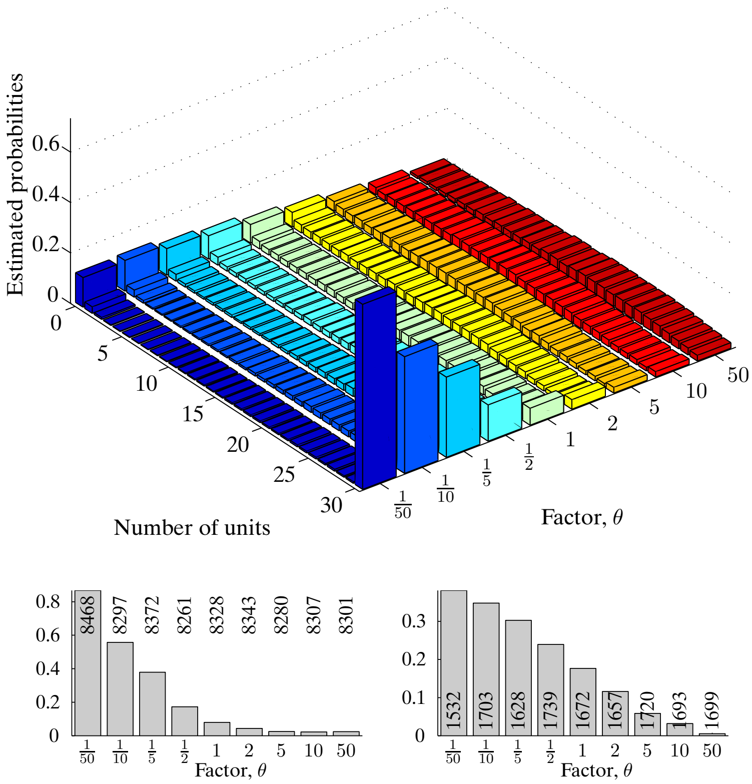

Figure 4.31.
Top: The estimated distribution of the number of units at the middle node for values of θ. All distributions are estimated from 10 000 samples each.
Bottom: The estimated probabilities for an empty and a full middle buffer for each of two cases. To the left is $\pi_3(b_3)$ given that the server is off and to the right, $\pi_3(0)$ given that the server is on. Inside the bars, the number of observations with the corresponding environment 'on' and 'off' are shown. The fraction of samples in the 'on' state is close to $1/6$.

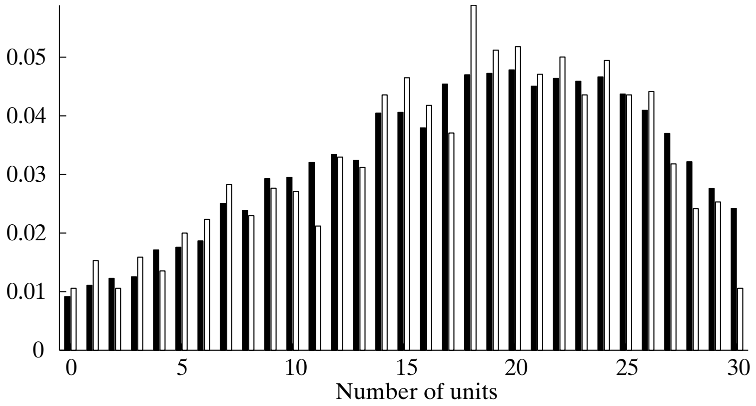

Figure 4.32. *The empirical distribution $\widehat{\pi}_3$ conditioned on an operating server (white bars) and an inoperational one (black bars) for $\theta = 50$. The former distribution was estimated from 1 699 samples, and the latter from 8 301.*

increased batch sizes. Indeed, the arrival instants of batches are coupled, and the stochastic dominance of the batch size distribution gives rise to a stochastic dominance in the number of units present in the network. The coupling of Theorem 2.5.5 is illustrated in Figure 4.33.

With the desire to keep the expected number of arriving units constant over a time interval, the arrival rates are scaled with respect to the batch size distributions. Suppose that the batch sizes at the entrance nodes are deterministic, and the arrival rate of batches of size k, $\beta_1^{(k)}$, is

$$\beta_1^{(k)} = \beta_1/k$$

for some constant k. Then, per time unit, the same expected number of units will try to enter the network, irrespective of the batch size k. Letting α_k be a batch size distribution for node k and $X \stackrel{\mathscr{D}}{=} \alpha_k$ a sample of a batch size, a reasonable choice of arrival rate is $\beta_1^{(k)} = \beta_1/\mathsf{E}[X]$.

Scenario 7 was revised to let the entrance nodes have batch arrivals. Two different settings were considered: one where the expected batch size is 2, and one where it is b_1. Of course, the arrival of a batch greater than, or equal to, b_1 fills the buffer instantaneously. For node 1, the batch sizes are deterministic, and the arrival rates of batches are scaled to $\beta_1^{(2)} = \beta_1/2$ and $\beta_1^{(15)} = \beta_1/15$. Node 2 is subject to random sized batch arrivals, for which the sizes follow a geometric distribution with parameter $q = 1/2$ in the first

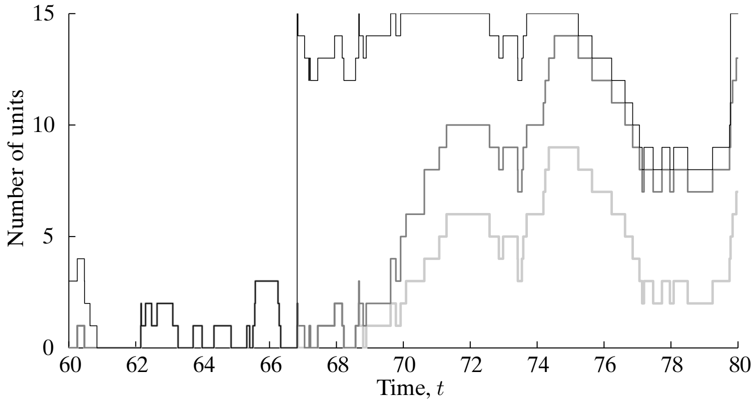

Figure 4.33. *A sample of trajectories of $(X_1(t))_{t\geqslant 0}$ for selected deterministic batch sizes: Single unit arrivals (light grey), double unit arrivals (medium grey), and, batch size 15 (thin black), where $\beta_1^{(k)} = \beta_1$ for $k = 1, 2, 15$. An illustration of the coupling in Theorem 2.5.5.*

setting, and $q = 1/15$ in the second. The symmetry in the network enables the entrance nodes to be mutually compared, and differences between deterministic and randomised batch sizes examined.

Figure 4.34 serves as an illustration of how the trajectories of the queue length process $(X_2(t))_{t\geqslant 0}$ may look for different batch sizes. Note that, in this rerouting network, single units may arrive at the entrance nodes, namely, those units transferring from nodes 4 and 5.

Take 100 000 samples from the stationary distribution of the network for the two settings of batch sizes, and compare them with the distribution estimated from 10 000 samples in scenario 7. In Figure 4.35 the empirical distribution of the number of units at node 1 is shown.

At entrance node 1, the geometrically distributed batch sizes lead to a higher probability of an empty buffer as the expected batch size increases, as well as a higher probability of many units present when $q = 1/15$.

Recall from Table 4.3.a that in scenario 7 the average number of units at node 1 was $m_1 = 1.047$. Here, the average number with randomised batch sizes at node 1, is $m_1^{(2)} = 1.156$ and $m_1^{(15)} = 1.038$. The averages m_1 and $m_1^{(15)}$ are of equal magnitude, but the distribution in the latter case is more disperse with higher probabilities for large and small values of the number

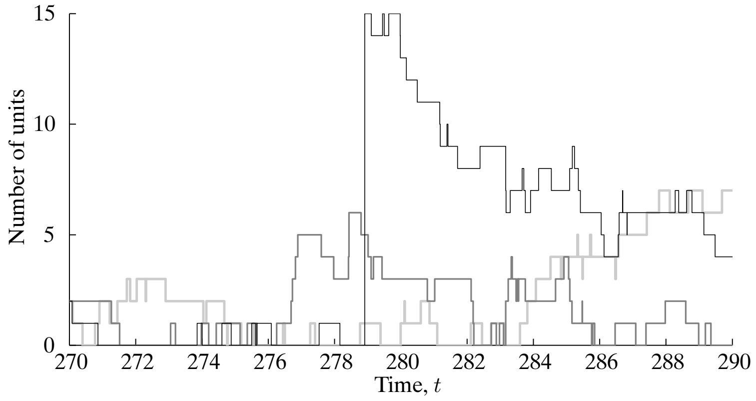

Figure 4.34. *A sample of trajectories of the queue length process of the number of units at node 2,* $(X_2(t))_{t\geqslant 0}$*. The batch sizes are deterministic and trajectories are shown for single unit arrivals (light grey), double unit arrivals (medium grey), and, batch size 15 (thin black).*

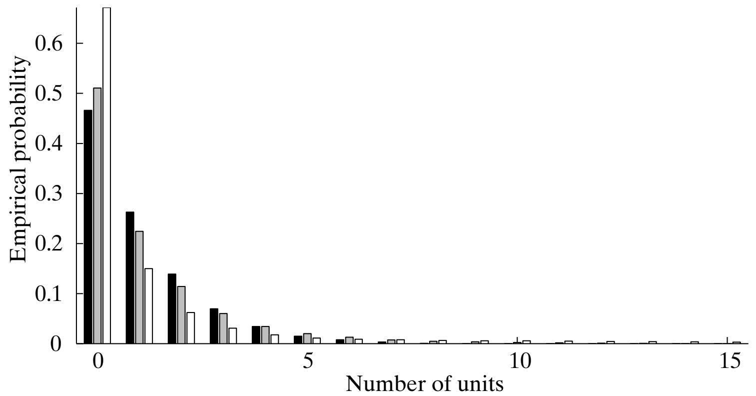

Figure 4.35. *The empirical distributions of the number of units at entrance node 1. Black bars are the distribution in scenario 7, i.e., deterministic single unit arrivals, estimated from 10 000 samples. Grey and black bars are the distributions for the batch sizes that are geometrically distributed random variables with expected values 2 and 15, respectively.*

of units present. The observed average is greatest for geometric batches of expected size 2; the observed standard deviation of X_1 is, using self explanatory notation, $s_1 = 1.36$, $s_1^{(2)} = 1.83$ and $s_1^{(15)} = 2.42$, which increases with the expected batch size.

Node 2 has batch arrivals of deterministic sizes. The qualitative behaviour of $\widehat{\pi}_2$ is the same as for $\widehat{\pi}_1$. Calculating the averages and observed standard deviations of our samples on π_2 for the different settings gives

$$\begin{aligned} m_1 &= 1.075 & s_1 &= 1.39 \\ m_1^{(2)} &= 1.108 & s_1^{(2)} &= 1.74 \\ m_1^{(15)} &= 1.397 & s_1^{(15)} &= 3.06. \end{aligned}$$

Again, the averages and observed standard deviation increase with the batch size.

Now, since not more than $b_1 = 15$ units may be allowed at once at node 1, and the blocking mechanism allows as many units as will fit, the impact on the queueing network of a geometrically distributed batch size X, is the same as that of X', where $\mathsf{P}(X' = k) = \mathsf{P}(X = k),\ k = 1, \ldots, b_1 - 1$ and $\mathsf{P}(X' = b_1) = \mathsf{P}(X \geqslant b_1)$.

The arrival rate of batches of expected size b_1 is the same; since X' is stochastically dominated by the degenerated distribution having a point mass at b_1, according to Theorem 2.5.5, the number of units at each node of the network with geometrically distributed batch sizes is stochastically dominated by the one with deterministic batch sizes, when the expected size is b_1. The expected batch size of X' is $\mathsf{E}[X'] = 9.7$.

The empirical distributions $\widehat{\pi}_1$ and $\widehat{\pi}_2$ are compared in Figure 4.36. Here the number of units at node 1 with geometrically distributed batch sizes is stochastically dominated by the number of units at node 2. Note that the samples on π_1 and π_2 are not independent.

4.2 ANALYSIS BASED ON TRAJECTORIES

Results obtained by analysing trajectories of a stationary SBE network are dealt with briefly here. Samples taken from the stationary distribution of a network are used as initial values for a network process $X = (X_t)_{t \geqslant 0}$. Although the trajectories of X may be used to estimate various interesting

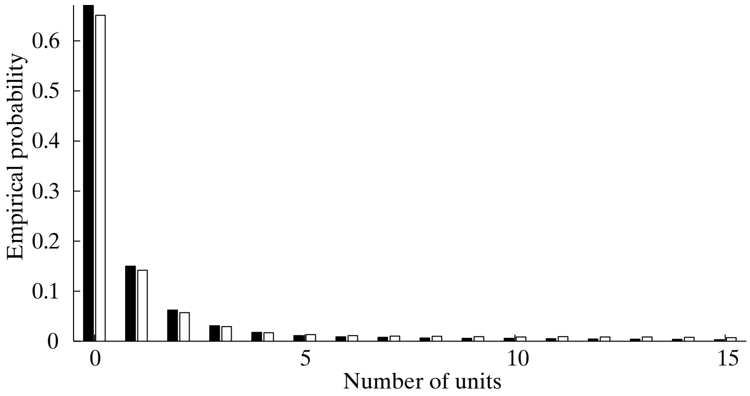

Figure 4.36. *The empirical distributions of the number of units at the entrance nodes 1 and 2 for batches with expected size 15. Black bars are* $\widehat{\pi}_1$*, and white bars are* $\widehat{\pi}_2$.

queueing quantities, only sojourn and waiting times of units are considered here. A question to be addressed is whether one fast server is better than several slow ones. In many situations, the waiting time of a unit is stochastically decreasing with the number of servers, but the sojourn time is stochastically increasing.

Consider the scenario pair 5/13. These rerouting networks have persistent units, large buffers at all the nodes and fast servers at the entrance nodes. The difference between the scenarios is the service rate at the middle node, μ_3, which is small in scenario 5 and large in 13. The scenarios are extended to let the middle node have multiple servers. For m servers, let $\nu_3(n) = \mu_3/m$ for $1 \leqslant n \leqslant m$, $\nu_3(n) = 0$ for $n > m$, and let the service rate at node 3 be

$$\mu_3(x) = \sum_{n=1}^{x_3} \nu_3(n).$$

When all m servers are active, the total departure rate at the node is μ_3. Suppose that the service facility handles units on a first–in, first–out basis, recall Section 2.4.1.

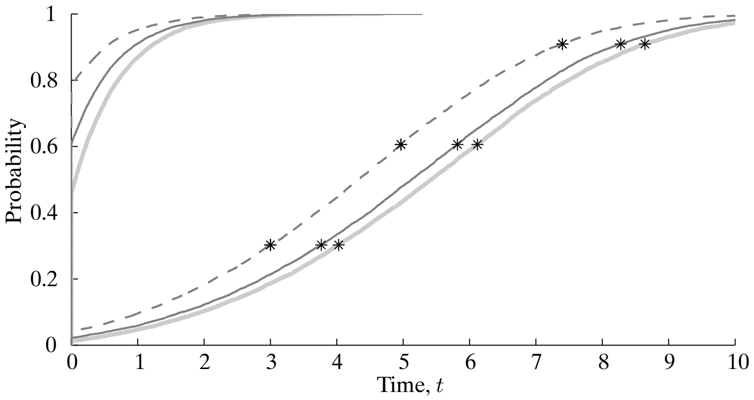

Figure 4.37. *Empirical distribution functions of the waiting times at node 3. The thick grey lines correspond to* $m = 1$ *server, black solid lines to* $m = 2$*, and dashed lines to* $m = 5$*. Lines marked with stars belong to scenario 5 where* μ_3 *is small, and those without to scenario 13.*

4.2.1 WAITING TIMES

The distribution function of the waiting times is estimated by taking n samples of the waiting time (as defined in Section 2.4.2) for the stationary SBE networks previously described. The revised Propp–Wilson algorithm gives a perfect sample, X_0, used as the initial value for the process $(X_t)_{t \geqslant 0}$ from which the first waiting time of an arriving unit to the middle node is measured. The procedure is repeated $n = 10\,000$ times for the different numbers of servers, $m \in \{1, 2, 5\}$, in both scenarios 5 and 13. When $m = 1$, it is as in the original scenarios in Section 4.1. The estimate of $\mathsf{P}(W \leqslant x)$ is the empirical distribution function

$$\widehat{F}(x) = \frac{1}{n} \#\{i : W_i \leqslant x\}.$$

Based on simulation, a graph of these can be found in Figure 4.37.

The waiting time is stochastically decreasing in the number of servers, and the waiting times in scenario 5 are much longer than the corresponding ones in scenario 13. This is natural, since the middle node in the former scenario is overloaded. Since there is a substantial probability of a unit finding the node empty in scenario 13, the waiting time distributions have an atom at zero.

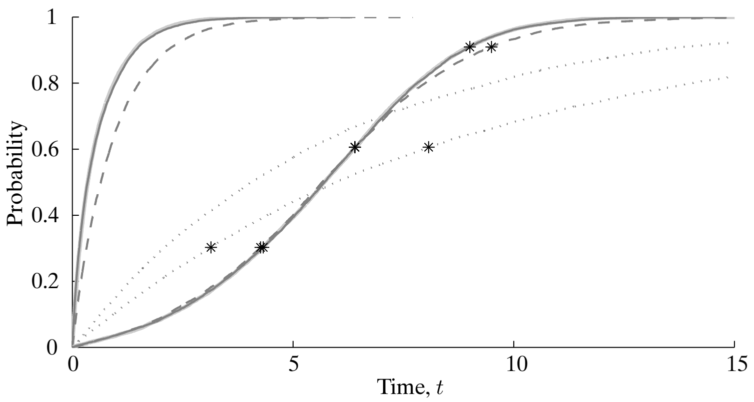

Figure 4.38. *Empirical distribution functions of the sojourn times at node 3. The thick grey lines correspond to* $m = 1$ *server, black solid lines to* $m = 2$*, dashed lines to* $m = 5$ *and dotted lines to* $m = 30$*. Lines marked with stars belong to scenario 5 where* μ_3 *is small, and those without to scenario 13.*

4.2.2 SOJOURN TIMES

Continue as in the previous section, and let S be the sojourn time of a unit at node 3 of the stationary SBE network in scenarios 5 and 13. The number of servers is $m \in \{1, 2, 5, 30\}$ in both scenarios, where $m = 1$ is the original scenarios 5 and 13. When $m = 30$ all arriving units are immediately assigned a server, and the waiting times are all zero. As before, the rate of the m servers is equal to μ_3/m. Based on $n = 10\,000$ samples, the distribution functions of S are estimated in each of the settings and scenarios. The resulting estimates are shown in Figure 4.38.

The difference between $m = 1$ and $m = 2$ is small for both scenarios. When node 3 is overloaded (scenario 5), the sojourn time is not stochastically increasing in the number of servers m. Approximately 47.5% of the units for $m = 30$ have a shorter sojourn time than when $m = 1$.

The distribution of the sojourn time become more dispersed for an increased number of servers. Estimating the standard deviations of S, $\widehat{\sigma}$, for different values of m in scenario 13, gives

m	1	2	5	30
$\widehat{\sigma}$	0.624	0.661	0.935	5.80.

Hence, the variance of the sojourn times is increasing in m. This is also true in scenario 5 where the estimated standard deviations are

m	1	2	5	30
$\widehat{\sigma}$	2.51	2.53	2.77	8.84.

This is beneficial for the units at node 3 in scenario 5. As indicated by the estimates, a low number of servers gives less variation in the sojourn time; hence, the sojourn time is somewhat more predictable. In an overloaded situation, as in scenario 5, the expected sojourn times are long, and units (customers) desiring a high probability of a short sojourn time benefit from high variability, such as having a large number of servers.

To conclude, in scenario-13-like settings where the service rate μ_3 is high, the waiting times of units at node 3 are stochastically decreasing in m but the sojourn times are stochastically increasing. In scenario 5, the former is true but here the sojourn times are not stochastically increasing in m. A small number of servers offers sojourn times with little variability, whereas a large number provides some units very small sojourn times, but at the cost of forcing other units to remain longer at the node.

Chapter 5

Model extensions

The SBE networks are used a model of a queueing system. The following sections contain extensions of the SBE network model to widen the applicability of the models. Not only do the extended models yield a more powerful tool for analysing network performance, but the extensions are also interesting in their own right by indicating what generalisations are readily available. The extensions in this chapter allow us to couple the trajectories of the network processes by embedding them in the same two-dimensional Poisson process, where the probability of common jumps is maximised, i.e., to let the networks have the same stream of arriving units etc.

This chapter starts with a section on different types of environments which may be used in conjunction with networks.

5.1 ENVIRONMENTS

As pointed out in Section 3.2.2, the technique of letting a random environment depend on a subenvironment can always be used. Having a monotone coupling of such an environment enables the use the revised Propp–Wilson algorithm to obtain samples from its stationary distribution. It is then possible to use the sample obtained as an initial state for all trajectories of an upper environment. Models employing this technique should be readily simulated within the framework of this thesis.

The environments need not, of course, be Poisson environments. For example, the coupling already used extensively works for any bounded birth and death process with non-increasing birth intensities and non-decreasing death intensities; indeed, it is simply a single isolated node of an SBE network. Letting an entire SBE network be an environment for another network could be useful in some applications. Suppose one can identify a part of a network, which can be modelled as a Markov chain by itself. It can then be used as an environment for the other part of the network. The service rates, arrival rates and basically everything in the other part of the network, can then be determined, by the functions f_k and g_k, for the state of the new network environment.

An example of such subnetworks is a network where some nodes are isolated from the rest, i.e., if A and B is a partition of $\{1, \ldots, d\}$, and $k \in A$ and $m \in B$ implies $p_{km} = p_{mk} = 0$. Then either subnetwork, the one with node set A or B, could be used as an environment to the other. Although this may be rather artificial, it serves as an example. A more useful criterion is $k \in A$ and $m \in B$ imply

$$p_{km}(1 - c_{km}) = 0 = p_{mk}(1 - c_{mk}),$$

stating that units may go from node set A to B, and vice versa, but if blocked they depart the network entirely: if p_{km} is positive then $c_{km} = 1$.

Consider the feed forward network in Figure 5.1 with a bound on the number of units at the first node and an arrival stream of units to node 2, that is β_2 is positive. Now, let $c_{12}(x) = 1$. Then $(X_1(t))_{t \geqslant 0}$ behaves as single node network in isolation, and it may be used as an environment of $(X_t)_{t \geqslant 0} = (X_1(t), X_2(t))$ on $\mathbb{Z}_+^{\boldsymbol{b}}$. Letting, for example, $f_2(x_1) = 1 - x_1/b_1$, for $x_1 = 0, \ldots, b_1$, and $\beta_2(x) = f_2(x_1)\beta_2 \mathbf{1}_{\{x_2 < b_2\}}$, the external arrival intensity to node 2 is linearly scaled with the number of units at the first node. The empirical distribution, π_2, obtained from $n = 100\,000$ samples

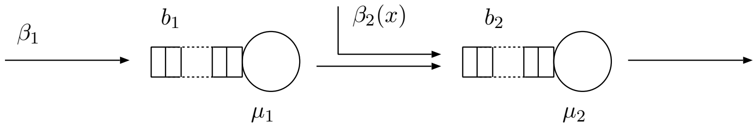

Figure 5.1. *This simple tandem queue is used as an example of an arrival process $\beta_2(x)$ that depends on the number of units at node 1.*

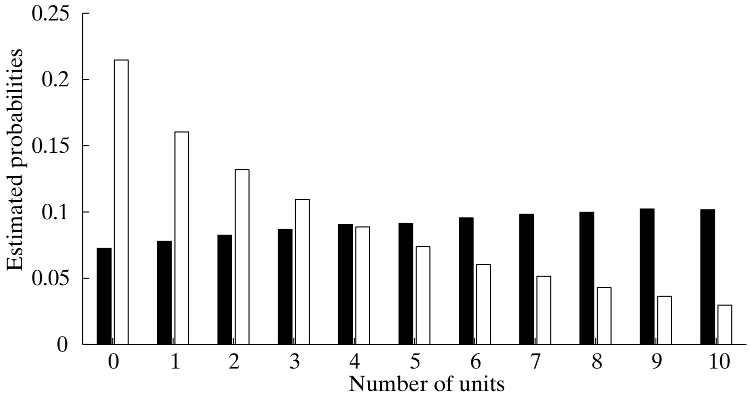

Figure 5.2. *The empirical distribution $\widehat{\pi}_2$ of the network in Figure 5.1 estimated from 100 000 samples. Black bars are $f_2(x_1) = 1 - x_1/b_1$ and white bars $f_2(x_1) = x_1/b_1$.*

of this tandem network is shown in Figure 5.2. The parameters used were $\beta_1 = \beta_2 = 2$, $\mu_1 = \mu_2 = 3$.

In the SBE networks, only the arrival and departure rates were allowed to depend on the environment, through the functions f_k and g_k, respectively. It is possible to let otherwise fixed constants depend on an environment, as is shown in the following section.

5.2 VARYING BUFFER SIZES

The buffer sizes of an SBE network are here described by birth and death processes, to model networks where the capacities of the waiting rooms vary over time.

Let $(B_t)_{t\geqslant 0}$ be a bounded birth and death process on $\{0,\ldots,b_B\}$ for some constant $b_B \in \mathbb{N}$. Assume that the birth rate $\beta^B(b)$ is non-increasing in b and the death rate $\mu^B(b)$ is non-decreasing in b. Clearly, $(B_t)_{t\geqslant 0}$ can be modelled as a single node SBE network, and the two-dimensional Poisson embedding in Section 3.1 can be used in conjunction with the revised Propp–Wilson algorithm to obtain samples on π_B, the stationary distribution of $(B_t)_{t\geqslant 0}$. Indeed, in this simple example, π_B can be readily calculated.

Here, $(B_t)_{t\geqslant 0}$ is a process which determines a time dependent buffer size of a network process $(B_t, X_t)_{t\geqslant 0}$ on $\{0, \ldots, b_B\}\times\mathbb{Z}_+^{\boldsymbol{b}}$. Fix $d \in \mathbb{N}$ and a node $k \in \{1, \ldots, d\}$. The vector $\boldsymbol{b} \in \mathbb{N}^d$ is the vector of bounds of buffer sizes, and $b_k = b_B$. Without further modifications for a d-node SBE network $(X_t)_{t\geqslant 0}$, it is possible to let

$$\beta_k(b, x) = \beta_k(x_k)\mathbf{1}_{\{x_k<b\}},$$

where $\beta_k(x_k)$ is the non-increasing arrival intensity to node k in the network. The previously constant b_k is now replaced by the state of the environment. Note that this is a restriction only on the external arrival of units. Units may still transfer from other nodes of the network when $b \leqslant x_k \leqslant b_B$. If this is not desirable,

$$\gamma_{mk}(x, b) = \mu_m(x_m)p_{mk}a_{mk}(x, b)$$

with $a_{mk}(x, b)$ as the acceptance probabilities, non-increasing in x_m and equal to zero if $x_m = b$.

Since the coupling used thus far, forces all trajectories to have the same environment trajectory, and holds at all times for all values of b_k, it also holds for a varying buffer size $(B_t)_{t\geqslant 0}$. Hence, the network components of the coupling, $((X_t, B_t), (X'_t, B_t))$, satisfy $X_t \preceq X'_t$ for all $t \geqslant 0$ whenever $X_0 \preceq X'_0$. At the current setting, no units are accepted at node k if $x_k \geqslant b$, but one may observe $x_k > b$ anyway. In the following modification of the network model, the coupling is modified as well.

The environment evolves independently of the network state and the buffer process $(B_t)_{t\geqslant 0}$ may have 'deaths', reducing the buffer size. Suppose that the environment is in state $b > 0$ and $x_k = b$ when the buffer size is reduced. This leaves the buffer process in state $b-1$ but the number of units at node k is still b. Thus, for $X_k(0) \leqslant B_0$, there is a need to modify the transition rates from (x, b) to ensure that $X_k(t) \leqslant B_t$ for all $t \geqslant 0$. A possible choice of transition rates of $(X_t, B_t)_{t\geqslant 0}$ from (x, b) shall be devised, where $(B_t)_{t\geqslant 0}$ is a d-node SBE network such that the number of units at node k determines the number of available buffer positions of node k in $(X_t)_{t\geqslant 0}$.

Fix $d \in \mathbb{N}$ and $\boldsymbol{b}^B \in \mathbb{N}^d$, the maximal buffer sizes at nodes $1, \ldots, d$. Let the SBE network $(B_t)_{t\geqslant 0}$ have state space $\mathbb{Z}_+^{\boldsymbol{b}^B}$ and denote its parameters by

β_k^B, μ_k^B. For simplicity, it is assumed that the environment $(B_t)_{t\geqslant 0}$ consists of d isolated nodes, $p_{km}^B = 0$ for all $k, m \in \{1,\ldots,d\}$. The reader should have no trouble generalising to an example with $p_{km}^B > 0$ for some k and m.

The transition intensities for $(X_t, B_t)_{t\geqslant 0}$ on $\mathbb{Z}_+^{\boldsymbol{b}^B} \times \mathbb{Z}_+^{\boldsymbol{b}^B}$ when $(X_t, B_t) = (x, b)$ are, for $k, m \in \{1,\ldots,d\}$,

to	**with intensity**	
$(x+e_k, b)$	$\beta_k(x,b)$	
$(x-e_k, b)$	$\delta_k(x,b)$	(5.1a)
$(x-e_k+e_m, b)$	$\gamma_{km}(x,b)$	
$(x, b+e_k)$	$\beta_k^B(b)$	(5.1b)
$(x, b-e_k)$	$\mu_k^B(b)\mathbf{1}_{\{x_k<b_k\}}$	
$(x-e_k, b-e_k)$	$\mu_k^B(b)\mathbf{1}_{\{x_k=b_k\}}p_k^*(x,b)$	(5.1c)
$(x-e_k+e_m, b-e_k)$	$\mu_k^B(b)\mathbf{1}_{\{x_k=b_k\}}p_{km}^*(x,b)$.	

Suppose that the intensities $\beta_k(x,b)$, for example, are the same as in an SBE network, except that b describes the current state of the buffers. For example, one may choose $\beta_k(x,b) = \beta_k\mathbf{1}_{\{x_k<b_k\}}$ as usual. This also applies for $\mu_k(x,b)$ and $\gamma_{km}(x,b)$. The only new part is (5.1c). This results in a network model in which surplus units at a node, i.e., units at a buffer that does not have room for them, are forced to leave the node. Here, there is a probability $p_k^*(x,b)$ of forcing them to leave the network entirely, and a probability $p_{km}^*(x,b)$ of rerouting them to node m.

It is required that $p_k^*(x,b) = 1 - \sum_{m=1}^d p_{km}^*(x,b)$, $p_{kk}^*(x,b) = 0$, for all $x, b \in \mathbb{Z}_+^{\boldsymbol{b}^B}$, and $p_{km}^*(x,b)$ is non-increasing in x_m. Under these assumptions p_{km}^* are arbitrary, but a natural choice is

$$p_{km}^*(x,b) = p_{km}a_{km}(x,b)$$

and

$$p_k^*(x,b) = p_k + p_{kk} + \sum_{\substack{m=1,\ldots,d \\ m\neq k}} p_{km}(1 - a_{km}(x,b)).$$

Here the expunged surplus units are routed forward in the network as if they had received service, with the exception that a unit may not be routed back to node k again, and the desistance probabilities are all one.

Examining the marginal processes $(X_t)_{t\geqslant 0}$, $(B_t)_{t\geqslant 0}$ of $(X_t, B_t)_{t\geqslant 0}$ reveals that, if $(B_t)_{t\geqslant 0}$ is constant, then $(X_t)_{t\geqslant 0}$ is a Markov chain behaving as an SBE network on $\mathbb{Z}_+^{\boldsymbol{b}}$ with $\boldsymbol{b} = B_0$. Again, what makes the difference is the transitions in (5.1c), which do not correspond to anything in an SBE network. The process $(B_t)_{t\geqslant 0}$ is a Markov chain, an SBE network, hence a random environment to $(X_t)_{t\geqslant 0}$. What remains is to couple $(X_t, B_t)_{t\geqslant 0}$ and $(X'_t, B'_t)_{t\geqslant 0}$, so that $(X_0, B_0) = (x_0, b)$, $(X'_0, B'_0) = (x'_0, b)$ and $x_0 \preceq x'_0$ implies $X_t \preceq X'_t$ for all $t \geqslant 0$. Note that only $(B_t)_{t\geqslant 0} = (B'_t)_{t\geqslant 0}$ [a.s.] needs to be taken into account.

The coupling of (X_t, B_t) and (X'_t, B'_t), $(\widetilde{X}_t, \widetilde{B}_t, \widetilde{X}'_t, \widetilde{B}_t)$, has transition intensities (5.2) from state (x, b, x', b), abbreviated (x, x', b). Compare this with equation (2.19).

For $k, m = 1, \ldots, d$

$$
\begin{aligned}
&(x+e_k, x', b) && \beta_k(x,b) - \beta_k(x',b) \\
&(x+e_k, x'+e_k, b) && \beta_k(x',b)
\end{aligned}
\tag{5.2a}
$$

$$
\begin{aligned}
&(x, x-e_k, b) && [\mu_k(x') - \mu_k(x)]\left(p_k + \textstyle\sum_{m=1}^{d} p_{km}(1-a_{km}(x',b))c_{km}(x',b)\right) \\
&(x-e_k, x', b) && \mu_k(x)\left(p_k + \textstyle\sum_{m=1}^{d} p_{km}(1-a_{km}(x,b))[c_{km}(x,b) - c_{km}(x',b)]\right) \\
&(x-e_k, x'-e_k, b) && \mu_k(x)\left(p_k + \textstyle\sum_{m=1}^{d} p_{km}(1-a_{km}(x,b))c_{km}(x',b)\right)
\end{aligned}
\tag{5.2b}
$$

$$
\begin{aligned}
&(x, x'-e_k+e_m, b) && [\mu_k(x') - \mu_k(x)]p_{km}a_{km}(x',b) \\
&(x-e_k+e_m, x', b) && \mu_k(x)p_{km}[a_{km}(x,b) - a_{km}(x',b)](1-c_{km}(x',b)) \\
&(x-e_k+e_m, x'-e_k, b) && \mu_k(x)p_{km}[a_{km}(x,b) - a_{km}(x',b)]c_{km}(x',b) \\
&(x-e_k+e_m, x'-e_k+e_m, b) && \mu_k(x)p_{km}a_{km}(x',b)
\end{aligned}
\tag{5.2c}
$$

$$
\begin{aligned}
&(x, x', b+e_k) && \beta_k^B(b) \\
&(x, x', b-e_k) && \mu_k^B(b)\mathbf{1}_{\{x'_k<b_k\}} \\
&(x, x'-e_k, b-e_k) && \mu_k^B(b)\mathbf{1}_{\{x_k<b_k\}}\mathbf{1}_{\{x'_k=b_k\}}p_k^*(x',b) \\
&(x, x'-e_k+e_m, b-e_k) && \mu_k^B(b)\mathbf{1}_{\{x_k<b_k\}}\mathbf{1}_{\{x'_k=b_k\}}p_{km}^*(x',b) \\
&(x-e_k, x'-e_k, b-e_k) && \mu_k^B(b)\mathbf{1}_{\{x_k=b_k\}}p_k^*(x,b) \\
&(x-e_k+e_m, x'-e_k, b-e_k) && \mu_k^B(b)\mathbf{1}_{\{x_k=b_k\}}[p_{km}^*(x,b) - p_{km}^*(x',b)] \\
&(x-e_k+e_m, x'-e_k+e_m, b-e_k) && \mu_k^B(b)\mathbf{1}_{\{x_k=b_k\}}p_{km}^*(x',b)
\end{aligned}
\tag{5.2d}
$$

equations (5.2a)–(5.2c) are the same as (2.19a)–(2.19c). Using $\mathbf{1}_{\{x_k=b_k\}} = \mathbf{1}_{\{x_k=b_k\}}\mathbf{1}_{\{x'_k=b_k\}}$, simple examination reveals that the marginals have the correct distributions,

$$(\widetilde{X}_t, \widetilde{B}_t)_{t\geqslant 0} \stackrel{\mathscr{D}}{=} (\widetilde{X}'_t, \widetilde{B}_t)_{t\geqslant 0}$$

with transition rates (5.1). What remains is to show that the coupling has the desired monotone property in the trajectories $(\widetilde{X}_t)_{t\geqslant 0}$ and $(\widetilde{X}'_t)_{t\geqslant 0}$.

The coupling of transitions for which b is constant is the same as in (2.19), hence, these transitions all preserve the ordering. The only potentially troublesome transitions are from (x, b, x', b) to $(x, b-e_k, x'-e_k, b-e_k)$, $(x, b-e_k, x'-e_k+e_m, b-e_k)$ and $(x-e_k+e_m, b-e_k, x'-e_k, b-e_k)$. They are analysed in that order.

- $(x, b-e_k, x'-e_k, b-e_k)$ can occur only if $x_k < b_k = x'_k$, hence $x \preceq x' - e_k$.
- $(x, b-e_k, x'-e_k+e_m, b-e_k)$ uses the same argument as above.
- $(x-e_k+e_m, b-e_k, x'-e_k, b-e_k)$ has positive intensity only if $p^*_{km}(x, b-e_k) > p^*_{km}(x', b-e_k)$. Remembering that $p^*_{km}(x, b)$ is non-increasing in x_m, this implies $x_m < x'_m$, hence $x - e_k + e_m \preceq x' - e_k$.

To conclude, networks $(X_t, B_t)_{t\geqslant 0}$ and $(X'_t, B_t)_{t\geqslant 0}$, with transition intensities (5.1) can be coupled by (5.2), which preserves the ordering of the trajectories $(X_t)_{t\geqslant 0}$ and $(X'_t)_{t\geqslant 0}$ at all times t.

5.3 REROUTING OF BLOCKED UNITS

A model which handles rerouting of blocked external arriving units will now be devised. In this type of networks, a unit arriving from the the outside and finding the destination node to be full does not have to leave but may choose to enter another node.

Take $d \in \mathbb{N}$ and let $(r_{km})_{m=1}^d$ be a probability measure on $\{1, \ldots, d\}$ for every $k \in \{1, \ldots, d\}$. Suppose that $r_{kk} = 0$ and let $r_k = 1 - \sum_{m=1}^d r_{km}$, for every $k = 1, \ldots, d$. The group $(r_{km})_m$ is called *rerouting probabilities*, analogous to the routing probabilities $(p_{km})_{m=1}^d$.

While the multivariate birth and death process about to be constructed is similar to an SBE network, it has rerouting of blocked external arrivals according to the rerouting probabilities. Units arriving at a full node, for example $x_k = b_k$, have a probability r_{km} of being rerouted to node m. The arrival intensity to node k then becomes

$$\beta_k(x) = \left(\beta_k(x_k) + \sum_{m=1}^{d} \beta_m(x_m) \mathbf{1}_{\{x_m = b_m\}} r_{mk} \right) \mathbf{1}_{\{x_k < b_k\}}, \tag{5.3}$$

$k = 1, \ldots, d$. Note that this set of intensities can be interpreted as if only one rerouting attempt of a blocked unit is made.

To achieve a coupling similar to (2.19), take $\boldsymbol{b} \in \mathbb{N}^d$ and let $E = \mathbb{Z}_+^{\boldsymbol{b}}$. Suppose that $x, x' \in E$ satisfies $x \preceq x'$. Then let $(\widetilde{X}_t, \widetilde{X}'_t)$ have transition intensities given by (2.19), but where

to	with intensity
$(x, x' + e_k)$	$\beta_k(x') - \beta_k(x) \wedge \beta_k(x')$
$(x + e_k, x')$	$\beta_k(x) - \beta_k(x) \wedge \beta_k(x')$
$(x + e_k, x' + e_k)$	$\beta_k(x) \wedge \beta_k(x')$

(5.4)

replaces (2.19a).

It should be clear that the marginal distributions of $\widetilde{X}_t$ and $\widetilde{X}'_t$ are the desired ones. Further, note that $x_k = x'_k$ implies $\beta_k(x) \leqslant \beta_k(x')$. Hence, $\beta_k(x) - \beta_k(x) \wedge \beta_k(x') = 0$, and transitions to $(x + e_k, x')$ can occur only if $x_k < x'_k$, after which $x + e_k \preceq x'$ still holds. Since the remaining transition intensities, given by (2.19b)–(2.19d), preserve the ordering, the conclusion is that $\widetilde{X}_0 \preceq \widetilde{X}'_0$ implies $\widetilde{X}_t \preceq \widetilde{X}'_t$ for all $t \geqslant 0$.

Thus, the coupling of networks constructed with arrival intensities given by (5.3) preserves the ordering of the trajectories $(\widetilde{X}_t)_{t \geqslant 0}$ and $(\widetilde{X}'_t)_{t \geqslant 0}$ at all times t. The revised Propp–Wilson algorithm may now be used to obtain samples from the stationary distribution.

5.4 COMMON ARRIVALS

Recall Section 1.2.2 with the definition of a multivariate birth and death process $(X_t)_{t \geqslant 0}$ on $\mathbb{Z}_+^d$. For subsets $B, D \subseteq \{1, \ldots, d\}$ satisfying

$$B \cap D = \varnothing \qquad B \cup D \neq \varnothing$$

the transition rates of $(X_t)_{t\geqslant 0}$ were positive only for transitions from x to y where

$$x + e_B = y + e_D.$$

An imagined application where more than atomic birth sets B are considered is given next.

There exist many models of parallel processor systems. One is that an SBE network with d nodes represents the state of multiple processors, where units are calculations to be performed by the particular processor in the buffer of which they are placed. A job consists of several calculations; it is the jobs that arrive to the processor system according to a Poisson process. When a job arrives, the calculations are simultaneously spread out to the processors, i.e., units arrive simultaneously to all the nodes of our network.

Perhaps a superposition of interrupted Poisson processes is a better model of the arrival process, if one imagines applications as being started by independent users, running for an exponential time generating calculations to the processor system, and then being terminated. For the sake of simplicity, however, the environment needed to accomplish this is not constructed, although the reader should be aware that, in practise, this could easily be done.

Suppose that the birth set is $B \subseteq \{1, \dots, d\}$, and the intensity of jobs is $\gamma_{\varnothing,B}(x) = \beta_B(x)$. One must determine how the blocking mechanism should work, if some of the nodes are full when a joint arrival to all the nodes in B occurs. The policy of accepting units at all nodes at which they will fit, and thus reducing the birth set B when some nodes are full, leads to the following model.

Fix a birth set $B \subseteq \{1, \dots, d\}$ and let the arrival rate satisfy

$$\beta_B(x) = \beta_B \prod_{k\in B} \mathbf{1}_{\{x_k < b_k\}}$$

for some constant $\beta_B > 0$, and for all non-empty subsets $B' \subseteq B$,

$$\beta_{B'}(x) = \beta_B \prod_{k\in B'} \mathbf{1}_{\{x_k < b_k\}} \prod_{k\in B\setminus B'} \mathbf{1}_{\{x_k = b_k\}}.$$

In other words, for a fixed birth set B, units do not enter nodes forming a subset of B, unless the buffers, at those nodes not contained in the subset, are full. This is a statement about the blocking mechanism, not the nature of the arrival process. The intensity $\beta_{B'}(x)$, induced by the intensity $\beta_B(x)$

and the birth set B, should not be considered an intensity of another birth set B'. The factor $\prod_{k\in B\setminus B'} \mathbf{1}_{\{x_k=b_k\}}$ in $\beta_{B'}(x)$ makes all the difference.

To ease the notation, for each birth set B, define functions

$$\Delta_G^B(x) = \prod_{k\in G} \mathbf{1}_{\{x_k<b_k\}} \prod_{k\in B\setminus G} \mathbf{1}_{\{x_k=b_k\}},$$

for non-empty subsets $G \subseteq B$. The notation

$$\beta_B \Delta_{B'}^B(x) \text{ for } \beta_{B'}(x)$$

is henceforth used to emphasise that the latter is induced from a particular birth set B. Note also that $\Delta_B^B(x) = \prod_{k\in B} \mathbf{1}_{\{x_k<b_k\}}$, hence, $\beta_B\Delta_B^B(x) = \beta_B(x)$.

Since each component of a vector $x \in \mathbb{Z}_+^{\boldsymbol{b}}$ must satisfy either $x_k < b_k$ or $x_k = b_k$, there exists for each x a unique set G for which $\Delta_G^B(x) = 1$; for $H \neq G$, $\Delta_H^B(x) = 0$.

An SBE network analog with simultaneous arrivals may be defined as the d-dimensional birth and death process, with transition rates as in an SBE network, but where $\beta_B(x) > 0$ for some birth sets $B \subseteq \{1, \dots, d\}$. A transition generated by $\beta_B(x)$ is called a *set transition* to distinguish it from our usual atomic transitions. One may have several birth sets $B_1, \dots, B_n$, and they may overlap. Each of them gives rise to induced birth rates $\beta_{B'}^{B_i}(x)$, $B' \subseteq B_i$, $i = 1, \dots, n$, of set transitions. The desired coupling is devised only for a single birth set B; the generalisation to several sets is immediate.

Two identical copies $(X_t)_{t\geqslant 0}$ and $(X'_t)_{t\geqslant 0}$ of an SBE network, with set transitions induced by a birth set B, may be coupled by letting the transition intensities for $(\widehat{X}_t, \widehat{X}'_t)_{t\geqslant 0}$ from state (x, x'), in addition to (2.19), be

to	**with intensity**
$(x + e_G, x' + e_H)$	$\beta_B\Delta_G^B(x)\Delta_H^B(x')$

for all non-empty subsets $G, H \subseteq B$.

Admittedly, this is a curious looking coupling, but if one considers the marginals, it can be seen that $(\widehat{X}_t)_{t\geqslant 0}$ has transition rates from x to $x + e_G$:

$$\sum_{H\subseteq B} \beta_B\Delta_G^B(x)\Delta_H^B(x') = \beta_B\Delta_G^B(x) \sum_{H\subseteq B} \Delta_H^B(x') = \beta_B\Delta_G^B(x),$$

remembering that only one set H satisfies $\Delta_H^B(x') > 0$, and then $\Delta_H^B(x') = 1$. The transition rates for $(\widehat{X}_t')_{t\geqslant 0}$ from x' to $x' + e_H$ are similar. Hence, the marginals have the correct transition rates.

Now, for each state (x, x'), $x \preceq x'$, of $(\widehat{X}_t, \widehat{X}_t')_{t\geqslant 0}$, there is a unique pair G, H corresponding to a state $(x + e_G, x' + e_H)$ with positive intensity. Keeping in mind how the blocking mechanism works for simultaneous arrivals, it should be clear what the state is. What is important is that $(x + e_G, x' + e_H)$ has positive intensity only if $\Delta_G^B(x)\Delta_H^B(x') = 1$. Further, $x \preceq x'$ implies $H \subseteq G$. Thus, the chain has a set transition to $(x + e_H + e_{G\backslash H}, x' + e_H)$. However, $(G\backslash H) \cap H = \varnothing$ so $x'_{G\backslash H} = b_{G\backslash H}$ and hence, $x + e_G \preceq x' + e_H$, show dominance between the states after the transition. Since all other (atomic) transitions in (2.19) preserve the ordering between the trajectories, a proof that $\widehat{X}_0 \preceq \widehat{X}_0'$ implies $\widehat{X}_t \preceq \widehat{X}_t'$ for all $t \geqslant 0$, is complete.

To see an example of a network with common arrivals, consider the network in Figure 4.1, and let the birth set $B = \{1, 2\}$. Thus, the entrance nodes have simultaneous arrivals, and as many units as will fit may enter. The parameter setting used is the one corresponding to scenario 61: that is, $\boldsymbol{b} =$ (15, 15, 30, 10, 10), $\beta_B = 2$, $\mu_1 = \mu_2 = 2.5 = \mu_4 = \mu_5$, $\mu_3 = 5.0$, no rerouting of units, $p = 0$, and persistent users $c = 0$. Since $\beta_B = 2$,

$$\beta_B(x) = \beta_B\Delta_B^B(x) = \beta_B(x) = 2 \cdot \mathbf{1}_{\{x_1<b_1\}}\mathbf{1}_{\{x_2<b_2\}},$$

and the induced transition rates are $\beta_{B'}(x) = \beta_B\Delta_{B'}^B(x)$ or, plainly,

$$\beta_{\{1\}}(x) = 2 \cdot \mathbf{1}_{\{x_1<b_1\}}\mathbf{1}_{\{x_2=b_2\}} \qquad \beta_{\{2\}}(x) = 2 \cdot \mathbf{1}_{\{x_2<b_2\}}\mathbf{1}_{\{x_1=b_1\}}.$$

This means that units arrive at both of the entrance nodes, 1 and 2, at the same time, and all units that fit into a buffer are accepted.

Taking $n = 10\,000$ samples $X_1, \ldots, X_n$, from the stationary distribution of the network, allows us to estimate the distribution of the number of units at the different nodes. This is compared with the distribution of units in an SBE network with arrival rates

$$\beta_1(x) = 2 \cdot \mathbf{1}_{\{x_1<b_1\}} \qquad \beta_2(x) = 2 \cdot \mathbf{1}_{\{x_2<b_2\}}.$$

The rates are chosen so that the traffic intensity in both of the networks is the same: $\rho = 0.80$, i.e., a slightly underloaded setting. In Figure 5.3 the estimated tail distribution

$$\widehat{\mathsf{P}}(X_1 > x) = 1 - \widehat{F}(x) = \frac{1}{n}\#\{i : X_{i,1} > x\}.$$

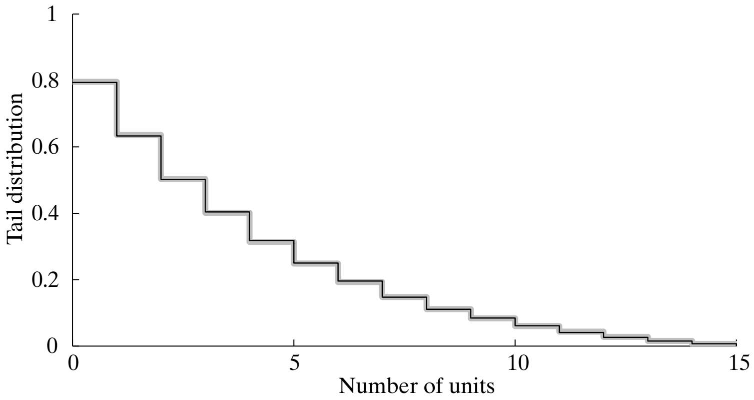

Figure 5.3. *The thick grey line shows the estimated tail probability* $\widehat{\mathsf{P}}(X_1 > x)$ *of the buffer length at the first node in the network with common arrivals at the entrance nodes. The thin black line is the corresponding estimated probability for the SBE network in scenario 61. The estimate is based on 10 000 samples.*

is shown for the two networks. Here $x_{k,l}$ is the l^{th} component of observation X_k. Figure 5.4 shows the estimated tail distribution for the 3$^{\mathrm{rd}}$ node of the networks.

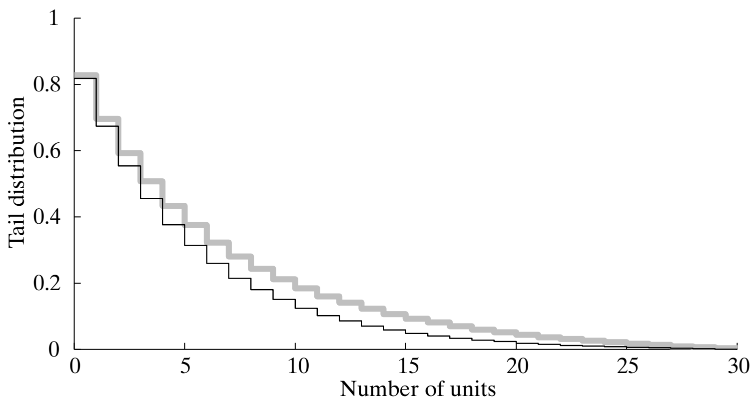

Figure 5.4. *The thick grey line shows the estimated tail probability* $\widehat{\mathsf{P}}(X_3 > x)$ *of the buffer length at the middle node in the network with common arrivals. The thin black line is the corresponding estimated probability for the SBE network. Based on 10 000 samples.*

Chapter 6

Conclusions

To conclude, a model of queueing networks with blocking was defined, which can be used for a wide range of applications satisfying the Markovian assumptions. For the model, an algorithm to obtain samples from the stationary distribution of the number of units in the network was constructed.

The network model has a blocking mechanism which works in such way that units are allowed to transfer between two nodes with a probability that depends on the number of units at the destination node. Blocked units are assigned a probability to remain at the originating node or leave the network entirely. This model was also extended to cover the situation of blocked units that have another set of rerouting probabilities. In the fundamental SBE model of the networks, a random environment was introduced as a series of independent ON/OFF processes, and the arrival and service intensities were allowed to be functions of the environment process. The model also provides for batch arrivals.

Properties of the network model were established by coupling techniques. The networks were shown to have certain monotonic properties, for example, that the number of units is stochastically increasing in the arrival rates and decreasing in the desistance probabilities. Some of the monotonic properties featured in Jackson networks are not preserved for the blocking networks. Variation of parameter settings for a sample network changed its fundamental properties; an increased buffer size could lead to either more or fewer units in the network. In some parameter configurations, the number of units in the network increased with the service rate. When a few fast servers were compared with several slow ones, the latter did not necessarily give stochastically longer sojourn times of units at the nodes.

A coupling of trajectories of network processes was constructed by means of a Poisson embedding which enabled the efficient application of a monotone Propp–Wilson algorithm to obtain samples from the stationary distribution. The efficiency of the algorithm was examined by a series of simulation experiments for different buffer sizes, traffic loads and network topologies.

The algorithm is effective for the SBE networks, in the sense that samples from the stationary distributions can be obtained for large and complex networks in a reasonable amount of time. Although the increase of simulation time measured in the number of iterations was hard to predict for some network topologies, it was always reasonably small, so that an actual implementation of the algorithm can be run on ordinary computer equipment.

Future research could include other types of environment processes as well as establishing bounds on the convergence times of the perfect sampling algorithm. The couplings devised rely heavily on the assumptions of exponential service times and Markovian routing of the network processes. It is possible, however, while retaining these assumptions, to work with network models in which units are subdivided into classes, and routing probabilities are assigned differently to units in separate classes. Other extensions could include other state spaces, for example spaces corresponding to closed networks, and other types of blocking mechanisms.

During a workshop in perfect simulation held in Denmark in 1997, a joint work between Lund and Wilson, Lund and Wilson [14], 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 units. Their generalisation to an infinite dam would lead 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. *Oper. Res. Lett.*, **16**, 177–181, 1994.

[2] Anantharam, V. & Tsoucas, P. Stochastic concavity of throughput in series of queues with finite buffers. *Adv. Appl. Prob.*, **22**, 761–763, 1990.

[3] Barndorff-Nielsen, O. E., Jensen, J. L., & Kendall, W. S. (editors). *Networks and Chaos—Statistical and Probabilistic Aspects*. Chapman & Hall, London, 1993.

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

[5] Brémaud, P. *Markov Chains : Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York, 1999.

[6] Daley, D. J. & Vere-Jones, D. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York, 1988.

[7] Dimakos, X. K. A guide to exact simulation. *Int. Stat. Rev.*, **69**(1), 27–48, 2001.

[8] Fill, J. A. An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.*, **8**(1), 131–162, 1998.

[9] GOTLIEB, C. C. & GOTLIEB, L. R. *Data Types and Structures*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1978.

[10] HÄGGSTRÖM, O. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, Cambridge, 2002 (to appear).

[11] LINDVALL, T. *Lectures on the Coupling Method*. John Wiley & Sons Inc., New York, 1992.

[12] LINDVALL, T. Stochastic monotonicities in Jackson queueing networks. *Probab. Engrg. Inform. Sci.*, **11**, 1–9, 1997.

[13] LINDVALL, T. On Strassen's theorem on stochastic domination. *Elect. Comm. in Probab.*, **4**, 51–59, 1999.

[14] LUND, R. B. & WILSON, D. B. Exact sampling algorithms for storage systems. 1997. Manuscript.

[15] LYONS, R. & PERES, Y. *Probability on Trees and Networks*. Cambridge University Press, Cambridge, 2002 (to appear). Preprint available at `http://www.math.gatech.edu/~rdlyons`.

[16] MATTSSON, D. On perfect simulation of certain queueing networks. 2000. Licentiate thesis, Department of Mathematics, Chalmers University of Technology.

[17] MEESTER, L. E. & SHANTHIKUMAR, J. G. Concavity of the throughput of tandem queueing systems with finite buffer storage space. *Adv. Appl. Prob.*, **22**, 764–767, 1990.

[18] NORRIS, J. R. *Markov Chains*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

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

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

[21] SERFOZO, R. *Introduction to Stochastic Networks*. Springer, New York, 1999.

[22] SHAKED, M. & SHANTHIKUMAR, J. G. *Stochastic Orders and Their Applications*. Academic Press, San Diego, 1994.

[23] SHANTHIKUMAR, J. G. & YAO, D. D. Stochastic monotonicity in general queueing networks. *J. Appl. Prob.*, **26**, 413–417, 1989.

[24] SPARAGGIS, P. D. & CASSANDRAS, C. G. Throughput monotonicity in communication networks with blocking: Properties and counter examples. In *Infocom proceedings 91 April 7–11, 1991, Bal Harbour, FL, USA IEEE computer society*, p. 101–110. IEEE, New York, 1991.

[25] SPRNG. Scalable Parallel Pseudo Random Number Generators Library. 6 Sep. 2001. URL `http://sprng.cs.fsu.edu/`. (Retrieved 26 Feb. 2002).

[26] THORISSON, H. *Coupling, Stationarity, and Regeneration*. Springer, New York, 2000.

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

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

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

[30] WOLFF, R. W. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1989.

Index

List of symbols

Symbol	Definition	Context
$\mathbb{Z}_+$	$\{0, 1, 2, \dots\}$	non-negative integers
$\mathbb{N}$	$\{1, 2, \dots\}$	positive integers
$\mathbf{1}_{\{A\}}$	$\begin{cases} 1 & \text{when } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$	indicator
δ_x	$\mathbf{1}_{\{x \in \cdot\}}$ $\mathbf{1}_{\{x = \cdot\}}$	Dirac's delta as a set function or as a function on $\mathbb{R}$
δ_{xy}	$\delta_x(y)$	δ_x evaluated at y
$<x, e_A>$	$(x_k)_{k \in A}$	coordinate mapping
$\preceq$		partial ordering of $\mathbb{R}^d$, see page 24
$\overset{\mathscr{D}}{\preceq}$		stochastic domination with respect to a partial ordering, see page 25
$\leqslant^*$		random ordering, see page 71
$\#A$		number of elements in the set A
$x \wedge y,\ x \vee y$	$\min(x, y),\ \max(x, y)$	minimum and maximum, respectively