This dissertation has been
microfilmed exactly as received

69-18,131

WALLACE, Victor Lew, 1933-

THE SOLUTION OF QUASI BIRTH AND DEATH PROCESSES ARISING FROM MULTIPLE ACCESS COMPUTER SYSTEMS.

The University of Michigan, Ph.D., 1969
Computer Science

University Microfilms, Inc., Ann Arbor, Michigan

THE SOLUTION OF QUASI BIRTH AND DEATH PROCESSES ARISING FROM MULTIPLE ACCESS COMPUTER SYSTEMS

by

Victor Lew Wallace

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
The University of Michigan
1969

Doctoral Committee:

Professor Keki B. Irani, Chairman
Associate Professor Kuei Chuang
Professor Ralph L. Disney
Associate Professor Eugene L. Lawler
Associate Professor Arch W. Naylor

ABSTRACT

THE SOLUTION OF QUASI BIRTH AND DEATH PROCESSES ARISING FROM MULTIPLE ACCESS COMPUTER SYSTEMS

by

Victor Lew Wallace

Chairman: Keki B. Irani

Quasi birth and death processes (QBD-processes) are continuous time Markov chains whose transition intensity matrices can be represented in tri-diagonal block form, with the blocks along a diagonal being identical submatrices after the first. They are a general class of infinite state Markov chains which are useful as models of the "high traffic" behavior of large, multiple-access computer systems. They are particularly useful in analysis and design of those systems which are often described as "computer utilities".

Classical queueing theory offers no general technique for determination of equilibrium distributions of QBD-processes, particularly not for QBD-processes of the degree of complexity likely to be encountered in computer applications. Furthermore, because the transition intensity matrix of a QBD-process is of infinite order, commonly used numerical techniques for determination of equilibrium probabilities are also not applicable. However the equilibrium equations can be shown to have a formal similarity to the equilibrium equations of simple birth and death processes with matrices and vectors in the

former replacing scalars in the latter. R. V. Evans has proposed that this similarity be used to derive a numerical procedure which involves several separate finite operations, each of which does not tax the memory or computing capabilities of a moderately large digital computer. Fundamental to this procedure is the evaluation of matrix roots of a matrix "characteristic equation", and the use of those matrix roots to generate "geometric" solutions to the equilibrium equations. The chief problem is to determine weak sufficient conditions for a QBD-process to be amenable to this procedure.

This dissertation rigorously develops the procedure, primarily through the use of modern analysis, matrix theory, and generalized difference equations. In the process, the notion of the "boundary leading" QBD-process is introduced. The boundary leading QBD-process is a very general QBD-process which includes all those which have a finite number of essential classes. In particular it includes all irreducible QBD-processes. It is shown that if a QBD-process is boundary leading then a suitable matrix root of the characteristic equation exists, as well as a solution of the equilibrium equations having the geometric form. This result is used to show that, for irreducible processes, the procedure can be applied, will determine an equilibrium distribution whenever one exists, and will provide an appropriate indication of failure when no equilibrium distribution exists. Other useful results regarding marginal equilibrium distributions, uniqueness of equilibrium distributions and roots of characteristic equations,

and positive recurrence in QBD-processes are also developed.

From this theory, one can conclude that the numerical solution of a very broad class of QBD-processes is feasible, and is well within the capability of a moderately large digital computer. Thus, these processes become available as models for the practice of computer system design and analysis and for other disciplines which use Markov chain models.

Acknowledgments

I wish to express my appreciation to Professors Keki B. Irani, Kuei Chuang, Ralph L. Disney, Eugene L. Lawler, and Arch W. Naylor for their service on my doctoral committee, and for their advice and counsel. I particularly thank Professor Irani for his sustained interest and encouragement as chairman of the committee. For her expert job of typing the final manuscript, I also express my appreciation to Miss Sharon Bauerle.

This work was supported by the Systems Engineering Laboratory through Air Force contract AF 30(602)-3953, granted by the Rome Air Development Center. This support was sincerely appreciated.

Finally, I thank my wife, Mary, for the patience and understanding which made this possible.

Victor Lew Wallace

Ann Arbor, Michigan
March 27, 1969

Table of Contents

		Page
List of Tables		vi
List of Figures		vii
Section		
1	Introduction	1
2	The QBD Process	11
3	Q Matrices and Notational Conventions	20
4	Existence of a Quadratic Root	32
5	Boundary Leading QBD Processes	44
6	Determination of R*	65
7	Equilibrium Distributions	72
8	Recurrent Processes	82
9	Uniqueness of S*	92
10	Conclusions	94
References		97

List of Tables

Table		Page
1.1	States of Figure 1.1, when m=2, and m'=1	6
1.2	The Transition Intensity Matrix of the Model of Figure 1.1	7

List of Figures

Figure		Page
1.1	Illustrating a Model of a Multiple Access System	3
2.1	Illustrating the States of $N = < N_1, N_2 >$	12
5.1	Graphs of Boundary Leading QBD Processes	45
5.2	Illustrating Example 5.1	49
5.3	Illustrating Example 5.2	50
5.4	Illustrating Example 5.3	51
5.5	Illustrating Example 5.4	52
5.6	Graphs of U_3 for Example of Figure 5.1	54
5.7	Graphs of U_k for a Particular N	59
7.1	A Graph of a Process for which $\rho(R) > 1$ and π is Convergent	74
8.1	A Reducible QBD Process and One of its Subgraphs	87

1. Introduction

Markovian models have found such considerable application [8, 9, 14, 15, 24] to the art of computer system design for two significant reasons. First, they have been applied because there is a sizable theory upon which to draw, and many powerful theoretical results which are available. Second, computer systems have come to a stage of evolution where they are predominantly "high-traffic" systems and, as a consequence, the modeling and estimation of statistical performance has become a crucial aspect of their design.

However, there are limitations to the models of this type which are capable of analysis. The models which must be solved are often too complex for classical queueing theory, and often have too many states for modern numerical techniques [23]. The precision required over a wide range of parameter variations often rules out simulation methods as too costly.

The study of multiple-access computer systems and the so-called "computer utilities" is particularly difficult in this regard. Furthermore, as these systems grow larger and more complex (through provision of more user circuits, introduction of multiprocessing, use of satellite computers and data-concentrators, etc.) this situation can be expected to become worse. Complexity of models, and largeness of state spaces will prevail. Yet these systems are the very systems for which analysis of Markov chain models can be

of greatest use to the system architect. Thus, new techniques are urgently needed.

A typical example of such a model for a multiple-access computer system will serve to identify some of the problems. Suppose one were considering a multiple-access computer system with two thousand common-carrier trunk-lines leading to a bank of fast processors operating on a large multiprogrammed, paged main core memory. Suppose also that the memory is demand-paged with a large bulk-core "virtual" memory. Finally, suppose our purpose is to study the relationship between the expected number of queued unserviced requests for computation and the size of main core memory, the degree of multiprogramming (i. e. the number of service requests which have some main core memory committed to them), or the number of processors committed.

A queueing model which could be used for this study is illustrated diagrammatically in Figure 1.1. This model anticipates that congestion in the channel to the bulk core will be a major problem [24]. (If it is not, of course, other models, similar in nature, would be examined.) It is assumed that connected users generate requests which "arrive" at a mean rate λ per second, at times approximately governed by a Poisson process. Each request is queued in a "wait" queue (first in, first out), until it can be serviced. A maximum of m requests is permitted to be assigned main core memory at one time, and so requests are taken from the wait queue only if the

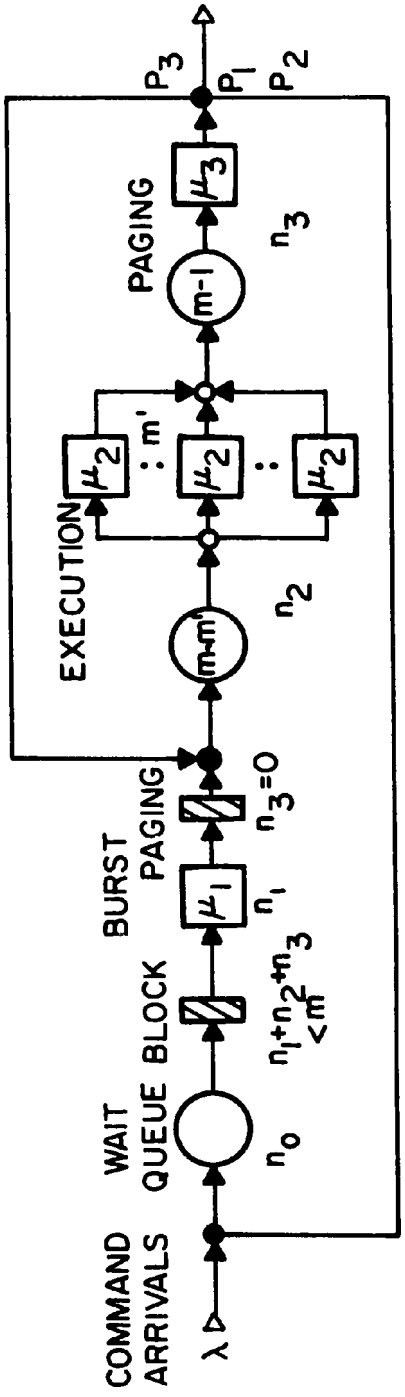


Figure 1.1

Illustrating a Model of a Multiple Access System

number so assigned is less than m (i. e. the "block" in the diagram is open). Once a request is given main core it will quickly demand that a number of pages be swapped ("burst paging") before serious execution can begin, and the bulk-core channel time for this to be accomplished has a mean $1/\mu_1$ seconds. If all m' processors are busy, the request queues for processing after the burst. (This queue can assume at most the length m-m'.) Execution proceeds for $1/\mu_2$ seconds (mean) on any one of the processors, until another page transfer is demanded. Again there is a queue for page transfers. The latter require $1/\mu_3$ seconds (mean) of bulk core transfer time. If both paging and burst-paging demands exist simultaneously, the former is given preemptive priority (which is represented on the diagram by a blocked completion of the "burst paging"). The paging may have been the result of final output of a completed command, an executive-induced interruption of the process which suspends eligibility for main core, or (most likely) a simple page demand to be followed by more execution. We assume each of these occurs with probability p_1, p_2, and p_3, respectively, and result in exit, return to the wait queue, or return to the execution queue, respectively. This model will be a continuous-parameter Markov chain provided only that the time intervals described are independent, conditionally on state, and have one of the "nice" distributions: exponential, Erlang, hyperexponential, etc. For most purposes, the exponential distribution will be simplest and most expedient.

To make things as simple as possible for this illustration, we let the degree of multiprogramming m be equal to 2, the number of processors m' to be equal to 1, and we scale time to the mean page swap-time, so that μ_3=1. The states will consist of quadruples $< n_0, n_1, n_2, n_3 >$, where n_0 is the number of commands in the wait queue, n_1 is the number in the burst phase (0 or 1), n_2 is the number in or waiting for execution, and n_3 is the number in or waiting for paging. Those states which are possible are listed and consecutively numbered in Table 1.1. Since n_0 can assume all nonnegative integer values, there are an infinite number of states, and states 10 to 15 repeat with n_0=2, 3, ... The transition intensity matrix for this model is represented in Table 1.2. It repeats the pattern within the dotted lines indefinitely. (The symbol x represents the diagonal entry, which for convenience is not evaluated. The dotted lines have been added to show the grouping of states for which n_0=0, n_0=1, n_0=2, etc.) Note that this model offers small hope of successful closed-form analytical solution. Furthermore, truncation to a finite matrix to approximate the process by one which could be solved numerically with standard methods is not possible, since n_0 can be expected to have a mean value of several hundred. For truncation to be a good approximation the range of n_0 would need to be much larger. Even for the small m, m' used for illustration, the number of states would quickly become prohibitive for numerical work.

Table 1.1

States of Figure 1.1, when m=2, and m'=1

	n_0	n_1	n_2	n_3
1.	0	0	0	2
2.	0	0	1	1
3.	0	0	2	0
4.	0	1	0	0
5.	0	1	0	1
6.	0	1	1	0
7.	0	0	0	0
8.	0	0	0	1
9.	0	0	1	0
10.	1	0	0	2
11.	1	0	2	0
12.	1	0	2	0
13.	1	1	0	0
14.	1	1	0	1
15.	1	1	1	0
16.	2	0	0	2
17.	2	0	1	1
18.	2	0	2	0
19.	2	1	0	0
20.	2	1	0	1
21.	2	1	1	0
22.	etc.			
⋮				

Table 1.2

The Transition Intensity Matrix of the Model of Figure 1.1

$$
\begin{bmatrix}
x & p_3 & 0 & 0 & p_2 & 0 & 0 & p_1 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
\mu_2 & x & p_3 & 0 & 0 & p_2 & 0 & 0 & p_1 & 0 & \lambda & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & x & 0 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & \lambda & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & p_1 & x & p_3 & 0 & 0 & 0 & 0 & 0 & 0 & p_2 & \lambda & 0 & & & & & & & & & \\
0 & 0 & \mu_1 & 0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda & & & & & & & & & \\
0 & 0 & 0 & \lambda & 0 & 0 & x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & p_2 & \lambda & 0 & p_1 & x & p_3 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & \lambda & 0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & 0 & p_1 & 0 & 0 & 0 & 0 & x & p_3 & 0 & 0 & p_2 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & p_1 & 0 & 0 & 0 & \mu_2 & x & p_3 & 0 & 0 & p_2 & 0 & \lambda & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & 0 & 0 & 0 & x & 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_1 & x & p_3 & 0 & 0 & 0 & p_2 & \lambda & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 & 0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 & \lambda & & & \\
 & & & & & & & & & 0 & 0 & 0 & 0 & p_1 & 0 & x & p_3 & 0 & 0 & p_2 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & p_1 & \mu_2 & x & p_3 & 0 & 0 & p_2 & 0 & \lambda & 0 & 0 & 0 \\
 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\
 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & x & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \\
 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_1 & x & p_3 & 0 & 0 & 0 & p_2 & \lambda \\
 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 & 0 & \mu_2 & x & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & p_1 & 0 & x & p_3 & 0 & 0 & p_2 & 0 \\
 & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & p_1 & \mu_2 & x & p_3 & 0 & 0 & p_2 \\
 & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_2 & x & 0 & 0 & 0 \\
 & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 & x & 0 & 0 \\
 & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_1 & x & p_3 \\
 & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 & 0 & \mu_2 & x \\
 & & & & & & & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & p_1 & 0 \\
 & & & & & & & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & p_1 \\
 & & & & & & & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & \mu_1 \\
 & & & & & & & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda \\
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda \\
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda
\end{bmatrix}
\quad
\begin{bmatrix}
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda \\
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda \\
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda
\end{bmatrix}
\quad
\begin{bmatrix}
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda \\
\lambda & 0 & 0 & 0 & 0 & 0 \\
0 & \lambda & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 \\
0 & 0 & 0 & p_2 & \lambda & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda
\end{bmatrix}
$$

etc.

The need for large state spaces demonstrated in this example is a natural consequence of treating multiple-access systems in which the number of users simultaneously active is great. While the earliest multiple-access systems had about ten active users, it will not be surprising to see, in the near future, systems with several thousand users simultaneously connected and generating requests at a rapid rate. This in turn implies that queues of waiting requests will be correspondingly long while performance is maintained at the same levels. For example, if one assumes an average system response time of one second, and an average request rate of one request every ten seconds for every user, then a system which has two thousand simultaneous users must be able to handle at least two hundred requests every second. It will also have an average of two hundred user requests awaiting service at any given time. Thus, it is often more reasonable, under such circumstances, to use a model which allows the number of waiting requests to assume any nonnegative integer value, as was done in the above example.

Fortunately, it turns out that these problems frequently exhibit useful structural characteristics in common. In particular, the pattern of repetition in the matrix of Table 2. 2 should be particularly noted. The dotted blocks repeat diagonally, and the form of the matrix is block-wise tri-diagonal. It should also be observed that although the size of the blocks would increase considerably for larger values of m or m', the pattern described would be preserved. This

pattern is, in fact, common to a large quantity of useful models for multiple -access computing systems (and other applications). It is a characteristics feature of the infinite state Markov chains which we call quasi birth and death processes (QBD processes) and which are the primary subject of this dissertation. Disney [3] observed the occurrence of this pattern in some finite Markov queueing problems, and offered a method for their solution which appears to be most effective when the number of repetitions is not too large. The infinite (or QBD) process was first identified and treated by Evans [4], and his work represents the genesis of our approach.

The chief purpose of this dissertation is to increase the arsenal of Markovian models available to system designers for the treatment of computer system (and other) design problems. A numerical technique has been proposed [4] for the solution of equilibrium distributions for QBD processes. These processes represent good models for many processes where one state variable assumes a very large number of possible values, as do computer utility problems when the "number of tasks awaiting completion of service" is a state variable. This variable can change by at most unity for any single state transition, and the probability intensity of each possible transition in state does not depend upon the number of tasks awaiting completion of service unless there is less than a certain number of them. Using structure imposed by these properties it will be seen that the equilibrium distributions of such a process can be solved by a procedure which

involves several separate, finite operations, each of which does not tax the capabilities of a moderately large digital computer.

In this dissertation, the procedure will be rigorously derived, and sufficient conditions will be established for its application to a given QBD process to be valid. To this end, the notion of a "boundary leading" QBD process is introduced and shown to have very useful characteristics. (Almost every interesting QBD process is boundary leading.)

2. The QBD-Process

Let m_o and m be positive integers, with

$$m_o \geq m. \tag{2.1}$$

Let $N = \{N(t), t \geq 0\}$ be a continuous parameter Markov chain with stationary transition probability functions continuous at $t = 0$, and with state space $\mathcal{N}$, where

$$\mathcal{N} = \{<0,1>, <0,2>, \ldots, <0,m_o>, <1,1>, <1,2>, \ldots, <m,1>, <2,1>, \ldots, <2,m>, \ldots, \ldots\}. \tag{2.2}$$

The process N is called a QBD process if its transition intensities $\{u_{jk}; j,k \in \mathcal{N}\}$, satisfy the properties:

(a) If $j = <j_1, j_2>$ and $k = <k_1, k_2>$ are two states for which $|j_1 - k_1| \geq 2$, then $u_{jk} = 0$.

(b) If $j = <j_1, j_2>$ and $k = <k_1, k_2>$ are two states for which $|j_1 - k_1| \leq 1$, if $(j_1 + k_1) \geq 3$, and if $j' = <j_1 - 1, j_2>$ and $k' = <k_1 - 1, k_2>$, then $u_{jk} = u_{j'k'}$.

A continuous-parameter Markov chain will also be called a QBD process if a mere one-one mapping of its states results in a QBD process. However, no loss in generality will result from consideration of only state spaces of the form (2.2), represented by the set $\mathcal{N}$.

The state space $\mathcal{N}$ is more graphically described as a two-dimensional integer space whose first dimension is countably infinite, whose second dimension is finite, and which has the configuration of

Figure 2.1. States in the set $\mathcal{N}_o = \{<0, 1>, <0, 2>, \ldots, <0, m_o>\}$ will be called boundary states and the set $\mathcal{N}_o$ will be called the boundary of N.

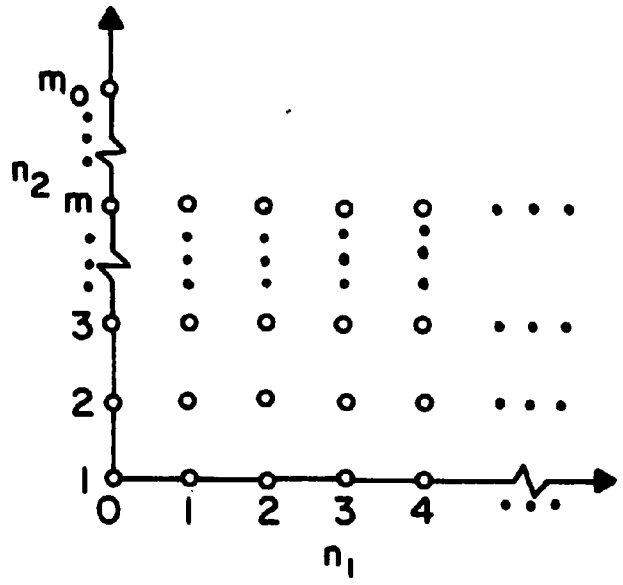


Figure 2.1 Illustrating the States of $N = < N_1, N_2 >$

If the lexicographic ordering of the states used in equation (2.2) is preserved, then the conditions (a) and (b) are equivalent to a specification that the transition intensity matrix $U = \{u_{jk};\ j, k \in \mathcal{N}\}$ is an infinite matrix which can nevertheless be displayed in the partitioned form

$$U = \begin{bmatrix} E & \hat{A} & & & \\ \hat{C} & B & A & & 0 \\ & C & B & A & \\ & & \ddots & \ddots & \ddots \\ 0 & & & & \end{bmatrix} \qquad (2.3)$$

where the submatrix E is an $m_0 \times m_0$ matrix; where A, B, and C are $m \times m$ matrices; and where $\hat{A}$ and $\hat{C}$ are $m_0 \times m$ and $m \times m_0$ matrices, respectively, and are further partitioned into the forms

$$\hat{A} = \begin{bmatrix} A \\ 0 \end{bmatrix} \tag{2.4}$$

and

$$\hat{C} = [C \;\; 0] \tag{2.5}$$

Many QBD processes whose transition intensity matrices do not have the appearance of Equation (2.3) can nevertheless be identified as QBD processes by simply regrouping the states and reidentifying them. Two examples are suggestive of the power of the model.

Example 2.1: The transition intensity matrix

$$\begin{bmatrix} E' & F' & & & & \\ G' & B & A & & & \\ & C & B & A & & \\ & & C & B & A & \\ & & \cdot & \cdot & & \cdot \\ & & & \cdot & \cdot & \cdot \end{bmatrix}$$

where E is an $m_0' \times m_0'$ matrix, where A, B, and C are $m \times m$, where F' and G' are $m_0' \times m$ and $m \times m_0'$ respectively, and where m_0' is not necessarily greater than m, can be converted to the form of Equation (2.3) by reidentifying states so that

$$E = \begin{bmatrix} B & G' \\ F' & E' \end{bmatrix}$$

Then E is $m_o \times m_o$, where $m_o = m_o' + m > m$.

Example 2.2: The transition intensity matrix

$$\begin{bmatrix} E' & F' & & & \\ G' & H' & A' & & \\ D' & C' & B' & A' & \\ & D' & C' & B' & A' \\ & \cdot & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \end{bmatrix}$$

can be converted to the form of Equation (2.3) by reidentifying its states so that

$$E = \begin{bmatrix} E' & F' \\ G' & H' \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 0 \\ A' & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} B' & A' \\ C' & B' \end{bmatrix}$$

and

$$C = \begin{bmatrix} D' & C' \\ 0 & D' \end{bmatrix}$$

provided the degrees of E' and H' together are greater than twice the degree of B'. (If not, the result is the given form in example 2.1, and can be further altered.)

Of course, the transition matrices of QBD processes need not be so easily converted to the standard form. Practically speaking, though, the application will usually suggest the appropriate state identification whereby every possible jump in state is at most a unit step in the first coordinate.

When $m_0 = m = 1$, the QBD process is readily recognized to be a simple birth and death process. The "QBD" in the name "QBD process" is an acronym for the somewhat barbaric, though apt descriptor "quasi birth and death", in recognition of this fact. (The prefix "quasi" is frequently used in matrix theory in describing matrix forms created by replacing scalar elements by submatrices, as in "quasi diagonal" and "quasi triangular" matrices.)

In order to calculate an equilibrium distribution of a QBD process N, reasoning reminiscent of that commonly followed in solving birth and death processes can be followed. This reasoning and the calculation technique that results is essentially the same, in outline, as that proposed by Evans [4, 5, 6].

An equilibrium distribution is a sequence $\pi = \{\pi_k, k \in \mathcal{N}\}$ of

nonnegative real numbers which satisfies the equation $\pi \mathcal{U} = \underline{0}$ and whose components have unit sum. If we let $\pi = [\phi_0, \phi_1, \phi_2, \ldots]$ be a partitioning of that sequence for which ϕ_0 is an m_o-component row vector, and $\phi_1, \phi_2, \ldots$, are m-component row-vectors, then the equation $\pi \mathcal{U} = \underline{0}$ is equivalent to the equations

$$\phi_o E + \phi_1 \hat{C} = \underline{0} \qquad (2.6a)$$

$$\phi_o \hat{A} + \phi_1 B + \phi_2 C = \underline{0} \qquad (2.6b)$$

$$\phi_{n-1} A + \phi_n B + \phi_{n+1} C = \underline{0} \qquad (2.6c)$$

$$n=2, 3, \ldots .$$

Now suppose that there is a solution $[\phi_0, \phi_1, \phi_2, \ldots]$ to Equations (2.6) of the form

$$\phi_1 = \phi_o \hat{R} \qquad (2.7a)$$

$$\phi_n = \phi_1 R^{n-1} \qquad n=2, 3, \ldots, \qquad (2.7b)$$

where R is an $m \times m$ matrix, where $\hat{R}$ is an $m_o \times m$ matrix partitioned to the form

$$\hat{R} = \begin{bmatrix} R \\ \underline{0} \end{bmatrix}, \qquad (2.8)$$

and where ϕ_o is some row-vector whose components are not all zero. Then it develops that, by substitution of Equations (2.7) in Equation (2.6), the matrix R and the vector ϕ_o must satisfy the equations

$$\phi_o (E + \hat{R}\hat{C}) = \underline{0} \qquad (2.9a)$$

$$\phi_o (A + \hat{R}B + \hat{R}\,RC) = \underline{0} \qquad (2.9b)$$

$$\phi_0 \hat{R} R^{n-2}(A + RB + R^2C) = \underline{0}. \qquad (2.9c)$$

Furthermore, since a solution of Equations (2.6) where $\phi_0, \phi_1, \phi_2, \ldots$, have only nonnegative components is sought, ϕ_0 and R must guarantee that nonnegativity as well. If we can show that such ϕ_0 and R exist, and yield a solution π which is an equilibrium distribution, then the assumption of Equations (2.7) will be vindicated. The Equations (2.9) suggest a sufficient condition for such a solution to be found. This condition is that a matrix R exists having the properties:

(a) All entries of R are nonnegative

(b) $E + \hat{R}\hat{C}$ is singular, and has a nonnegative, nonzero vector ϕ_0 in its null-space (so that Equation (2.9a) is satisfied).

(c) All entries of $A + RB + R^2C$ are zero.

(This latter property is reminiscent of the classical characteristic roots of a difference equation, while the first two are reminiscent of the application boundary conditions to difference equation solutions.)

It will be our purpose to show sufficient conditions that a matrix R does exist having the properties (a), (b) and (c). Remarkably, those conditions are very general, and include almost all interesting QBD processes. Under those conditions, an equilibrium distribution can be calculated by the four steps

(1) Determine a matrix R having properties (a), (b) and (c).

(2) Determine a vector ϕ_0.

(3) Normalize ϕ_0, if possible, so that π will have unit sum.

(4) Determine as many of the ϕ_n as desired using Equations (2.7a) and (2.7b).

It will be shown also that such a matrix R can be determined by an iterative process in every interesting case, and that ϕ_0 can also be calculated by known techniques.

Thus, the problem of determining an equilibrium distribution of a QBD process (which has an infinite state space) will be shown to reduce to several routine finite problems, each of which does not tax the memory or computing capabilities of a digital computer of moderate-to-large size.

This dissertation represents a significant extension of the very original work of Evans. We extend his work in several ways. In the first place, he considered only QBD processes for which A was a scalar multiple of the identity matrix ($A = \lambda I$). Secondly, he required that the process N be "irreducible with minimum path of order 2m". Thirdly, he considered only positive recurrent processes. In what follows, the matrix A is arbitrary (consistent, however, with U being a transition intensity matrix), the QBD process N is assumed only to be "boundary leading" (an assumption weaker than "irreducible"), and existence of a matrix R having the properties (a), (b), (c) is proved without any reference to "recurrence". (We do, however, still require positive recurrence, or something very close to it, for finding R.)

In addition, results are given here providing (1) necessary and sufficient conditions on R for an irreducible QBD process to be positive recurrent, (2) sufficient conditions for R and π to be unique, with π equal to the limiting distribution.

These relaxations of Evans' assumptions considerably increase the depth of the results. The presentation, relying as it does almost exclusively on algebra and modern analysis rather than probability theory, also offers a sharper, more rigorous definition for computational purposes.

3. Q-matrices and Notational Conventions

The proofs of this work are primarily algebraic, and are concerned in large measure with a type of matrix known as a Q-matrix. In this chapter, a convenient notation for some frequently used matrix relations will be established, the term "Q-matrix" will be defined, and several of its more important properties described. The machinery thus provided is necessary to the work that follows.

For any two matrices $X = \{x_{jk},\ j \in J,\ k \in K\}$ and $Y = \{y_{jk},\ j \in J,\ k \in K\}$ of equal dimensions (perhas infinite) the relations $X = Y$, $X > Y$, $X \geq Y$ will be defined to signify

$X = Y$: $x_{jk} = y_{jk}$ for every $j \in J$ and $k \in K$,

$X > Y$: $x_{jk} > y_{jk}$ for every $j \in J$ and $k \in K$,

$X \geq Y$: $x_{jk} \geq y_{jk}$ for every $j \in J$ and $k \in K$.

The special relation $X \geq Y$ will signify that the relation $X \geq Y$ holds, but that the relation $X = Y$ does not. In other words, when $X \geq Y$, every element x_{jk} is greater than or equal to y_{jk}, but at least one element exhibits inequality.

The symbol $\underline{0}$ will be used for a matrix whose elements are all zero. Similarly, the symbols $\underline{1}$ and $\underline{0}$ will be used for row or column vectors whose components are all 1 or 0, respectively. The square matrix I, of course, is the identity matrix. Where it will not be ambiguous, the context will determine dimensions, and whether row- or column-vectors are intended.

A matrix or vector X will be called nonnegative (non-positive) if every element is nonnegative (non-positive). In other words X is called nonnegative if $X \geq \underset{\sim}{0}$. A matrix will be called positive (negative) if $X > \underset{\sim}{0}$ ($X < \underset{\sim}{0}$). The row-sums of a matrix X are the components of the vector $X\underline{1}$, and are found by adding all the elements in each row of the matrix. A square matrix $X = \{x_{jk};\ j,k \in K\}$ is said to be stochastic if it is nonnegative and its row-sums are all unity (i.e. $X\underline{1} = \underline{1}$).

A square matrix $Q = \{q_{jk};\ j,k \in K\}$ is called a Q-matrix when its non-diagonal elements are not negative, and its row sums are not positive. In other words, the matrix Q is a Q-matrix when $q_{jk} \geq 0$ whenever $j \neq k$, and $Q\underline{1} \leq \underline{0}$. The transition intensity matrix of any continuous parameter Markov chain is always a Q-matrix, as are the matrices U, B, and E of the QBD process N. By the same token, the matrices A and C are nonnegative.

A matrix $X = \{x_{jk};\ j,k \in K\}$ is said to be reducible if its index set K can be divided into two disjoint nonempty subsets K_1 and K_2 such that $x_{jk} = 0$ for all $j \in K_1$ and $k \in K_2$. The matrix X is said to be irreducible otherwise. We say, further, that index j leads to index k (under X) whenever the indices K cannot be divided into two disjoing nonempty subsets K_1, K_2 containing j and k, respectively, and such that $x_{j'k'} = 0$ for all $j' \in K_1$ and $k' \in K_2$. Further, we say that j leads directly to k (under X) if $q_{jk} \neq 0$. Two indices j, k are said to communicate, on the other hand, if j leads to k and k leads to j. By

fiat, an index of a matrix will be said to communicate with itself. Therefore, the relation "i communicates with j under X" is symmetric, transitive, and reflexive, and partitions the indices into disjoint subsets called communicating index classes (or, simply, index classes) such that two indices belong in the same class if and only if they communicate. An index which communicates with every index it leads to is called essential; otherwise inessential. An index class which contains an essential index must contain only essential indices, and therefore the class is called an essential index class. (Some authors [17] prefer the term "minimal closed".) An irreducible matrix must have but a single index class, and that class must be essential.

An index class K of a Q-matrix Q is called a conservative index class if

$$\sum_{k \in K} q_{jk} = 0$$

for every $j \in K$. The matrix Q is called a conservative Q-matrix if every index class of Q is conservative, or in other words if $Q\underline{1} = \underline{0}$. For every finite conservative Q-matrix Q, the equation

$$\underline{x}\, Q = \underline{0}.$$

is true for at least some row-vector $x \geq \underline{0}$.

It should be noted that a continuous parameter Markov chain is irreducible (reducible) if its transition intensity matrix is irreducible (reducible). Furthermore, the indices of the transition

intensity matrix are "states" of the chain, and thus all terms for indices of a Q-matrix apply also to states of a chain. Note, however, that this terminology has been developed so that it can be used on Q-matrices regardless of whether or not they can be interpreted as transition intensity matrices.

The remainder of this chapter is concerned only with finite matrices.

The singularity or non-singularity of a finite Q-matrix is, of course, an important property. Since every row of a conservative Q-matrix has zero sum, its columns must be linearly dependent vectors, and Q clearly must be singular. A much more useful result is the following theorem:

Theorem 3.1

A Q-matrix is nonsingular if and only if it has no essential index class which is conservative.

The proof of this theorem will make use of two lemmas, the first of which is very minor.

Lemma 3.1

Let M be a reducible matrix, partitioned in the form

$$M = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix} \qquad (3.1)$$

where M_{11} and M_{22} are square submatrices. Then M is nonsingular iff M_{11} and M_{22} are both nonsingular.

Proof: This lemma is only a minor alteration of the known result [11, page 431] that the determinant of a quasi triangular matrix (as Equation (3.1)) is equal to the product of the determinants of its diagonal submatrices. Obviously M has nonzero determinant only if both M_{11} and M_{22} do also, and conversely if M_{11} and M_{22} have nonzero determinant then M does also. Thus, the lemma is true.

The second lemma is due to Collatz [2] and Taussky [20], among others. It will not be proved here.

Lemma 3.2

An irreducible nonconservative Q-matrix is nonsingular, and its inverse is (strictly) negative.

With that we can proceed to the proof of Theorem 3.1.

Proof (Theorem 3.1): Suppose that a Q-matrix Q has an essential index class K_1 which is conservative. If K_1 contains all indices of Q, then Q is a conservative matrix, hence singular. If, on the other hand, K_1 does not contain all indices of Q, then we can permute its rows and columns so that it can be partitioned to the form

$$Q = \begin{bmatrix} Q_{11} & 0 \\ Q_{21} & Q_{22} \end{bmatrix} \qquad (3.2)$$

where the row sums of Q_{11} are all 0. Q_{11} is clearly a conservative Q-matrix, and singular. By Lemma 3.1, then, Q is singular. This proves (modus ponens) that Q is nonsingular only if it has no

essential index class which is conservative.

To prove the converse, let Q be $n \times n$ and have no essential index class which is conservative. We proceed by induction on the number of index classes. Thus, if it has only one index class then Q is irreducible and the index class must be essential. Therefore, the matrix must be nonconservative and,by Lemma 3.2, nonsingular. Now, suppose Q has i+1 ($1 \leq i \leq n-1$) index classes, and it is true that a Q-matrix is nonsingular if it has i index classes, none of which is both essential and conservative. At least one class must be essential, and therefore nonconservative. Then the rows and columns of Q can again be permuted so that it can be partitioned to the form of Equation (3.2) where Q_{11} is irreducible and nonconservative (its indices corresponding to an essential index class of Q), and where $Q_{21} \geq \underset{\sim}{0}$. Every index class of Q_{22} must correspond to an index class of Q, and therefore Q_{22} must have i index classes. It can have no essential index class which is conservative, for if it did the corresponding rows of Q_{21} would necessarily be zero, and the corresponding indices of Q would be both essential and conservative, contradicting the main hypothesis that Q has no such classes. Consequently, Q_{22} is nonsingular. Since Q_{11} and Q_{22} are nonsingular, by Lemma 3.1 Q is nonsingular. Therefore, by induction, if the Q-matrix Q has <u>any</u> number of index classes and has no essential index class which is conservative, it is nonsingular. This completes the proof of Theorem 3.1.

We will have a need for more detailed properties of the inverse of a Q-matrix which do not appear to be common currency, mostly because they do not assume irreducibility of the matrix. But first, we quote a well-known lemma which is the basis for the "method of partitioning" (or "escalator method") for the numerical inversion of matrices. We refer to [13, p. 78] for proof.

Lemma 3.3

Let M be a partitioned nonsingular matrix

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \tag{3.3}$$

Let M_{11} and M_{22} be nonsingular, and let the matrix $W = M^{-1}$ be partitioned with the same partitioning as M,

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \tag{3.4}$$

Then W_{11} and W_{22} are nonsingular and

$$W_{11} = (M_{11} - M_{12}M_{22}^{-1}M_{21})^{-1} \tag{3.5}$$

$$W_{12} = -M_{11}^{-1}M_{12}W_{22} \tag{3.6}$$

$$W_{21} = -M_{22}^{-1}M_{21}W_{11} \tag{3.7}$$

$$W_{22} = (M_{22} - M_{21}M_{11}^{-1}M_{12})^{-1}. \tag{3.8}$$

The next two lemmas follow almost directly.

Lemma 3.4

Let Q be a nonsingular Q-matrix. Then the inverse of Q is non-positive

$$Q^{-1} \leq 0 \tag{3.9}$$

and its diagonal elements are strictly negative.

Proof: This proof also follows an induction on the number of index classes in Q. Because of its similarity to the proof of the converse part of Theorem 3.1, this proof will only be sketched. For one index class, the result follows from Lemma 3.2. Assuming that nonsingular Q-matrix with i index classes is non-positive with negative diagonals, we observe that a (i+1) index class nonsingular Q-matrix can always be partitioned (after permutation) as

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

where Q_{11} is irreducible, and Q_{22} has i classes. Then we can proceed to show that the matrices Q_{11}, $(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})$, and $(Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})$ all are nonconservative Q-matrices and have non-positive inverses. This can be done by observing row-sums, under the constraint that $(Q_{11} + Q_{12})$ and $(Q_{22} + Q_{21})$ must have non-positive row-sums and that at least one of those sums must be non-zero, and by observing that the number of index classes in a

Q-matrix (i.e. Q_{11} and Q_{22}, in this case) cannot be increased by adding a nonnegative matrix. Application of Equations (3.5)-(3.8) of Lemma 3.3 then demonstrates the non-positivity. The diagonal property is the result of the fact that $(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})^{-1}$, and Q_{11}^{-1} are strictly negative (because they are irreducible), and that $(Q_{22} - Q_{21}Q_{11}^{-1}Q_{12})^{-1}$ and Q_{22}^{-1} have negative diagonals (by the induction hypothesis). This completes the sketch of the induction.

Lemma 3.5

Let M be a nonsingular matrix. Then index j leads to index k under M if and only if index j leads to index k under M^{-1}.

Proof: Suppose j does not lead to k under M. Then (by definition of "leads to") M can be partitioned into the form

$$M = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix}$$

where j is in the upper partition of indices and k in the lower. By Lemma 3.3, the inverse of M can be shown to be correspondingly partitioned to

$$M^{-1} = W = \begin{bmatrix} W_{11} & 0 \\ W_{21} & W_{22} \end{bmatrix}.$$

Clearly j cannot lead to k under Q^{-1}. The converse is proven by

identical reasoning, replacing M by W and vice versa. This completes the proof (modus ponens).

With these results, we have finally:

Theorem 3.2

Let Q be a nonsingular Q-matrix. Then the entries w_{jk} of its inverse $W = Q^{-1}$ are nonzero if and only if j leads to k under Q.

Proof: By Lemma 3.5, $w_{jk} \neq 0$ implies that j leads to k under Q, so that only the converse needs to be proved. We therefore suppose that $w_{jk} = 0$ and examine Q. Since WQ = I, we can write

$$\sum_{\substack{i \\ i \neq \ell}} w_{ji} q_{i\ell} + w_{j\ell} q_{\ell\ell} = 0 \quad \text{for every } \ell \neq j. \qquad (3.10)$$

Let I_j be the set

$$I_j = \{i : w_{ji} \neq 0\}$$

and note that since $w_{jj} < 0$ (by Lemma 3.4), it follows that $j \in I_j$. Note further that $w_{ji} \leq 0$ (by Lemma 3.4) for all i and j, that $q_{\ell\ell} < 0$ (Q is nonsingular!) for every ℓ, and that $q_{i\ell} \geq 0$ for all $i \neq \ell$. It follows that the summation over $i \neq j$ in Equation (3.10) must be zero for every $\ell \notin I_j$, and that $q_{i\ell} = 0$ for every $\ell \notin I_j$ and every $i \in I_j$. (Observe that $\ell \notin I_j$ implies $\ell \neq j$.) Thus no index in I_j leads to one not in I_j. But, by hypothesis $w_{jk} = 0$ so that $k \notin I_j$, and, as observed previously, $j \in I_j$. Thus, the theorem is completed.

Putting together the equivalences from Lemma 3.5 and Theorem 3.2, we get the obvious corollary:

Corollary 1

> If a Q-matrix Q is nonsingular, then index j leads to index k under Q^{-1} if and only if j leads directly to k under Q^{-1}. (Equivalently, j leads to k under $W = Q^{-1}$ if and only if the entry w_{jk} is nonzero.)

We conclude this section with a simple theorem about sums of Q-matrices or nonnegative matrices.

Theorem 3.3

> If X is either a nonnegative matrix or a Q-matrix, and if Y is either a nonnegative matrix or a Q-matrix having the same index set as X, then every essential index class of X + Y contains at least one essential index class of X and at least one essential index class of Y.

Proof: Each index class of X and each index class of Y must be contained in an index class of X+Y, since the addition of positive, non-diagonal entries to a matrix cannot zero-out non-diagonal entries already present in the original and since, therefore, states which communicate under X or Y communicate under X+Y. Let K be an essential index class of X+Y. Then no indices of K lead to indices not in K under X (nor, separately, under Y either) for otherwise the same

would be true for X+Y. Then K must contain an essential index of X (and, separately, Y), and therefore one of the index classes of X (or Y) contained in K must be essential. This proves the theorem.

4. Existence of a Quadratic Root

A central question in this theory is whether or not a matrix R exists having the properties (a), (b), and (c) described in Chapter 2. These properties can be rewritten, using the notation of Chapter 3, as the properties:

(a) $R \geq \underset{\sim}{0}$,

(b) $(E + \hat{R}\hat{C})$ is nonsingular, and a row-vector $\pi_0 \geq \underline{0}$ exists for which

$$\pi_0(E + \hat{R}\hat{C}) = \underline{0},$$

(c) $A + RB + R^2C = \underset{\sim}{0}$.

If a matrix R having those properties does not exist, then our construction scheme cannot be applied to finding the equilibrium distribution π of a QBD process N, and the analogy between the solution of QBD processes and the solution of simple birth and death processes breaks down. In this chapter and the next we are concerned with some sufficient conditions that such a matrix exists.

Put in other terms, the question is whether there exists a root of the quadratic equation (c) which satisfies properties (a) and (b). In the scalar (birth and death process) case, with $A=\lambda$, $B = -(\lambda+\mu)$, $C=\mu$, and $E = -\lambda$, the quadratic becomes the familiar one

$$\lambda - R(\lambda+\mu) + R^2\mu = 0,$$

and has the two roots $R = \lambda/\mu$ and $R=1$. Only the root λ/μ satisfies condition (b), so that there is exactly one 1×1 matrix R satisfying all

three conditions. The question is not so simple when $m > 1$.

Unlike a scalar quadratic, a matrix quadratic equation $A + XB + X^2C = 0$ may have more than two roots. In fact, it has potentially as many roots as there are combinations of 2m objects taken m at a time, where m is the degree of the coefficient matrices. (See, for example, the treatment of matrix polynomials in [11, pp. 227-239].) This multiplicity is not of direct concern, since only a restricted subset of the roots (those satisfying condition (b)) is of interest. However, it does offer an indication that a study of the roots of $A + XB + X^2C = \underline{0}$ must proceed by means different from those which apply to scalar quadratics.

The conditions (a) and (b) will be strengthened somewhat in this section, for convenience. (Since the existence of roots satisfying stronger conditions implies their existence under the weaker conditions, no loss of significance results. Our proof will be adequately general.) Let $\mathcal{R}$ be the set of nonnegative matrices R'

$$R' \geq \underline{0} \tag{4.1}$$

for which the condition

$$R'C\underline{1} = A\underline{1} \tag{4.2}$$

is true. Then one can say:

Lemma 4.1

If $R \in \mathcal{R}$, then conditions (a) and (b) are true.

Proof: Condition (a) is true by definition. To prove condition (b) is true, we first observe that the transition intensity matrix $\mathcal{U}$ is an infinite, conservative Q-matrix. Therefore, by Equation 2.3, the following properties hold for the matrices A, B, C, and E:

$$(A + B + C)\underline{1} = \underline{0}, \tag{4.3}$$

$$(E + \hat{A})\underline{1} = \underline{0}. \tag{4.4}$$

Equation (4.4) implies that

$$-E\underline{1} = \begin{bmatrix} A\underline{1} \\ \underline{0} \end{bmatrix} \tag{4.5}$$

and Equations (2.5 and (2.8) imply that

$$\hat{R}\hat{C}\underline{1} = \begin{bmatrix} RC & \underline{0} \\ \underline{0} & \underline{0} \end{bmatrix} \underline{1} = \begin{bmatrix} RC \\ \underline{0} \end{bmatrix} \underline{1}. \tag{4.6}$$

Since R is in $\mathcal{R}$, then $RC\underline{1} = A\underline{1}$, and

$$\hat{R}\hat{C}\underline{1} = \begin{bmatrix} A\underline{1} \\ \underline{0} \end{bmatrix} = -E\underline{1}. \tag{4.7}$$

Thus

$$(E + \hat{R}\hat{C})\underline{1} = \underline{0}. \tag{4.8}$$

But E is a Q-matrix, and R and C are nonnegative. It follows, then, that $E + \hat{R}\hat{C}$ has nonnegative off-diagonal entries, and zero row sum. Thus $E + \hat{R}\hat{C}$ is an $m_o \times m_o$ conservative Q-matrix, and a nonnegative row-vector $\pi_o \geq \underline{0}$ exists for which

$$\pi_0(E + \hat{R}\hat{C}) = \underline{0}.$$

This proves the lemma.

The existence of a quadratic root $R \in \mathcal{R}$ is explored through a study of a fixed point of a certain mapping. Let $\mathcal{A}$ be the set of nonnegative $m \times m$ matrices S which satisfy the restrictions

$$S\underline{1} = A\underline{1} \tag{4.9}$$

and

$$(B + S) \text{ is nonsingular.} \tag{4.10}$$

(This set may be empty, but we will show later that in all interesting cases it is not.) The matrix B+S is clearly a Q-matrix for every $S \in \mathcal{A}$, since (1) the addition of a nonnegative matrix cannot make the off-diagonal elements of the Q-matrix B nonnegative, and (2) the row sums of B+S are non-positive by equations (4.3), (4.9) and (4.10), and by the construction

$$-C\underline{1} = (B+A)\underline{1} = (B+S)\underline{1} \tag{4.11}$$

The matrix $-(B+S)^{-1}C$ is also stochastic. First, a multiplication of Equation (4.11) by $(B+S)^{-1}$ shows

$$-(B+S)^{-1}C\underline{1} = \underline{1}.$$

Second, Lemma 3.4 shows that $(B+S)^{-1} \leq \underline{0}$. Thus $-(B+S)^{-1}C$ is nonnegative, and has unit row-sums. These results will be expressed, for future reference, as the lemma:

Lemma 4.2

If S is a nonnegative matrix $S \geq \underline{0}$ for which $S\underline{1} = A\underline{1}$, then the matrix (B+S) is a Q-matrix. If, in addition, (B+S) is nonsingular (and therefore S is in $\mathcal{A}$), then $-(B+S)^{-1}C$ is a stochastic matrix.

Now, let T be a mapping on $\mathcal{A}$ defined by

$$T(S) = -A(B+S)^{-1}C \tag{4.12}$$

Then we have a third lemma which will be found useful in what follows:

Lemma 4.3

For each $S \in \mathcal{A}$, T(S) is nonnegative, and

$T(S)\underline{1} = A\underline{1}$.

Proof: Since B+S is a nonsingular Q-matrix for all $S \in \mathcal{A}$, then by Lemma 3.4, $-(B+S)^{-1}$ is nonnegative. Therefore $T(S) = -A(B+S)^{-1}C$ is a product of nonnegative matrices, and is nonnegative. Multiplying the left and right of Equation (4.11) by $A(B+S)^{-1}$ gives

$$-A(B+S)^{-1}C\underline{1} = A\underline{1},$$

showing that $T(S)\underline{1} = A\underline{1}$, and proving the lemma.

The connection between fixed points of T and quadratic roots in $\mathcal{R}$ can now be shown. A fixed point S* of T is a matrix $S^* \in \mathcal{A}$ for which

$$S^* = T(S^*) = -A(B+S^*)^{-1}C. \tag{4.13}$$

Theorem 4.1

If T has a fixed point $S^* \in \mathcal{S}$, then $R^* = -A(B+S^*)^{-1}$ is a root of the quadratic $A + XB + X^2C = \underline{0}$ and is in $\mathcal{R}$.

Proof: Replacing X by $-A(B+S^*)^{-1}$ in the quadratic, and applying Equation (4.13) once, gives

$$\begin{aligned} A + R^*B + (R^*)^2C &= A - A(B+S^*)^{-1}B + A(B+S^*)^{-1}A(B+S^*)^{-1}C \\ &= A - A(B+S^*)^{-1}B - A(B+S^*)^{-1}S^* \\ &= A[I - (B+S^*)^{-1}(B+S^*)] \\ &= A[I - I] = \underline{0}, \end{aligned} \tag{4.14}$$

proving that R^* is a root of the quadratic. To prove that R^* is in $\mathcal{R}$, we must prove that it is nonnegative, and that $R^*C\underline{1} = A\underline{1}$. The matrix S^* is in $\mathcal{S}$. Consequently, by Lemma 4.2 and Lemma 3.4, $(B+S^*)$ is a Q-matrix and its inverse is non-positive. Thus $R^* = -A(B+S^*)^{-1}$ is nonnegative. Replacing S by S^* in Equation (4.11), multiplying the right side of Equation (4.11) by $-A(B+S^*)^{-1}$, and multiplying the left side of (4.11) by R^* demonstrates that

$$-R^*C\underline{1} = -A\underline{1} \tag{4.15}$$

and completes the proof.

That such a fixed point exists will be proved using the well-known theorem of Brouwer [18, pp. 42-45] which will not be proved here.

Theorem 4.2 (Brouwer's Fixed-Point Theorem)

Let $\mathcal{F}$ be a continuous mapping of a non-empty, convex, compact subset $\mathcal{M}'$ of a finite-dimensional normed space $\mathcal{N}$ into itself. Then $\mathcal{F}$ has a fixed point in $\mathcal{M}'$.

It will be shown that a particular subset of $\mathcal{L}$ plays the role of $\mathcal{M}'$ in this theorem, and that T plays the role of $\mathcal{F}$. When that can be done, we will have shown that a fixed point S* exists in the subset of $\mathcal{L}$, (consequently, also in $\mathcal{L}$), and that therefore a root $R^* = -A(B+S^*)^{-1}$ of the quadratic also exists.

The set $\mathcal{N}$ of m × m matrices is a normed space (cf. [Fa, p. 58]), under the norm

$$||M|| = \max_{i=1,\ldots,m} \sum_{j=1}^{m} |m_{ij}|, \qquad (4.16)$$

where $M = (m_{ij})$ is a matrix in $\mathcal{M}$. There is a class of subsets of $\mathcal{M}$ which are compact and convex. Let M be an m × m matrix. Let $\mathcal{L}_M$ be the set of m × m matrices S which satisfy the restrictions

$$S\underline{1} = A\underline{1} \qquad (4.17)$$

$$S \geqq M \geqq \underline{0} \qquad (4.18)$$

Theorem 4.3 demonstrates the compactness and convexity of $\mathcal{L}_M$.

Theorem 4.3

For every nonnegative m × m matrix M, $\mathcal{L}_M$ is a compact, convex subset of $\mathcal{M}$.

Proof: Since $\mathcal{A}_M$ is finite dimensional, in order to show compactness it is sufficient to show that it is closed and bounded. That it is convex is readily shown, since for each $S_1, S_2 \in \mathcal{A}_M$, and each $a \in [0,1]$, the matrix $S = aS_1 + (1-a)S_2$ has the properties

$$S\underline{1} = aS_1\underline{1} + (1-a)S_2\underline{1} = aA\underline{1} + (1-a)A\underline{1} = A\underline{1}$$

and

$$S = aS_1 + (1-a)S_2 \geqq aM + (1-a)M = M.$$

That $\mathcal{A}_M$ is bounded is similarly easily shown, since every member of $\mathcal{A}_M$ is nonnegative and has the same row sums, and therefore every member of $\mathcal{A}_M$ has the same norm. I.e. $S_1, S_2 \in \mathcal{A}_M$ implies $||S_1|| = ||S_2|| = ||A||$.

That $\mathcal{A}_M$ is closed is proved by showing that if $S_o = (s^o_{ij})$ is a limit point of $\mathcal{A}_M$, then it is in $\mathcal{A}_M$. First, it is noted that, for every $s \in \mathcal{A}_M$, $S\underline{1} = A\underline{1}$. Consequently, any limit point S_o of $\mathcal{A}_M$ clearly has $S_o\underline{1} = A\underline{1}$. Second, if S_o is a limit point but $S_o \ngeqq M$, then there is at least one entry of $S_o - M$ which is negative. Say,

$$s^o_{i^*j^*} - m_{i^*j^*} = -\delta < 0.$$

Then $||(S-M)-(S_o-M)|| > \delta$ for every $S \in \mathcal{A}_M$, since $(S-M)$ is nonnegative, and the absolute sum of the i^*-th row is at least δ. Consequently, $||S-S_o|| > \delta$ for every $S \in \mathcal{A}_M$, and there is some ϵ-neighborhood (e.g. $\epsilon = \delta/2$) of S_o which does not contain any member of $\mathcal{A}_M$. This contradicts that S_o is a limit point, proves that

$S_o \geq M$ and that $S_o \in \mathcal{A}_M$, and completes the proof of the theorem.

Having shown that $\mathcal{A}_M$ qualifies as the subset for the Brouwer theorem for arbitrary $M \geq Q$, the next requirement is to examine the continuity of T.

Theorem 4.4

Let M be a nonnegative m × m matrix. If $\mathcal{A}_M \subseteq \mathcal{A}$ then T is a continuous mapping on $\mathcal{A}_M$.

Proof: By the definition of $\mathcal{A}$, (B+S) is nonsingular for every $S \in \mathcal{A}_M$. Let S_o be any member of $\mathcal{A}_M$, and consider any S in $\mathcal{A}_M$ such that $||S-S_o|| < \delta_\epsilon = \epsilon / ||A(B+S_o)^{-1}||$. Then

$$\begin{aligned} T(S) - T(S_o) &= -A(B+S)^{-1}C + A(B+S_o)^{-1}C \\ &= -A(B+S_o)^{-1}[(B+S_o)^{-1}(B+S)^{-1} - I]C \\ &= -A(B+S_o)^{-1}[(B+S_o)-(B+S)](B+S)^{-1}C \\ &= -A(B+S_o)^{-1}(S_o-S)(B+S)^{-1}C \end{aligned}$$

and

$$\begin{aligned} ||T(S)-T(S_o)|| &\leq ||A(B+S_o)^{-1}||\ ||S-S_o||\ ||(B+S)^{-1}C|| \\ &< \epsilon\, ||(B+S)^{-1}C|| \end{aligned}$$

However, by Lemma 4.2, $-(B+S)^{-1}C$ is a stochastic matrix. It follows that $||(B+S)^{-1}C|| = 1$ and that

$$||T(S) - T(S_o)|| \leq ||A(B+S_o)^{-1}||\ ||S-S_o|| < \epsilon$$

which proves continuity at every $S_o \in \mathcal{A}_M$, concluding the theorem

Several statements are useful alternatives to the antecedent of Theorem 4.4.

Theorem 4.5

Let M be a nonnegative $m \times m$ matrix. Then the following are equivalent:

(a) (B+S) is nonsingular for every $S \in \mathcal{L}_M$.

(b) $\mathcal{L}_M \subseteq \mathcal{L}$.

When $M\underline{1} \leqq A\underline{1}$, these are equivalent to:

(c) B+M has no essential index class whose corresponding entries to $C\underline{1}$ are identically zero.

Proof: The equivalence of (a) and (b) follows trivially from the definitions of $\mathcal{L}$ and $\mathcal{L}_M$. The equivalence of (c) is the important result. Since M is nonnegative, and $M\underline{1} \leqq A\underline{1}$, then $(B+M)\underline{1} \leqq -C\underline{1}$ and B+M is a Q-matrix. For every $S \in \mathcal{L}_M$, $B+S \geqq B+M$, and both B+S and B+M are Q-matrices. Therefore, by Theorem 3.3, every essential index class of B+S contains at least one essential index class of B+M. Thus condition (c) implies that for every $S \in \mathcal{L}_M$, B+S has no essential index class whose corresponding entries to $C\underline{1}$ are identically zero. But $(B+S)\underline{1} = -C\underline{1}$. Consequently, for every $S \in \mathcal{L}_M$, B+S has no such index class which is conservative. By Theorem 3.1, B+S is nonsingular and thus (a) is implied by (c). Conversely, let D be a matrix formed by adding the entries of $(A\underline{1}-M\underline{1})$ to the corresponding diagonal entries of M. Then $D\underline{1} = A\underline{1}$ and $D \geqq M$, so that $D \in \mathcal{L}_M$. By theorem 3.1 and Lemma 4.2, conditions (a) and (c) imply that B+D has no conservative index class, and hence no class for which corresponding entries of $C\underline{1}$ are zero.

But the addition of a diagonal matrix does not change the communication properties of a matrix, hence B+M has the same properties as B+D. This proves that (a) and (b) imply (c), and completes the theorem.

We are now prepared to routinely prove the principal result of this section:

Theorem 4.6

If a nonnegative matrix M exists such that

(a) $\mathcal{L}_M$ is not empty,

(b) $\mathcal{L}_M \subseteq \mathcal{L}$,

(c) T maps $\mathcal{L}_M$ into $\mathcal{L}_M$,

then a root R of the quadratic $A + XB + X^2C = \underline{0}$ exists in $\mathcal{R}$.

Proof: The conditions of the Brouwer theorem (Theorem 4.2) are met for the normed space $\mathcal{M}$, the nonempty, convex, compact subset $\mathcal{L}_M$, and the continuous mapping T of $\mathcal{L}_M$ into itself. Therefore there exists a fixed point S* of T, where $S^* \in \mathcal{L}_M$. Since $\mathcal{L}_M \subseteq \mathcal{L}$, then S* is also in $\mathcal{L}$, and by Theorem 4.1 the root R exists in $\mathcal{R}$. This proves the theorem.

The condition of Theorem 4.6 is easily shown to be met by QBD-process for which B is irreducible, and for others where B has other readily identifiable properties. These classes of processes are quite useful, and are special cases of an even broader class which will be shown in Chapter 5 to also meet the conditions of

Theorem 4.6. However, since the proof of the more general case is quite complicated, it will be instructive and somewhat satisfying to prove the special case now, as a simple corollary of Theorem 4.6.

Corollary 1

If B has no essential index class whose corresponding entries to $C\underline{1}$ are all zero, then a root $R \in \mathcal{R}$ of the quadratic equation $A + XB + X^2C = \underset{\sim}{0}$ exists.

Proof: Choose $M = \underset{\sim}{0}$. The set $\mathcal{A}_0$ contains the nonnegative matrix A, and is therefore not empty. Furthermore, it is clear that $M\underline{1} \leq A\underline{1}$ and thus, by Theorem 4.5, $\mathcal{A}_0 \subseteq \mathcal{A}$. Finally, by Lemma 4.2, $T(S) \geq 0$ and $T(S)\underline{1} = A\underline{1}$ for every $S \in \mathcal{A}_0$. Consequently, T maps $\mathcal{A}_0$ into $\mathcal{A}_0$ and the antecedent of Theorem 4.6 is satisfied, proving the corollary.

A special case of QBD processes satisfying the antecedent of this corollary occurs when B is irreducible and C is not a zero matrix ($\underset{\sim}{0}$). Such problems occur frequently in applications.

5. Boundary Leading QBD Processes

The sufficient conditions for the existence of the quadratic root R which were defined by Theorem 4.6 are satisfied by a very general type of QBD process, the "boundary leading" process.

Definition

A QBD process N is boundary leading if every non-boundary state leads to a boundary state.

This condition is a very weak one, and encompasses all irreducible QBD-processes as well as a considerable variety of reducible ones.

Transition graphs of a number of simple boundary leading QBD processes are shown in Figure 5.1, demonstrating this variety. These graphs represent every state by a vertex, and every nonzero, non-diagonal transition intensity u_{jk} $(j \neq k)$ by a directed branch from state j to state k. The states are arranged as in Figure 2.1. Two way paths in Figure 5.1 are shown by heavy lines, and boundary states by square vertices. These graphs are useful in visualizing the communication properties of particular QBD-processes, since a state j "leads to" a state k if and only if there is a directed path in the graph which can be followed from j to k. Note that only the graphs (e) and (f) describe an irreducible process. The rest are reducible.

It should be particularly noted that the definition of boundary leading processes makes no mention of paths between boundary

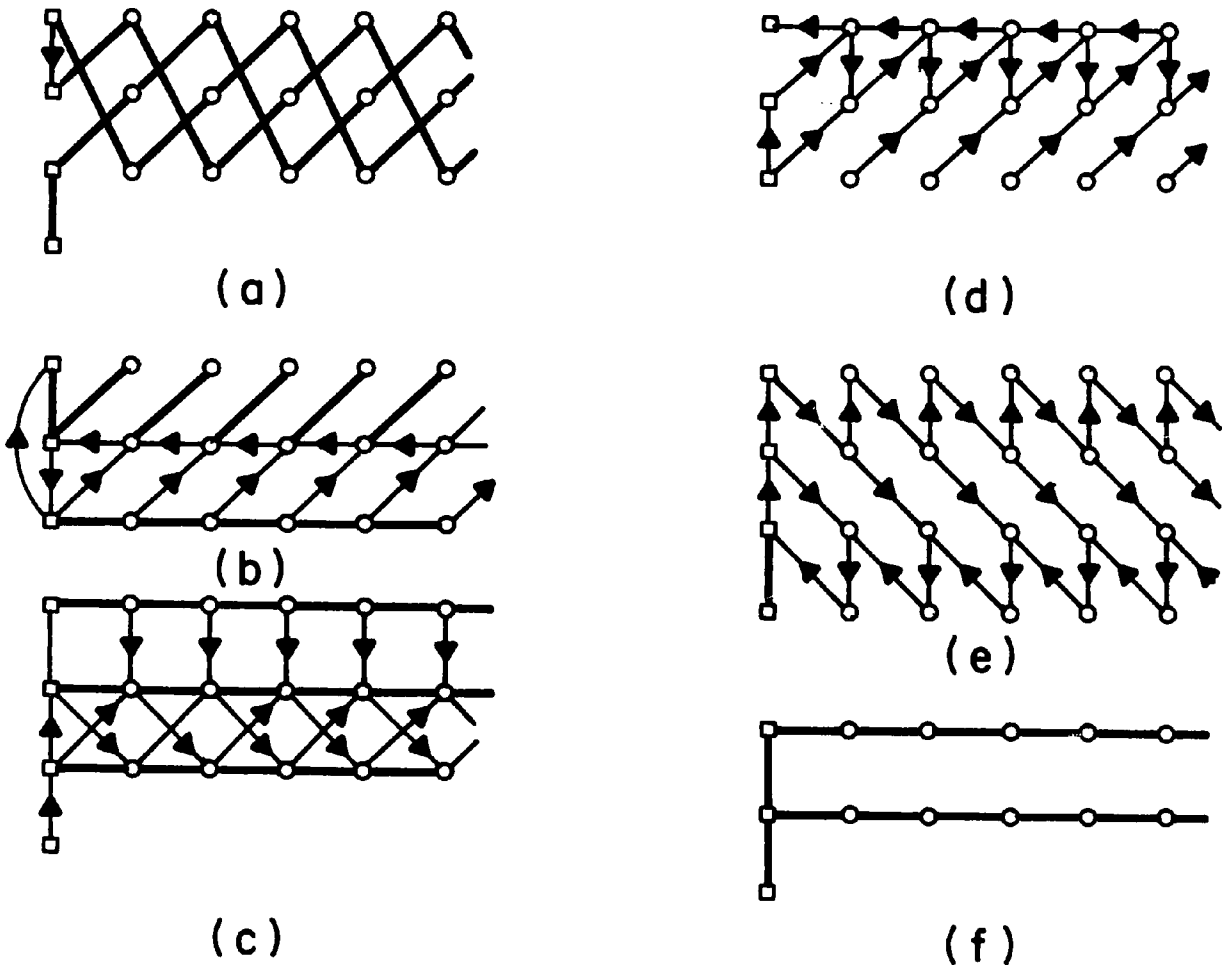


Figure 5.1

Graphs of Boundary Leading QBD-processes

states. Thus, branches between square vertices in Figure 5.1 can be removed or added at will without altering the boundary leading character of the examples. This exercise offers still other interesting examples. The most interesting aspect of all the examples, however, is the way that the horizontal repetitiveness, characteristic of QBD-processes, forces certain communication properties. The most pervasive of these properties can be stated in the obvious "lemma".

Lemma 5.1

Let j and k be two states, not both boundary states, of a QBD-process. Then $j = \langle j_1, j_2 \rangle$ leads to $k = \langle k_1, k_2 \rangle$ if and only if $\langle j_1+\ell, j_2 \rangle$ leads to $\langle k_1+\ell, k_2 \rangle$, where ℓ is a positive integer.

(We omit proof.)

The assumption that N is boundary leading is so general that it automatically includes every QBD process which has a finite number of communicating classes, as seen by the theorem:

Theorem 5.1

If a QBD process N has a finite number of communicating classes of states, then it is boundary leading.

Proof: Let $\mathcal{N}_{n_1} = \{\langle n_1, n_2 \rangle : n_2=1, 2, \ldots, m\}$ for each $n_1=1, 2, \ldots$. Let $\mathcal{N}_o = \{\langle 0, n \rangle : n_2=1, 2, \ldots, m_o\}$. Suppose N is not boundary leading. Then there is a set of nonboundary states which do not lead to a boundary state. Let $\langle \alpha_1, \alpha_2 \rangle$ be a member of that

set which has least value $\alpha_1 > 0$. It cannot lead to any state in $\mathcal{N}_{\alpha_1-1}$, since those are either boundary states or lead to boundary states. The state $<\alpha_1+1, \alpha_2>$ can not lead to any state in $\mathcal{N}_{\alpha_1}$, by Lemma 5.1. Similarly states $<\alpha_1+2, \alpha_2>$ cannot lead to states in $\mathcal{N}_{\alpha+1}$, etc. Each of these states, therefore, cannot communicate with one another, and must be in distinct communicating classes. This proves the existence of an infinite number of such classes, contradicting the antecedent and proving the theorem.

The converse is not true, however, since Figure 5.1(d) describes a boundary leading process which (degenerately) has an infinite number of communicating classes.(Note that each state in the lowest row is a communicating class by itself.) A weaker form of the converse is somewhat interesting, and easily proved.

Theorem 5.2

> If N is boundary leading, then N has a finite number of essential communicating classes. Furthermore, each such class includes at least one boundary state.

Proof: By definition, every essential state communicates with every state it leads to. Since every state of a boundary leading process leads to a boundary state, every essential state communicates with a boundary state and every essential class contains a boundary state. No state may be in more than one class, so there must be no more essential classes than there are boundary states, a finite number. This proves the theorem.

This theorem shows that when a boundary leading process has an infinite number of communicating classes, the classes must necessarily be inessential (except for a finite number). A useful corollary shows that every boundary leading QBD process has some essential states.

Corollary 1

If N is boundary leading, then N has at least one essential communicating class of states.

Proof: It is sufficient, by the theorem, to show that at least one boundary state is essential. Every boundary state is either essential, or leads to another state which does not lead to it. If the latter, it must lead to another boundary state, because if the state it leads to is a non boundary state, that state must lead to a boundary state. Since the boundary states are finite in number, all cannot lead to another without communicating. Thus one must be essential, and the corollary is proved.

The boundary leading QBD process has been singled out for study here because it represents one of the most general processes for which a quadratic root matrix R exists in $\mathcal{R}$. That the root in $\mathcal{R}$ does not exist in general is illustrated by a pair of examples. The first has an infinite number of essential communicating classes of states, and the second has an infinite number of inessential classes. (Recall that non-boundary leading processes must have an infinite number of communicating classes.)

Example 5.1: Let

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & -4 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 \\ 2 & 1 \end{bmatrix} \quad E = \begin{bmatrix} 0 & 0 \\ 1 & -2 \end{bmatrix}$$

Then the graph of N becomes that of Figure 5.2.

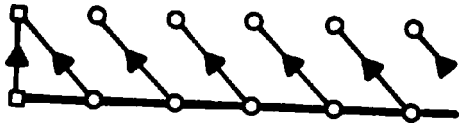


Figure 5.2

Illustrating Example 5.1

An equilibrium solution π can be expressed as

$$\pi = < \xi_0, 0, \xi_1, 0, \xi_2, 0, \ldots >$$

where $\xi_0, \xi_1, \xi_2, \ldots$ are arbitrary positive constants. A brute-force calculation of roots to $A + XB + X^2C$ (found by expanding the matrix product and solving the resulting four equations in A_{ij}) shows that every root is of the form

$$R = \begin{bmatrix} \zeta & 0 \\ -\frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

where ζ is an arbitrary constant. There is no root for which $RC1 = A1$, and therefore no root $R \in \mathcal{R}$. Furthermore, since $A\underline{1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $B+S$ is necessarily singular for every $S \in \mathcal{L}$, and $\mathcal{L}$ is therefore empty.

Example 5.2: Let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}, \quad C = \underset{\sim}{0}, \quad E = \begin{bmatrix} -2 & 0 \\ 1 & -2 \end{bmatrix}$$

so that N has the graph of Figure 5.3. In this case the quadratic

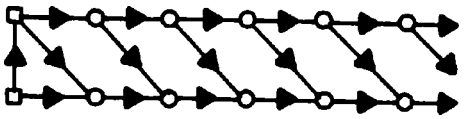


Figure 5.3

Illustrating Example 5.2

reduces to $A + XB = \underset{\sim}{0}$, so that $R = -AB^{-1}$ is the only root:

$$R = \begin{bmatrix} .5 & 1 \\ 0 & 1 \end{bmatrix}.$$

Since $C = \underset{\sim}{0}$, $RC\underset{\sim}{1}$ is obviously not equal to $A\underset{\sim}{1}$, so that there are no roots in $\mathscr{R}$. In this case, however, $\mathscr{S}$ is not empty, since it contains at least

$$S = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}.$$

These examples notwithstanding, boundary leadingness is not a necessary condition for the existence of a quadratic root in $\mathscr{R}$. This is attested by the further examples:

Example 5.3: Let

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad B = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad E = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix},$$

so that N has the graph of Figure 5.4. Using the procedure outlined

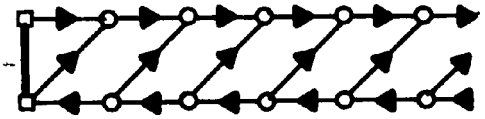


Figure 5.4

Illustrating Example 5.3

by Gantmacher [11, pp. 227-230], all of the roots of $A + XB + X^2C$ are found to be:

$$R = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}.$$

Of these, the second (and only the second) is in $\mathcal{R}$, and for that one

$$E + \hat{R}\hat{C} = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}.$$

Thus, $\pi_o = <1, 2>$, and an equilibrium solution is

$$\pi = <1, 2, 3, 3, 6, 6, 12, 12, \ldots>$$

We have also, the example

Example 5.4: Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \quad E = \begin{bmatrix} -2 & 1 \\ 0 & -1 \end{bmatrix}$$

with the graph of N given in Figure 5.5. In this case, using Gantmacher's procedure, we have the roots

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & .5 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ 0 & .5 \end{bmatrix}.$$

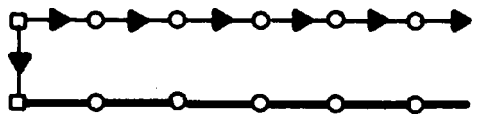


Figure 5. 5

Illustrating Example 5. 4

In this case, the last is in $\mathcal{R}$.

These latter examples appear to indicate that perhaps boundary leading should be extended to permit a "path" via infinity, so as to include these cases. However, the proof which follows uses only the definition originally given. We will show that for boundary leading QBD processes there exists a matrix V such that (a) $\mathcal{A}_V$ is not empty, (b) $\mathcal{A}_V \subseteq \mathcal{A}$, and (c) $T(\mathcal{A}_V) \subseteq \mathcal{A}_V$. This proof unfortunately requires considerable machinery for the construction of V.

Let n be a positive integer, and let U_n be the $mn \times mn$ matrix

$$U_n = \begin{bmatrix} B & A & & & \\ C & B & A & & \\ & \cdot & \cdot & \cdot & \\ & & & \cdot \cdot \cdot & \\ & & & C & B \end{bmatrix} \qquad (5.1)$$

This matrix is a Q-matrix. To construct V, we consider the inverse W(n) (assuming it exists) of U_n. Let W(n) be partitioned so that its diagonal submatrices are $m \times m$ and $m(n-1) \times m(n-1)$

$$U_n^{-1} = W(n) = \begin{bmatrix} W_1(n) & W_3(n) \\ W_2(n) & W_4(n) \end{bmatrix} \qquad (5.2)$$

We contend that the matrix $B - AW_1(n)C$ is a Q-matrix, and that for some $n=\sigma$ the nonnegative $m\times m$ matrix $B - AW_1(\sigma)C$ has no essential index class whose corresponding entries to $C\underline{1}$ are identically zero. (By Theorem 4.4 this property is sufficient to show that $B+S$ is nonsingular for all $S \in \mathcal{A}_{-AW_1(\sigma)C}$.) We let

$$V = -AW_1(\sigma)C, \qquad (5.3)$$

and contend that $\mathcal{A}_V$ has the desired properties. These contentions will be proved in a sequence of six lemmas.

But first, observe that the matrix U_n represents the transitions between the states in a subset $\mathcal{N}_{\alpha,n}$ of the state set of N, where α is a positive integer and

$$\mathcal{N}_{\alpha,n} = \{\alpha, \alpha+1, \ldots, \alpha+n-1\} \times \{1, 2, \ldots, m\}.$$

To aid visualization, the transitions of U_n can be interpreted by a graph which is a "slice" of width n taken from the middle of the graph of N. (For example, the graphs of Figure 5.6 interpret U_3 for each corresponding graph of N in Figure 5.1.) With this background, we are ready to prove the first lemma:

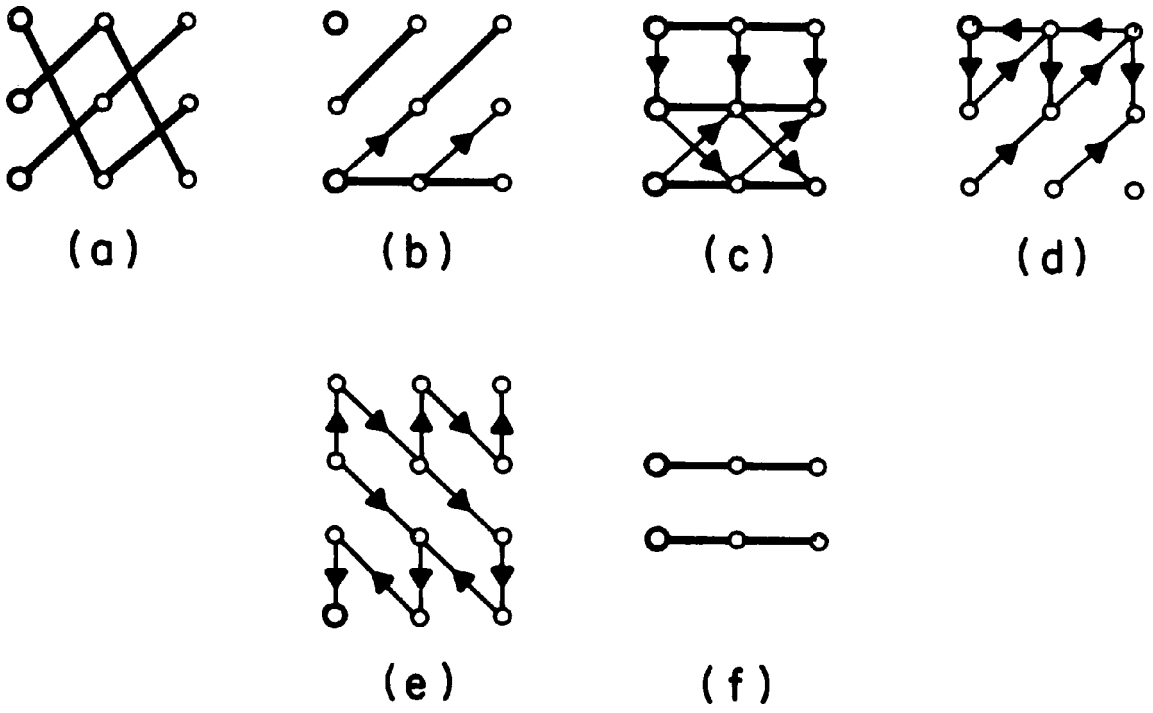


Figure 5.6

Graphs of U_3 for Examples of Figure 5.1

Lemma 5.2

If N is boundary leading, then U_n is nonsingular for all $n \geq 1$. In particular, B is nonsingular.

Proof: Suppose U_n is singular. Since it is a Q-matrix, by Theorem 3.1 it has an essential index class which is conservative. This would imply that N has an essential class of states entirely within $\mathcal{N}_{\alpha, n}$, since no transitions from states in $\mathcal{N}_{\alpha, n}$ are possible without producing row sums of U which are greater than the corresponding ones of U_n. $\mathcal{N}_{\alpha, n}$ does not contain any boundary states. Therefore there is an essential class which does not include a boundary state, contradicting Theorem 5.2. This proves the main consequent of the lemma. The remainder is proved by merely noting that $U_1 = B$. Thus, the lemma is proved.

A recursive relation for the matrices $W_1(n)$, $n=1, 2, \ldots$, can be found by making use of the formula for inverting partitioned matrices.

Lemma 5.3

If N is boundary leading, then $W_1(n)$ can be determined from the recursion

$$W_1(n+1) = (B - AW_1(n)C)^{-1},$$

where $W_1(1) = B^{-1}$ and $n=1, 2, \ldots$.

Proof: That $W_1(1) = B^{-1}$ is a trivial result of the definitions. When U_{n+1} is partitioned so that its diagonal submatrices are $m \times m$ and $mn \times mn$, the resulting matrix is written

$$U_{n+1} = \begin{bmatrix} B & \hat{A}_n \\ \hat{C}_n & U_n \end{bmatrix} \tag{5.4}$$

where $\hat{A}_n$ is the $m \times mn$ matrix

$$\hat{A}_n = [A \; \underset{\sim}{0} \ldots \underset{\sim}{0}] ; \tag{5.5}$$

and $\hat{C}_n$ is the $mn \times m$ matrix

$$\hat{C}_n = \begin{bmatrix} C \\ \underset{\sim}{0} \\ \vdots \\ \underset{\sim}{0} \end{bmatrix} . \tag{5.6}$$

But both B and U_n are nonsingular, and the matrix $W_1(n+1)$ is the upper diagonal submatrix of the inverse of U_{n+1}. Therefore, by Lemma 3.3, $W_1(n+1)$ is given by the formula

$$W_1(n+1) = (B - \hat{A}_n U_n^{-1} \hat{C}_n)^{-1} \tag{5.7}$$

for n=1, 2, ... (the non singularity of $B - \hat{A}_n U_n^{-1} \hat{C}_n$ being guaranteed by the lemma). In view of the zero elements of $\hat{A}_n$ and $\hat{C}_n$ only the $m \times m$ upper diagonal submatrix of U_n^{-1} need be evaluated, and Equation (5.7) can be simplified to

$$W_1(n+1) = (B - AW_1(n)C)^{-1} \tag{5.8}$$

(Nonsingularity is, again, guaranteed.) This proves the lemma.

The matrix U_n is a nonsingular Q-matrix. Therefore its inverse is non-positive, and so also is its submatrix $W_1(n)$. Furthermore, $B - AW_1(n)C$ is a Q-matrix, by the lemma:

Lemma 5.4

If N is boundary leading, the following inequalities hold for all n=1, 2, . . .

(a) $-AW_1(n)C \geqq \underset{\sim}{0}$

(b) $-AW_1(n)C\underline{1} \leqq A\underline{1}$,

and therefore $B - AW_1(n)C$ is a Q-matrix.

Proof: The non negativity of $-AW_1(n)C$ (inequality (a)) follows immediately from the non positivity of $W_1(n)$, and the non negativity of A and C. Recalling (Equation 4.3) that

$$(A + B + C)\underline{1} = \underline{0},$$

it follows that

$$-C\underline{1} - A\underline{1} = B\underline{1}$$

and, multiplying on the left by AB^{-1},

$$-AB^{-1}C\underline{1} - AB^{-1}A\underline{1} = A\underline{1}.$$

Since B^{-1} is non positive, $-AB^{-1}A\underline{1}$ is nonnegative, and therefore

$$-AB^{-1}C\underline{1} \leqq A\underline{1}. \tag{5.9}$$

Since $W_1(1) = B$, this proves (b) for n=1. If $-AW_1(n)C\underline{1} \leqq A1$, then

$$(B - AW_1(n)C)\underline{1} \leqq (B + A)\underline{1},$$

$$(B - AW_1(n)C)\underline{1} \leqq -C\underline{1}$$

and, multiplying on the left by $A(B - AW_1(n)C)^{-1}$, which is non-positive

$$A\underline{1} \geqq -A(B - AW_1(n)C)^{-1}C\underline{1} = -AW_1(n+1)C\underline{1}.$$

This completes the induction for (b). The matrix $B - AW_1(n)C$ is the sum of a nonnegative matrix and a Q-matrix, so that its off-diagonal elements are nonnegative. Its row sums are non positive, so it is clearly a Q-matrix. This completes the proof.

The next step is to show that for some σ the matrix $W_1(\sigma)$ has no essential index class for which the corresponding entries of the vector $C\underline{1}$ are zero. (This is needed in order to show that B+S is nonsingular for all $S \in \mathcal{L}_V$, which in turn will be used to show that $\mathcal{L}_V \subseteq \mathcal{L}$.) Our objective will be pursued by closer examination of the communication properties of U_n.

The matrix $W_1(n)$ has nonzero entries $w_{ij}(n)$ if and only if i leads to j under U_n, where i and j are members of the first m indices of U_n. This is a consequence of Theorem 3.2, Equations (5.1) and (5.2), and the fact (Lemma 5.2) that U_n is a nonsingular Q-matrix. Now, what we are interested in is whether or not all those first m indices of U_n lead to other indices whose entries in $C\underline{1}$ are not zero. To make this clearer, observe Figure 5.7, which shows graphs of U_1, U_2, U_3, and U_4 for a particular N. The index "4" is the only one of the first m (m=4) whose entry in $C\underline{1}$ is not zero, since it has a leftward leading branch leaving it. In this graph it is seen that every other index in $\{1, 2, 3, 4\}$ of U_3 leads to index 4 (see the darkened paths), but that index 1 does not in U_2, and indices 1 and 2 do not in U_1. Thus, $W_1(3)$, has nonzero entries such that every index leads to an index whose entry in $C\underline{1}$ is not zero.

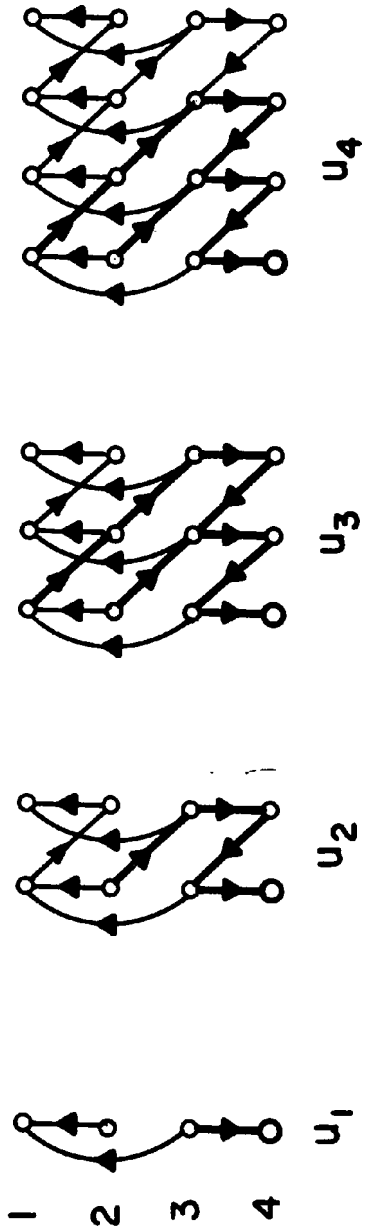


Figure 5.7

Graphs of U_k for a Particular N.

Let J_C be the set of indices (in $1, 2, \ldots, m$) which have nonzero entries to C1. (I. e., in the above example $J_C = \{4\}$.) Let I(n) be the set of indices in $1, 2, \ldots, m$ which do not lead (under U_n) to any index in J_C. Then we can generalize what we have just shown in example by the following lemma.

Lemma 5. 5

If N is boundary leading, then there exists an integer $\sigma < m$ such that $I(\sigma+1)$ is empty.

Proof: First, we note that if an index is a member of $I(n+1)$ it must have been a member of I(n), since the increase of n to n+1 merely appears to add branches to the right of the graph to those already present in U_n and can only add to the indices which lead to J_C. Symbolically $I(n+1) \subseteq I(n)$. Second, we note that if $I(n+1) = I(n)$, then $I(n+\ell) = I(n)$ for every positive ℓ, since: the increase of n to n+1 can also be regarded as an addition of a column of vectors and their branches to the left of the graph U_n, and if $I(n+1) = I(n)$, then no new paths are added. Consequently, the addition of a new column to the left of that can do no more than the last, because of the repetitive character of the graphs, etc. (Strictly we are here appealing to Lemma 5. 3). Third, we observe that if $I(n+\ell) = I(n)$ for all positive ℓ, then the indices in I(n) do not lead to an index in J_C for any value of n. This implies that the states $\{<\alpha, i>,\ i \in I(n)\}$ do not lead to any states which lead "leftward" from $\mathcal{K}_\alpha$, and hence implies that the process N is not boundary leading. Consequently, $I(n+1) \subset I(n)$ unless

$I(n) = \phi$. Finally, we observe that since I(1) could contain at most m members, and each successive set has less members, then the first set $I(\sigma+1)$ which is empty must occur before n=m, and thus $\sigma < n$. This proves the lemma.

The next lemma completes the construction of V.

Lemma 5.6

If N is boundary leading and σ is the integer referred to in Lemma 5.5, then $B - AW_1(\sigma)C$ has no essential class whose corresponding entries to $C\underline{1}$ are identically zero.

Proof: The index i leads to index j under $W_1(\sigma+1)$ if and only if i leads to j under $U_{\sigma+1}$. If it does, then, by the previous lemma, every index of $W_1(\sigma+1)$ leads to one in J_C. But by Equation (5.8)

$$W_1(\sigma+1) = (B - AW_1(\sigma)C)^{-1}.$$

Again using Theorem 3.2, the matrix $B - AW_1(\sigma)C$ also has no index which does not lead to one in J_C. Every index in an essential class must lead to one in J_C, and the class must contain at least one. If an index in an essential class leads to another index, that latter index must be in the class. Therefore every essential class contains at least one index in J_C, proving the lemma.

One last lemma is required before the main theorem. It shows that $\{W_1(n), n=1, 2, \ldots\}$, is a monotone sequence.

Lemma 5.7

If N is boundary leading, $-W_1(n+1) \geqq -W_1(n)$, for all $n=1, 2, \ldots$.

Proof: The lemma follows by induction. First, noting from Equation (5.2) that $W_1(1) = B^{-1}$, we have

$$\begin{aligned} W_1(2) - W_1(1) &= W_1(2)[(W_1(1))^{-1} - (W_1(2))^{-1}]W_1(1) \\ &= W_1(2)[B - (B-AB^{-1}C)]W_1(1) \\ &= W_1(2)(AB^{-1}C)W_1(1) \\ &\geqq \underset{\sim}{0} \end{aligned}$$

Second, if $W_1(n) - W_1(n-1) \leqq \underset{\sim}{0}$, then

$$\begin{aligned} W_1(n+1) - W_1(n) &= W_1(n+1)[(B-AW_1(n-1)C)-(B-AW_1(n)C)]W_1(n) \\ &= W_1(n+1)A(W_1(n) - W_1(n-1))C\,W_1(n) \\ &\leqq \underset{\sim}{0} \end{aligned}$$

This proves the lemma.

Theorem 5.3

If N is boundary leading, σ is the integer referred to in Lemma 5.5, and $V = -AW_1(\sigma)C$, then

(a) $\mathcal{L}_V \neq \phi$,

(b) $\mathcal{L}_V \subseteq \mathcal{L}$, and

(c) $T(\mathcal{L}_V) \subseteq \mathcal{L}_V$.

Proof: $\mathcal{L}_V$ is the set of all matrices S for which $S\underline{1} = A\underline{1}$ and $S \geqq V$. By Lemma 5.4,

$$V\underline{1} = -AW_1(\sigma)C\underline{1} \leqq A1.$$

Therefore, there is a nonnegative matrix (e. g. the diagonal matrix whose diagonal elements are the entries of $A\underline{1} - V\underline{1}$) which, when added to V, produces a matrix in $\mathcal{A}_V$. Thus (a) is true. By Lemma 5. 6, B+V has no essential index class for which the corresponding entries of $C\underline{1}$ are identically zero. If $S \in \mathcal{A}_V$, then $(B+S) \geq (B+V)$, and (Theorem 3. 3) every essential class of (B+S) contains at least one essential class of (B+V). But since $(B+S)\underline{1} = C\underline{1}$, then no essential class of (B+S) is conservative. With Theorem 4. 5, this proves (b). To prove (c), we first observe that (Lemma 4. 3) for every $S \in \mathcal{A}$, and therefore for every $S \in \mathcal{A}_V$, $T(S)\underline{1} = A1$. Second we observe (Lemma 5. 7) that $-W_1(\sigma+1) \geqq -W_1(\sigma)$, and that therefore $-AW_1(\sigma+1)C \geqq -AW_1(\sigma)C = V$. Finally, we observe that for all $S \in \mathcal{A}_V$, $S - V \geqq \underset{\sim}{0}$ and

$$\begin{aligned} T(s) + AW_1(\sigma+1)C &= -A(B+S)^{-1}C - A(B+V)^{-1}C \\ &= -A(B+S)^{-1}[I - (B+S)^{-1}A(B+V)^{-1}]C \\ &= -A(B+S)^{-1}[(B+V) - (B+S)](B+V)^{-1}C \\ &= A(B+S)^{-1}(S-V)(B+V)^{-1}C \\ &\geqq \underset{\sim}{0} \end{aligned}$$

and that, therefore

$$T(S) \geqq -AW_1(\sigma+1)C \geqq V.$$

Thus, for all $S \in \mathcal{A}_V$, $T(s) \in \mathcal{A}_V$. That completes the proof.

Theorem 5.4

If N is boundary leading, then

(a) There exists a fixed point S* of T, in $\mathcal{A}_V$, and therefore also in $\mathcal{A} \subseteq \mathcal{A}_V$.

(b) There exists a quadratic root $R \in \mathcal{R}$, satisfying Equation (4.8).

Proof: This theorem follows immediately from Theorems 4.1, 4.6 and 5.3.

6. Determination of R*

It is, of course, not very useful to know that a root matrix R exists in $\mathcal{R}$ unless there is the means to calculate it. One such means, proposed by Evans, is to repeatedly apply the mapping T to some initial iterate S_0, hoping that the sequence $T(S_0)$, $T^2(S_0)$, ..., will converge to a fixed point S*. Once S* has been determined, a root matrix R* is found by use of Theorem 4.1 and the formula

$$R^* = -A(B+S^*)^{-1}. \tag{6.1}$$

The question naturally arises whether the iteration will converge to a fixed point S* of T, and for what initial iterates.

Clearly, for $T^n(S_0)$ to converge, T must not map any iterate $T^n(S_0)$ into a matrix S for which (B+S) is singular, for otherwise $T^{n+1}(S_0) = T(S)$ would be undefined and the sequence S_0, $T(S_0)$, $T^2(S_0)$, $T^3(S_0)$, ..., is without meaning. To avoid this difficulty, the initial iterate must be chosen from a subset $\mathcal{H} \subseteq \mathcal{A}$ for which the mapping T has the property $T(\mathcal{H}) \subseteq \mathcal{H}$.

The strongest convergence theorem we have succeeded in proving follows. In it we adopt the notation $\rho(M)$ for the spectral radius of a matrix M.

Theorem 6.1

Let $\mathcal{H}$ be a subset $\mathcal{H} \subseteq \mathcal{A}$ for which $T(\mathcal{H}) \subseteq \mathcal{H}$.

Let S* be a fixed point of T for which $\rho(A(B+S^*)^{-1}) < 1$.

Then

$$\lim_{n\to\infty} T^n(S_0) = S^*$$

for every $S_0 \in \mathcal{H}$.

The following well known lemma is required for the proof of this theorem. The proof of the lemma will be found in [21, p. 13].

Lemma 6.1

If M is an $m \times m$ complex matrix, then

$$\lim_{n\to\infty} M^n = \underset{\sim}{0}$$

if and only if $\rho(M) < 1$.

With this lemma, we proceed to the proof of Theorem 6.1.

Proof (Theorem 6.1): The difference $T(S) - S^*$ is given by

$$T(S) - S^* = T(S) - T(S^*) = - A(B+S)^{-1}C + A(B+S^*)^{-1}C$$

$$= - A[(B+S)^{-1} - (B+S^*)^{-1}]C$$

$$= - A(B+S^*)^{-1}[(B+S^*) - (B+S)](B+S)^{-1}C,$$

for every $S \in \mathcal{S}$. Consequently,

$$T(S) - S^* = [- A(B+S^*)^{-1}](S-S^*)[- (B+S)^{-1}C] \qquad (6.2)$$

and, by repeated application of Equation (6.2) we have, for every $S_0 \in \mathcal{H}$,

$$T^n(S_0) - S^* = [- A(B+S^*)^{-1}]^n(S_0-S^*) \prod_{k=0}^{n-1} [-(B+T^k(S_0))^{-1}C]. \qquad (6.3)$$

However, since $-(B+S)^{-1}C$ is a stochastic matrix for every $S \in \mathcal{S}$, $-(B+T^k(S_0))^{-1}C$ is a stochastic matrix for every k=0, 1, 2, . . . , and

$$||- (B+T^k(S_o))^{-1}C|| = 1.$$

Therefore the norm of the product on the right is less than or equal to unity for every n. Consequently, from Equation (6.3) we get

$$||T^n(S_o) - S^*|| \leq ||[- A(B+S^*)^{-1}]^n|| \; ||S_o - S^*||. \quad (6.4)$$

But if $\rho(A(B+S^*)^{-1}) < 1$, then by Lemma 6.1

$$\lim_{n\to\infty}[- A(B+S^*)^{-1}]^n = \underline{0},$$

and thus

$$\lim_{n\to\infty} ||T^n(S_o) - S^*|| = 0. \quad (6.5)$$

For a finite dimensional normed space, convergence in norm is equivalent to convergence in components, so that the theorem is proved.

Using Theorem 5.3 , we get the corollary

Corollary 1

Let N be boundary leading. Then if a fixed point S^* of T exists for which $\rho(A(B+S^*)^{-1}) < 1$, we have

$$\lim_{n\to\infty} T^n(S_o) = S^*$$

for every $S_o \in \mathcal{S}_V$.

The theorem as given appears to be excessively restrictive. Experimentation with the iteration, with the purpose of discovering useful examples and counterexamples, produced remarkably consistent results. The only failures of convergence either occurred in non-boundary-leading processes or were the result of singularity

of $B+S_n$ where S_n was some iterate. The latter were presumably the result of choice of initial iterate which was not in $\mathcal{A}_V$, or some other suitable $\mathcal{A}$. In some cases this was checked and proved to be true. Specifically, the restriction "$\rho(A(B+S^*)^{-1}) < 1$" in the theorem appears to be unnecessary. Convergence to rather large root matrices were rather frequently obtained, as a result of deliberate challenges to the program.

However, the failure to find a stronger theorem, removing the spectral radius assumption, has not lead to serious difficulties in this theory. It will be seen (Chapter 8) that in almost all interesting cases, $\rho(A(B+S^*)^{-1}) < 1$ holds. Of that, more later.

There is one theoretical problem resulting from Theorem 6.1 which is more troublesome. It is very awkward to work with the set $\mathcal{A}_V$. A separate calculation of V is required and, with large m, it is rarely plain when σ iterations have been completed and V has been obtained.

One escape from this difficulty is to take one's chances with singularities in $B+T^n(S_o)$ and simply restart with a perturbed $T^n(S_o)$ when it occurs. This will seem more attractive with the observation that the interruption will only occur early in the iteration process. This will follow from the readily proved fact that:

Lemma 6.2

If N is boundary leading, then

$$T^n(S) \geqq -AW_1(n)C \tag{6.6}$$

for every $S \in \mathcal{A}$ and for every integer n for which $T^n(S)$ is defined.

Proof: The proof will be by induction on n. From Lemma 5.1 we have

$$W_1(n+1) = (B - AW_1(n)C)^{-1}$$

and

$$W_1(1) = B^{-1}.$$

Thus,

$$T(S) + AW_1(1)C = -A(B+S)^{-1}C + AB^{-1}C$$

$$= A[B^{-1} - (B+S)^{-1}]C = AB^{-1}[(B+S) - B](B+S)^{-1}C$$

$$= A[-B^{-1}]S[-(B+S)^{-1}]C$$

Since every factor is a nonnegative matrix, we have

$$T(S) + AW_1(1)C \geqq \underline{0}$$

and therefore

$$T(S) \geqq -AW_1(1)C.$$

For general n, we have, similarly,

$$T^{n+1}(S) + AW_1(N+1)C = -A(B+T^n(S))^{-1}C + A(B - AW_1(n)C)^{-1}$$

$$= A(B - AW_1(n)C)^{-1}[T^n(S) + AW_1(n)C](B+T^n(S))^{-1}C$$

$$= A[-W_1(n+1)][T^n(S) + AW_1(n)C][-(B+T^n(S))^{-1}]C.$$

But $[-W_1(n+1)]$ is nonnegative, as is $[-(B+T^n(S))^{-1}]$, and $[T^n(S) + AW_1(n)C]$ is nonnegative by the induction hypothesis.

Therefore

$$T^{n+1}(S) + AW_1(n+1)C \geqq \underline{0}$$

and the induction is complete, proving that

$$T^n(S) \geqq - AW_1(n)C$$

for all integers n for which $T^n(S)$ is defined (i. e. all n such that $T^k(S) \in \mathcal{A}$ for k=1, 2, . . . , n-1.)[1]

The set $\mathcal{A}_V$ is the set of matrices S' in $\mathcal{A}$ for which $S' \geqq V = - AW_1(\sigma)C$, where σ is the integer defined in Lemma 5. 5. Lemma 6. 2 shows that if N is boundary leading and if $T^k(S_o)$ is in $\mathcal{A}$ for every k=1, 2, . . . , σ, then $T^\sigma(S_o)$ is in V because $T^\sigma(S) \geqq V = - AW_1(\sigma)C$. By Theorem 5. 3 $T(\mathcal{A}_V) \subseteq \mathcal{A}_V$, so that $T^k(S_o)$ must also be defined for k=σ+1, σ+2, ... In short, if N is boundary leading, and the iteration "survives" σ iterations without leaving the set $\mathcal{A}$, then the iteration is "safe" from leaving $\mathcal{A}$ thereafter. Usually σ is a very small integer, since it is at most equal to the number of essential classes of B.

An alternative escape from the difficulty of finding $\mathcal{A}_V$ is to find another set which can play the role of $\mathcal{H}$ in Theorem 6. 1. It need only satisfy $\mathcal{A} \supseteq \mathcal{K} \supseteq T(\mathcal{K})$. Unlike $\mathcal{A}_V$, $\mathcal{K}$ does not have to be convex, compact, or closed. For example, if N is boundary leading, the set

$$\{T^n(A),\ n=1, 2, \ldots\}$$

can be shown to qualify, so that the iteration will remain in $\mathcal{A}$ if

---

[1]Note that this proof can also be used to prove that $- AW_1(n)C = T^n(\underline{0})$, if the domain of T is extended to the set $\{S:\ S \geqq 0$ and (B+S) is nonsingular$\}$.

A is chosen as the initial iterate. Of course, convergence depends on whether the spectral radius property holds.

7. Equilibrium Distributions

The construction of an equilibrium solution offers little theoretical difficulty. We have the theorem:

Theorem 7.1

If there exists a root $R \in \mathcal{R}$ of the matrix quadratic $A + XB + X^2C$, then

(a) There exists a sequence $\pi = \{\pi_k,\ k \in \mathcal{N}\}$, which satisfies $\pi \geq 0$ and $\pi U = 0$.

(b) The sequence π is of the partitioned form $\pi = [\phi_0, \phi_1, \ldots]$, where ϕ_0 satisfies $\phi_0 \geq 0$ and

$$\phi_0(E + \hat{R}\hat{C}) = \underline{0}, \tag{7.1}$$

and where

$$\phi_1 = \phi_0\hat{R} \tag{7.2}$$

$$\phi_n = \phi_{n-1}R = \phi_1(R)^{n-1} \tag{7.3}$$

$$n=2, 3, 4, \ldots .$$

(c) If $\rho(R) < 1$, then π is convergent.

The proof of this theorem requires the following lemma, due to Ostrowski. (See [21, p. 82] for proof.)

Lemma 7.1

Let M be an arbitrary complex matrix. Then the series $I + M + M^2 + \ldots$ converges if and only if $\rho(M) < 1$. If $\rho(M) < 1$, then (I-M) is nonsingular and the series converges to $(I-M)^{-1}$.

Having this lemma, we proceed with the proof.

Proof (Theorem 7.1): In the early discussion of Chapter 4, it was observed that if a root R of $A + XB + X^2C = \underline{0}$ exists in $\mathcal{R}$, then Equation (7.1) has a solution $\phi_0 \geq \underline{0}$. It was observed in Chapter 2 that the partitioned sequence $\pi = [\phi_0, \phi_1, \ldots]$ as defined by Equations (7.1), (7.2) and (7.3) satisfies $\pi U = 0$, and that $\pi \geq \underline{0}$. Thus, conclusions (a) and (b) were already proven. Clearly, π converges if $\phi_1 + \phi_2 + \ldots$ converges. Using Equation (7.3) we have

$$\phi_1 + \phi_2 + \ldots = \phi_1 + \phi_1 R + \phi_1 R^2 + \ldots \tag{7.4}$$

and, provided the series $I + R + R^2 + \ldots$ converges, we can factor out the vector ϕ_1 so that

$$\phi_1 + \phi_2 + \ldots = \phi_1(I + R + R^2 + \ldots) \tag{7.5}$$

Since by hypothesis $\rho(R) < 1$, then by Lemma 7.1 the series $I + R + R^2 + \ldots$ does converge. Therefore, π is convergent, (c) is true, and the proof is complete.

The converse of (c), i.e. that if π is convergent then $\rho(R) < 1$, is not generally true for boundary leading QBD processes.

Example 7.1: The process whose transition graph is shown in Figure 7.1 has a root

$$R = \begin{bmatrix} 2 & 0 \\ 0 & .5 \end{bmatrix},$$

so that $\rho(R) = 2$, and yet it has an equilibrium solution

$$\pi = \{0, 1, 0, .5, 0, .25, \ldots\}$$

which is convergent.

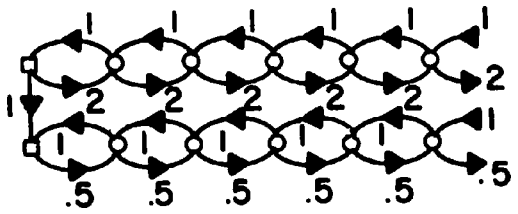


Figure 7.1

A Graph of a Process for which $\rho(R) > 1$ and π is Convergent

The matrix $(E+\hat{R}\hat{C})$ is a conservative Q-matrix for every $R \in \mathcal{R}$ (c.f. Chapter 4), and therefore Equation (7.1) could be regarded as defining a finite Markov chain with $E + \hat{R}\hat{C}$ as its transition intensity matrix. Consequently, any technique satisfactory for determining equilibrium solutions for finite Markov chains can be applied to determine a vector ϕ_0 satisfying (7.1). Once ϕ_0 is determined, Equations (7.2) and (7.3) can be applied routinely to obtain a solution π, or as much of that solution as one wishes to expand. Theorem 7.1 shows, further, that if $\rho(R) < 1$, then that solution is necessarily convergent.

One technique for determining ϕ_0 from Equation (7.1) follows that of the "Recursive Queue Analyzer" [23], and develops the iterations

$$\beta_{n+1} = \beta_n[\frac{1}{\Delta}(E + \hat{R}\hat{C}) + I] \qquad (7.6)$$

where β_n, $n=0, 1, 2, \ldots$, is a row vector with m_o entries, Δ is a

positive scalar slightly greater than the absolute value of the largest diagonal entry to $E + \hat{R}\hat{C}$, and β_0 is an otherwise arbitrary row-vector for which $\beta_0 \geq \underline{0}$. Proof that the iteration (7.6) converges to an equilibrium solution ϕ_0 is found in [23]. (If the matrix $E + \hat{R}\hat{C}$ is reducible, the ϕ_0 found is not unique and will depend on the choice of β_0. Nevertheless, any sequence π constructed from any ϕ_0 is an equilibrium solution, so that the solution π is similarly not unique.)

If $\rho(R) < 1$, then an equilibrium distribution can be found by normalization, through the use of the following theorem.

Theorem 7.2

If, under the conditions of Theorem 7.1, $\rho(R) < 1$, then $\bar{\pi} = [\bar{\phi}_0, \bar{\phi}_1, \ldots]$ is an equilibrium distribution, where

$$\bar{\phi}_0 = \phi_0/s \tag{7.7}$$

$$\bar{\phi}_1 = \bar{\phi}_0\hat{R} \tag{7.8}$$

$$\bar{\phi}_n = \bar{\phi}_{n-1}R = \bar{\phi}_1(R)^{n-1} \tag{7.9}$$

and where s is a scalar:

$$s = \phi_0 \begin{bmatrix} (I-R)^{-1}\underline{1} \\ \\ \underline{1} \end{bmatrix} \tag{7.10}$$

Proof: We show that

$$s = \sum_{k \in \mathcal{N}} \phi_k, \tag{7.11}$$

so that it normalizes $\bar{\pi} = \pi/s$ to a distribution. The sum (7.11) can

also be written

$$\sum_{k \in \mathcal{K}} \phi_k = \phi_0 \underline{1} + \phi_1 \underline{1} + \ldots = \phi_0 \underline{1} + (\phi_1 + \phi_2 + \ldots)\underline{1} \qquad (7.12)$$

since it is (absolutely) convergent. By Equations (7.2) and (7.5)

$$\phi_0 \underline{1} + (\phi_1 + \phi_2 + \ldots)\underline{1} = \phi_0 \underline{1} + \phi_0 \hat{R} (I + R + R^2 + \ldots)\underline{1} . \qquad (7.13)$$

The right hand side is equivalent to the partitioned form

$$\phi_0 \begin{bmatrix} 1 \\ \\ 1 \end{bmatrix} + \phi_0 \begin{bmatrix} (R + R^2 + \ldots)\underline{1} \\ \\ \underline{0} \end{bmatrix} = \phi_0 \begin{bmatrix} (I + R + R^2 + \ldots)\underline{1} \\ \\ \underline{1} \end{bmatrix} .$$

Finally, by Lemma 3.5, we write

$$\phi_0 \begin{bmatrix} (I+R + R^2 + \ldots)\underline{1} \\ \\ \underline{1} \end{bmatrix} = \phi_0 \begin{bmatrix} (I - R)^{-1}\underline{1} \\ \\ \underline{1} \end{bmatrix} ,$$

proving the theorem.

It is possible to determine the equilibrium distribution more directly, while at the same time determining "marginal" equilibrium distributions on the second (finite) component of the process $N = \langle N_1, N_2 \rangle$. Let γ be a row vector whose i-th entry is the sum of the i-th entries of all $\bar{\phi}_n$, $n=0, 1, 2, \ldots$.

$$\gamma = \bar{\phi}_0 + [\bar{\phi}_1 \ \underline{0}] + [\bar{\phi}_2 \ \underline{0}] + [\bar{\phi}_3 \ \underline{0}] + \ldots . \qquad (7.14)$$

We shall refer to γ as the marginal equilibrium distribution on N_2, since the sum of its entries is unity. Using Equations (7.8) and (7.9), and again using the absolute convergence of Equation (7.4), it is easily verified that

$$\gamma = \bar{\phi}_0 \begin{bmatrix} (I-R)^{-1} & 0 \\ 0 & I \end{bmatrix}, \tag{7.15}$$

and that (I-R) is nonsingular. Equation (7.15) implies that

$$\bar{\phi}_0 = \gamma \begin{bmatrix} I-R & 0 \\ 0 & I \end{bmatrix} . \tag{7.16}$$

Substitution of Equation (7.16) into Equation (7.1), and letting E be partitioned so that

$$E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}, \tag{7.17}$$

where E_{11} is $m \times m$ and E_{22} is $(m_o-m) \times (m_0-m)$, gives the equation

$$\gamma \begin{bmatrix} (I-R)(E_{11} - RC) & E_{12} \\ (I-R)E_{21} & E_{22} \end{bmatrix} = \underline{0} \tag{7.18}$$

It can be solved under the restriction

$$\gamma \underline{1} = 1, \tag{7.19}$$

whereupon $\bar{\phi}_0$ is automatically found from Equation (7.16). Usually, (7.18) and (7.19) are more easily solved than (7.1), (7.7) and (7.10) since at least one matrix inversion is eliminated. Sometimes γ is all that is desired.

We express these results as the Theorem 7.3 which has just been proved.

Theorem 7.3

If, under the conditions of Theorem 7.1, the spectral radius $\rho(R) < 1$, then for every equilibrium distribution $\bar{\pi}$ there is a marginal equilibrium distribution γ which is a solution of

$$\gamma \begin{bmatrix} (I-R)(E_{11}-RC) & E_{12} \\ (I-R)E_{21} & E_{22} \end{bmatrix} = \underline{0} \tag{7.18}$$

and

$$\gamma \underline{1} = 1. \tag{7.19}$$

Furthermore,

$$\bar{\phi}_0 = \gamma \begin{bmatrix} (I-R) & 0 \\ 0 & I \end{bmatrix}$$

It often happens that the equilibrium mean of N_1 conditioned on N_2 is desired. One can express this mean as an m-component row-vector γ_1 whose k-th entry is the equilibrium mean of N_1 conditional on $N_2=k$ (Observe that for $k > m$, N_1 can have only value zero):

$$\begin{aligned} \gamma_1 &= \bar{\phi}_1 + 2\bar{\phi}_2 + 3\bar{\phi}_3 + \ldots \\ &= \bar{\phi}_1(I + 2R + 3R^2 + \ldots). \end{aligned}$$

The calculation of γ_1 is facilitated by the theorem

Theorem 7.4

Let M be an arbitrary complex matrix. Then the series $I + 2M + 3M^2 + \ldots$ converges if and only if $\rho(M) = r < 1$. The series converges to $(I-M)^{-2}$.

Proof: This theorem is analogous to Lemma 7.1. We first assume that $\rho(M) < 1$. We have the identity

$$(I+2M+3M^2+\ldots+nM^{n-1})(I-M)^2 - I = [n(I-M) + I]M^n$$

By multiplying both sides by $[(I-M)^{-1}]^2$, (which we write $(I-M)^{-2}$), we have

$$(I+2M+3M^2+\ldots+nM^{n-1}) - (I-M)^{-2} = [n(I-M)+I]M^n(I-M)^{-2}$$

Thus

$$||(I+2M+3M^2+\ldots+nM^{n-1}) - (I-M)^{-2}||$$

$$\leq ||[n(I-M) + I]M^n||\ ||(I-M)^2||, \qquad (7.17)$$

for all $r \geq 1$. But $||[n(I-M)+I]M^n|| = O[(n+1)r^n]$ and therefore the left hand side of (7.17) converges to zero. Thus the series converges (it is finite dimensional) to $(I-M)^{-2}$. Conversely, if the series converges, let μ be an eigenvalue, corresponding to the eigenvector x. Then

$$(I + 2M + 3M^2 + \ldots)x = (1 + 2\mu + 3\mu^2 + \ldots)x,$$

and the convergence of the series implies convergence of $\mu + 2\mu^2 + 3\mu^3 + \ldots$ for any eigenvalue of M. It is well known that this implies $|\mu| < 1$ for all eigenvalues of M, and thus that $\rho(M) < 1$, completing the proof.

As a consequence of this theorem we have

$$\gamma_1 = \bar{\phi}_1(I-R)^{-2} \tag{7.18}$$

and, from Equations (7.2) and (7.16), we get

$$\gamma_1 = \gamma \begin{bmatrix} R(I-R)^{-1} \\ \\ 0 \end{bmatrix} \tag{7.19}$$

(recalling that $R(I-R) = (I-R)R$).

The equilibrium second moment of N_1 conditional on N_2, which is correspondingly:

$$\gamma_2 = \bar{\phi}_1 + 4\bar{\phi}_2 + 9\bar{\phi}_3 + \ldots$$

$$= \phi_1(I + 4R + 9R^2 + \ldots)$$

can similarly be calculated by the use of the theorem

Theorem 7.5

> Let M be an arbitrary complex matrix. Then the series $I + 4M + 9M^2 + 16M^3 + \ldots$ converges if and only if $\rho(M) = r < 1$. Then the series converges to $(I+M)(I-M)^{-3}$.

The proof of this theorem follows the same manipulations as previously, starting the with identity

$$(I-M)^3(I + 4M + 9M^2 + \ldots + n^2M^{n-1}) - (I+M) = [n^3M - (n^2+2n-1)I]M^n$$

It need not be repeated here. Using this theorem, we have

$$\gamma_2 = \bar{\phi}_1(I+R)(I-R)^{-3} \tag{7.20}$$

and

$$\gamma_2 = \gamma \begin{bmatrix} R(I+R)(I-R)^{-2} \\ \\ 0 \end{bmatrix} \tag{7.21}$$

(again recalling the commutativity of products of matrix polynomials).

8. Recurrent Processes

A state j of a continuous-parameter Markov chain Z is said to be recurrent if

$$\int_0^\infty p_{jj}(t)dt = \infty, \tag{8.1}$$

where, by definition,

$$p_{jk}(t) = pr[Z(t) = k | Z(0) = j]. \tag{8.2}$$

The state j is said to be non-recurrent otherwise. A recurrent state is said to be positive recurrent if, in addition,

$$\lim_{t\to\infty} p_{jj}(t) > 0, \tag{8.3}$$

and null recurrent if

$$\lim_{t\to\infty} p_{jj}(t) = 0. \tag{8.4}$$

These properties are shared by all states in a communicating class. Consequently, we can say that a communicating class of states is non-recurrent, positive recurrent, or null recurrent if the states in in the class are non-recurrent, positive recurrent, or null recurrent, respectively. Every recurrent class is essential, and every finite essential class is positive recurrent. However, essential classes containing an infinite number of states are not necessarily positive recurrent, or even null recurrent.

When a QBD process is used as a model of a physical system, the presence of an essential state (and therefore an essential class)

which is not positive recurrent signifies that the system is "unstable" in the sense that, with nonzero probability, the state of the system increases without bound as time elapses. This is a condition generally to be avoided in design of systems, and its mere occurrence generally makes further calculation unnecessary. However, the presence of an essential class which is not positive recurrent is often difficult to detect by casual observation or physical argument, and therefore an important capability of a computational technique is to provide such detection automatically. Concomitantly, it is not usually important that a computational technique provide solutions when such a class is detected.

A key result of this section will be to show that a necessary and sufficient condition for an irreducible QBD process to be positive recurrent is that there be a fixed point S^* of T for which $\rho(A(B+S^*)^{-1}) < 1$. Since the latter was already shown (Chapter 6) to be sufficient for the convergence of the iteration process T^n to a unique fixed point S^*, it follows that this procedure for calculating S^* will work for every irreducible, positive recurrent QBD process, and that failure to converge, or convergence to a limit S^* for which $\rho(\bar{A}(B+S^*)^{-1}) \geq 1$ is conclusive evidence that the process is not positive recurrent. Hence, we will have a test for positive recurrence, and also assurance that the computational technique works in every interesting case. For boundary leading QBD processes similar, but weaker, results will be demonstrated.

The following theorem, due to R. G. Miller [16] provides the key to (1) the relationship between positive recurrence of a continuous-parameter Markov chain and existence of convergent equilibrium solutions, (2) the uniqueness of equilibrium solutions, and (3) the relationship between the limiting distribution and the equilibrium solutions. It is an adaptation to continuous-parameter Markov chains of a very powerful theorem of F. G. Foster (see [10], and [17, p. 256]) for discrete-parameter Markov chains. The proof will be omitted here.

Theorem 8.1 (R. G. Miller)

> If a continuous-parameter Markov chain Z is uniquely determined by its transition intensity matrix Q, and if Q is irreducible and conservative, then Z is positive recurrent if and only if $\pi Q = \underline{0}$ has a convergent solution $\pi \geq \underline{0}$. The solution is unique (up to a multiplicative constant) and positive, and is equal to the limiting distribution (within a multiplicative constant).

Specializing this theorem to QBD processes, we obtain the following corollary.

Corollary 1

> If N is an irreducible QBD process, then it is positive recurrent if and only if $\pi U = \underline{0}$ has a convergent solution $\pi \geq 0$. The solution is unique (up to a multiplicative constant) and positive, and is equal to the limiting

distribution (within a multiplicative constant).

Proof: Every continuous parameter Markov chain whose transition intensity matrix is bounded (i. e. whose entries are bounded) and conservative is uniquely determined by that intensity matrix. (See Chung [1, p. 237]). Since no element of U, the intensity matrix of N, can be greater in absolute value than the diagonal elements of E or B, and since there are finitely many of those, then the intensity matrix of N is bounded. By definition, every QBD process is conservative. Hence N is uniquely determined by its intensity matrix, and the antecedent of Theorem 8.1 is satisfied, proving the corollary.

Corollary 2

If N is an irreducible QBD process, then it is positive recurrent if and only if there exists a fixed point S^* of T for which $\rho(A(B+S^*)^{-1}) < 1$.

Proof: If there exists such a fixed point S^*, then by Theorem 4.1, $R^* = -A(B+S^*)^{-1}$ is a root of the quadratic $A + XB + X^2C$. But then $\rho(R^*) < 1$, and by Theorem 7.1 a convergent sequence $\pi \geq 0$ can be constructed which satisfies $\pi U = \underline{0}$. By Corollary 1, N is positive recurrent. To prove the converse, we observe by Corollary 1 that if N is positive recurrent then there exists a convergent solution $\pi \geq 0$ to $\pi U = \underline{0}$, which is unique and positive. However, by Theorem 5.4, there exists a root $R \in \mathcal{R}$ of $A + XB + X^2C$ (recall that irreducible QBD processes are boundary leading). Theorem

7.1 then tells us that $\pi = [\phi_0, \phi_1, \phi_2, \ldots]$ is a solution to $\pi U = \underline{0}$, and $\pi \geq \underline{0}$. It must, therefore, be the unique, positive, convergent solution of Corollary 1. Moreover, since it is convergent $\sum_{n=1}^{\infty} \phi_n = \sum_{n=1}^{\infty} \phi_1 R^{n-1}$ must converge. Since ϕ_1 is positive, $\sum_{n=1}^{\infty} \phi_1 R^n$ can converge only if

$$\lim_{n \to \infty} R^n = 0$$

This implies that $\rho(R) < 1$ (cf. Lemma 6.1), completing the proof.

The preceding two corollaries give the desired assurances for irreducible processes. It is clear that if a QBD process is positive recurrent, the conditions for convergence of the iteration on T are satisfied, so that a solution can always be constructed. It is also clear that convergence of T and construction of a convergent π (or determination that $||R|| < 1$) is sufficient assurance that the solution is the desired limiting distribution, and that the process is positive recurrent. Conversely, failure of T to converge (an occurrence which has not been observed to date), or divergence of the constructed π, is sufficient evidence that the irreducible QBD process is not positive recurrent, and that its limiting distribution is therefore pathological.

These results extend to boundary leading QBD processes, with appropriate changes in terminology. First, for a model to be useful in estimating limiting probabilities, it is only necessary for all the essential states to be positive recurrent. Inessential states

are necessarily non-recurrent, but do not (at least in boundary leading processes) admit the possibility of "instability" or "continued growth" referred to earlier. Inessential states always have a limiting probability of zero, and the equilibrium solutions reflect this property, as will be demonstrated.

Secondly, when there is more than one essential class, the limiting distribution does not exist. Rather the limiting probabilities depend on the initial distribution. Thus, the equilibrium solution should not be expected to be unique.

In general, an essential communicating class of states of a reducible QBD process does not separately define an irreducible QBD process, as is the case for Markov chains in general. Figure 8.1a illustrates the transition graph of a boundary leading process whose subgraphs using any particular essential class of states is not a QBD process. Figure 8.1b shows a subgraph, and it is seen

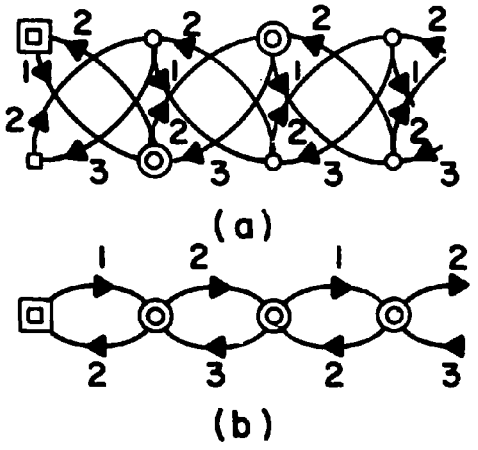


Figure 8.1

A Reducible QBD Process and One of its Subgraphs

that the transition intensities do not repeat properly. Therefore, the boundary leading QBD processes must be treated as a distinctive entity, rather than relying on the usual arguments for Markov chains which justify study of only irreducible processes by "splitting up" the original process. The analogue of Corollary 1 can be expressed as follows.

Theorem 8.2

If N is boundary leading, then every essential state of N is positive recurrent if and only if $\pi U = 0$ has a convergent solution $\pi = \{\pi_j : j \in \mathcal{N}\}$ such that the components π_j are positive for every essential j and zero for every inessential j.

Proof: Since the matrix $U = \{u_{jk} : j, k \in \mathcal{N}\}$ is conservative, then equivalently

$$\sum_{k \in \mathcal{N}} u_{jk} = 0, \quad j \in \mathcal{N} \tag{8.11}$$

Then let θ be the number of essential communicating classes in N and consider an arbitrary essential communicating class I_α, $\alpha = 1, 2, \ldots, \theta$. Since it is essential, then $u_{jk} = 0$ whenever $j \in I_\alpha$ and $k \notin I_\ell$. Therefore

$$\sum_{k \in I_\alpha} u_{jk} = 0, \qquad j \in I_\alpha, \tag{8.12}$$

and the matrix $Q_\alpha = \{u_{jk} : j, k \in I_\alpha\}$ is conservative. It is also

irreducible (by definition of "communicating class"), bounded, and uniquely defines a Markov chain Z_α. By Theorem 8.1 that Markov chain is positive recurrent if and only if there is a convergent vector $\nu_\alpha \geq 0$ for which $\nu_\alpha Q_\alpha = 0$.

Now, suppose that $\pi U = 0$ has a convergent solution $\pi = \{\pi_j : j \in \mathcal{N}\}$ for which $\pi_j > 0$ for every essential j and $\pi_j = 0$ for every inessential j. Then

$$\sum_{j \in \mathcal{N}} \pi_j u_{jk} = 0, \qquad k \in \mathcal{N}, \tag{8.13}$$

and since for every $j \notin I_\alpha$ either $\pi_j = 0$ (inessential j) or $u_{jk} = 0$ for all $k \in I_\alpha$ (essential j), then

$$\sum_{j \in I_\alpha} \pi_j u_{jk} = 0, \qquad k \in I_\alpha. \tag{8.14}$$

This demonstrates that $\nu_\alpha = \{\pi_j : j \in I_\alpha\}$ is a positive solution to $\nu_\alpha Q_\alpha = 0$. The sequence π is absolutely convergent, so that ν_α is necessarily convergent. Thus the Markov chain Z_α is positive recurrent and $\nu_\alpha > 0$. This, in turn, demonstrates that every state in I_α is positive recurrent and $\pi_j > 0$ for every essential j. (Recall that positive recurrence of $j \in I_\alpha$ is defined by a property of $\mathrm{pr}[N(t) = j \mid N(0) = j]$, and that since every sample function of N for which N(0) = j must remain forever in I_α, then $\mathrm{pr}[N(t) = j \mid N(0) = j] = \mathrm{pr}[Z_\alpha(t) = j \mid Z_\alpha(0) = j]$. As a consequence, positive recurrence of Z_α implies positive recurrence of each

$j \in I_\alpha$.) Every essential state is in some essential class I_α, so that this proves the necessity of positive recurrence.

Conversely, suppose every essential state is positive recurrent. Then every Z_α is also positive recurrent, and there exists a convergent solution $\nu_\alpha > 0$ to $\nu_\alpha Q_\alpha = 0$. Write that solution as $\nu_\alpha = \{\pi_j : j \in I_\alpha\}$. Then

$$\sum_{j \in I_\alpha} \pi_j u_{jk} = 0 \qquad k \in I_\alpha. \tag{8.15}$$

Now let $\bar{I}$ be the set of inessential states of N. Since $u_{jk} = 0$ for every $j \notin \bar{I}$ and $k \in \bar{I}$,

$$\sum_{j \in \bar{I}} \pi_j u_{jk} = 0, \qquad k \in \bar{I}. \tag{8.16}$$

The sequence $\{\pi_j : j \in \bar{I}\} = 0$ is surely a solution to Equation (8.16). By Equations (8.15) and (8.16) we have demonstrated that $\pi = \{\pi_j : j \in \mathcal{N}\}$ is a solution to $\pi U = 0$ for which $\pi_j > 0$ for every essential j, and $\pi_j = 0$ for every inessential j. This completes the proof.

It should be noted that the restriction that "$\pi_j > 0$ for every essential state" could be weakened to "$\pi_j > 0$ for at least one state in each essential class" without affecting the proof. In that case, "$\pi_j > 0$ for every essential state" becomes an additional conclusion of the theorem, as it is in Corollary 1 of Theorem 8.1. Every component solution ν_α is unique <u>within a constant</u>. If the above

qualification were not applied, it would be possible to demonstrate a convergent solution for a process which had a null-recurrent class I_α by the simple expedient of setting $\pi_j = 0$ for every $j \in I_\alpha$.

Example 7.1 illustrates why the boundary leading analogue of Corollary 2 of Theorem 8.1 must be significantly weaker. The existence of a fixed point S* for which $\rho(A(B+S^*)^{-1}) < 1$ is clearly not necessary for the essential states of N to be positive recurrent. It appears, however, to be sufficient. The difficulty stems from the existence of inessential states of N which might have been essential with a different E-matrix. (Observe that properties of T, S* and R* do not depend at all upon E.) It appears that if every state of a boundary leading N is positive recurrent, then it is necessary that $\rho(A(B+S^*)^{-1}) < 1$. But if every state is positive recurrent then every state is essential, so that we have a very different kind of a process. Weaker conditions that imply that $\rho(A(B+S^*)^{-1}) < 1$ can be found. For example, it appears to hold if one assumes simply that "every infinite communicating class (is) positive recurrent." (Proof of these suppositions requires development of relationships between communication properties of indices of S* and of boundary states of N, and will not be given.)

9. Uniqueness of S*

The question of whether or not there is more than one root of $A + RB + R^2C$ in $\mathcal{R}$ is not central to this theory. The central issue is whether the equilibrium solution π is unique , and whether it is given by the procedure of Chapter 7. If such a solution is unique, the existence of more than one R* which generates that solution is not disturbing. If the solution is not unique, it is similarly not too disturbing if the multiplicity comes from multiple solutions to the boundary equation (Equation 7.1), or from multiplicity of R*.

Nevertheless, a better intuitive feel for the mathematics of QBD-processes is a likely by-product of considering questions of uniqueness. Also, the answers to these questions may be useful in future development of QBD theory.

Theorem 6.1 implies that if T has a fixed point $S^* \in \mathcal{S}$ for which $\rho(A(B+S^*)^{-1}) < 1$, then there is no fixed point of $\mathcal{F}$, other than S*, in any set $\mathcal{H}$ for which $T(\mathcal{H}) \subseteq \mathcal{H} \subseteq \mathcal{S}$. (Otherwise $T^n(S_0)$ would not converge to S* for every $S_0 \in \mathcal{H}$.) This result can be extended to the theorem:

Theorem 9.1

If there exists a fixed point S* of T for which $\rho(A(B+S^*)^{-1}) < 1$, then there is exactly one fixed point of T .

Proof: Suppose that S' is a fixed point of T. Then $T(S') = S'$ and the set $\mathcal{H} = \{S'\}$ has the property $T(\mathcal{H}) \subseteq \mathcal{H} \subseteq \mathcal{A}$. Furthermore

$$\lim_{n \to \infty} T^n(S') = S'.$$

However, by Theorem 6.1

$$\lim_{n \to \infty} T^n(S') = S^*.$$

Therefore $S' = S^*$, and the theorem is proved.

Note that this theorem does not require any property of N other than that a fixed point of T exist with the given property. Even that property may be unnecessary, but a proof of a stronger theorem is not available. (An example of a QBD process with multiple fixed points of T has not been found, either.)

It does not follow from the uniqueness of S^* that there is, necessarily, a unique root of the quadratic $A + XB + X^2C$ in the set $\mathcal{R}$. The matrix $R^* = -A(B+S^*)^{-1}$ is, of course, a root. Furthermore, it can be readily shown that if R is a root and RC is in $\mathcal{A}$, then RC is a fixed point of T. (Thus $S^* = R^*C$.) However, since C is often singular, it is conceivable that $R = R^*$ is not the only root matrix for which $RC = S^*$. Furthermore, if R is a root in $\mathcal{R}$, it does not necessarily follow that RC is in $\mathcal{A}$, since B+RC may be singular. Thus, even if there is only one root with $RC \in \mathcal{A}$, there may still be more than one root in $\mathcal{R}$.

10. Conclusions

It has been shown that, under common and predictable circumstances, the QBD-processes lend themselves to expeditious numerical solution, and are therefore a useful class of stochastic models for computer system analysis (or other queueing application). The equilibrium distributions and other related measures are determined by a procedure which involves operations with only finite matrices, and which is analogous to procedures used in solution of the difference equation of simple birth and death processes.

This conclusion has been demonstrated by a comprehensive algebraic theory whose two main theoretical conclusions can be summarized as:

Conclusion 1.

If a QBD process is boundary leading, then the matrix quadratic $A + XB + X^2C$ is, indeed, analogous to the characteristic equation for the birth and death process, having a root matrix R which generates the characteristic solution

$$\phi_n = \phi_1 R^{n-1} \qquad n=2, 3, \ldots .$$

The m-vector ϕ_1 satisfies a boundary condition

$$\phi_0(E + \hat{R}\hat{C}) = \underline{0}$$

$$\phi_1 = \phi_0\hat{R},$$

and the equilibrium solution is given by

$$\pi = [\phi_0, \phi_1, \phi_2, \ldots]$$

The root matrix is in the known set $\mathcal{R}$. If π is convergent, the equilibrium distribution is merely a scalar multiple of π.

Conclusion 2.

If the QBD process is irreducible and positive recurrent, then the root R can be found by a well-defined iteration process

$$S_{n+1} = T(S_n)$$

$$R = C(B + S_\infty)^{-1}$$

and the resulting solution π will be convergent. Conversely, if the process is irreducible and a root R is found for which the resulting π is convergent, then the process is positive recurrent.

Much more has been said, of course, but these results are the most significant and general. The theory presented is broad enough so that there are numerous specific questions about QBD processes which can be answered by intelligent use of the theorems, corollaries, and lemmas explicitly supplied. Thus, this theory provides a good base on which to build even greater usefulness.

The most important area for further work is in the process for determining the root matrix. Firstly, it was observed that the interaction on T has been found, empirically, to converge even in non-recurrent irreducible, and in reducible, but boundary leading, QBD

processes. Both of these cases do not have the advantage of valid proofs of convergence (except, in the latter case, when the spectral radius property $\rho(A(B+S)^{-1}) < 1$ holds). Stronger proofs should be available. Secondly, it does not appear likely that the iteration T, requiring, as it does, an inversion on every iteration, is the most efficient procedure possible for finding the root R. A case in point is the iteration,

$$R_{n+1} = - D_n B^{-1}(A + CR_n^2),$$

which has been studied extensively. (D_n is a diagonal matrix chosen to keep R_{n+1} in $\mathcal{R}$.) A series of 1800 cases with diverse choices of A, B, and C have been run, and this iteration has converged in every boundary leading case. No proof that this should be true has been found. Nevertheless, the inverse of B can be calculated once, and no matrix inversion is required thereafter. Thus, this iteration is probably more efficient than the iteration of T^n, which requires inversion on every iteration.

In the last analysis, when one considers the variety of birth and death processes which have been studied, and the results concerning them which are available, one realizes the potential for future work that is here. How much of birth and death theory can be generalized by the act of replacing scalar transition-intensities by submatrices? This work has probably only scratched the surface in providing the answer. However, it should provide a firm base from which to generalize and develop a larger theory.

References

[1] Chung, K. L., Markov Chains with Stationary Transition Probabilities, (Hamburg: Springer-Verlag, 1960).

[2] Collatz, L., "Aufgaben monotoner Art," Arch. Math., Vol. III, 1952, pp. 366-376.

[3] Disney, R. L., "Some Multichannel Queueing Problems with Ordered Entry," J. Industrial Engineering, (Technical Notes) Vol. XIII, No. 1, January-February, 1962.

[4] Evans, R. V., "The Structure of Some Two-Dimensional Queueing Systems," Research Report 85, Management Sciences Research Project, University of California, Los Angeles, July 1963.

[5] Evans, R. V., "Numerical Methods for Queues with Complex Service Rates," Research Report 82, Management Sciences Research Project, University of California, Los Angeles.

[6] Evans, R. V., "Geometric Distribution in Some Two-Dimensional Queueing Systems," Operations Research, Vol. 15, No. 5, September-October 1967, pp. 830-846.

[7] Fadeeva, V. N., Computational Methods of Linear Algebra (Transl. by C. D. Benster), (New York: Dover, 1959).

[8] Fife, D. W. and Rosenberg, R. S., "Queueing in a Memory Shared Computer," Proc. of the 19th National Conf. of the ACM, Philadelphia, 1964.

[9] Foley, J. D., "A Markovian Model of The University of Michigan Executive System," Comm. of the ACM, Vol. 10, No. 9, September 1967.

[10] Foster, F. G., "On the Stochastic Matrices Associated with Certain Queueing Processes," Ann. Math. Stat., Vol. 24, 1953, pp. 355-360.

[11] Gantmacher, F. R., The Theory of Matrices, Vol. 1, (New York: Chelsea, 1959).

[12] Gantmacher, F. R., Applications of the Theory of Matrices (Transl. by J. L. Brenner), (London: Intersciences, 1959).

[13] Householder, A. S., Principles of Numerical Analysis, (New York: McGraw-Hill, 1953).

[14] Kleinrock, L., "Time Shared Systems: A Theoretical Treatment," J. ACM, Vol. 14, No. 2, April 1967, pp. 242-261.

[15] Krishnamoorthi, B. and Wood, R. C., "Time Shared Computer Operations with Both Interarrival and Service Times Exponential," J. ACM, Vol. 13, No. 3, July 1966, pp. 317-338.

[16] Miller, R. G., "Stationary Equations in Continuous Time Markov Chains," Technical Report No. 80, Applied Math. and Stat. Labs, Stanford University, Stanford, California.

[17] Parzen, E., Stochastic Processes, (San Francisco: Holden-Day, 1962).

[18] Saaty, T. L., and J. Bram, Nonlinear Mathematics, (New York: McGraw-Hill, 1964).

[19] Smith, J. L., "An Analysis of Time-Sharing Computer Systems Using Markov Models," Proc. AFIPS 1966 Spring Joint Computer Conference, Vol. 28, pp. 97-104.

[20] Taussky, O., "A Recurring Theorem on Determinants," Amer. Math. Monthly, Vol. 56, pp. 672-676.

[21] Varga, R. S., Matrix Interactive Analysis, (Englewood Cliffs, New York: Prentice-Hall, 1962).

[22] Vulikh, B. Z., Introduction to Functional Analysis, (Transl. by I. N. Sneddon), (Oxford: Pergamon Press, 1963).

[23] Wallace, V. L. and Rosenberg, R. S., "RQA-1, The Recursive Queue Analyzer," Technical Report No. 2, Systems Engineering Laboratory, The University of Michigan, Ann Arbor (1966).

[24] Wallace, V. L. and Mason, D. L., "The Optimal Degree of Multiprogramming in Page-on-Demand Computer Systems," Comm. of the ACM, Vol. 12, No. 5, May 1969.