

Analysis of Large Markovian Models by Parts. Applications to Queueing Network Models.

P.-J. Courtois

Philips Research Laboratory
Av. Van Becelaere 2 - Box 8
B-1170 Brussels - Belgium

In. Messung, Modellierung und Bewertung
von Rechenystemen. GI/NTH-Fachtung, Dortmund. October
1985.
Springer Verlag.

1. Introduction

The Markov chain remains one of the most useful and flexible techniques for the evaluation of the performances and of the reliability characteristics of models of communication and computer systems.

Often, however, in order to attain a sufficient level of realism in the model, a large number of states need to be taken into account. Models with 10.000 states and more are not unusual (see e.g. [7]). The definition, construction and manipulation of the corresponding large transition matrix may then be far from easy.

These large state spaces are especially frustrating as one is often interested in characteristics of the system that are primarily related to the behavior of a subchain only, and that depend only indirectly on the other states of the model. It becomes interesting in this case to restrict the analysis to the subchain of interest, even at the cost of an approximation, provided that the accuracy of this isolated analysis remains known and tolerable.

This paper outlines the principles of a simple method suggested by my colleague P. Semal and myself (see [2,3] for more details) to alleviate these difficulties. The method consists essentially in computing lower and upper bounds on the visiting rates and on the equilibrium state probabilities of a subset of states of a chain when the submatrix of transition probabilities between the states of that subset only is accessible. These bounds, based on recent results [3] in Linear Algebra, have been shown to be the tightest ones that exist in that case; a more intuitive presentation of their properties will be given here in terms of probability arguments.

The method should be especially useful in the steady state analysis of large or infinite matrices that can have many identical diagonal submatrices, such as those encountered in queueing models. To illustrate this point we will use the method to compute bounds on the probabilities of blocking in a network of finite capacity queues.

2. Bounds on Steady State Probabilities

Let us consider a Markov process described by the following matrix of transition probabilities

$$\begin{pmatrix} [\mathbf{Q}]_{n \times n} & [\mathbf{E}]_{n \times m} \\ [\mathbf{F}]_{m \times n} & [\mathbf{G}]_{m \times m} \end{pmatrix}, \quad (2.1)$$

where the states of interest are supposed to be collected in the aggregate S the transition probabilities of which are given by the submatrix \mathbf{Q} . We are interested in that part of the equilibrium probability vector of the Markov chain which corresponds to the states of S . If \mathbf{Q} is irreducible, this equilibrium subvector, say ν , is also solution of the matrix equation (see [6] for instance):

$$\nu^T = \nu^T [\mathbf{Q} + \mathbf{E}(\mathbf{I} - \mathbf{G})^{-1} \mathbf{F}] \stackrel{\Delta}{=} \nu^T \bar{\mathbf{Q}}, \quad (2.2)$$

where $\bar{\mathbf{Q}}$ is stochastic and $\mathbf{E}(\mathbf{I} - \mathbf{G})^{-1} \mathbf{F} \geq \mathbf{0}$, so that $\bar{\mathbf{Q}} \geq \mathbf{Q}$, elementwise.

If only the block \mathbf{Q} is available, and not \mathbf{E} , \mathbf{G} , and \mathbf{F} , equation (2.2) cannot be used to evaluate ν . But lower and upper bounds on ν can be obtained in the following way.

It follows from a theorem in [3] that the Perron-Frobenius vector ν of any non-negative matrix $\bar{\mathbf{Q}}$ which is of the same size as a non-negative irreducible matrix \mathbf{Q} , and which is element wise equal or larger than \mathbf{Q} , i.e. ($\bar{\mathbf{Q}} \geq \mathbf{Q}$), is a convex linear combination of the normalized rows of $(\mathbf{I} - \mathbf{Q})^{-1}$; more precisely:

$$\mathbf{M} \stackrel{\Delta}{=} (\mathbf{I} - \mathbf{Q})^{-1}, \quad (2.3)$$

and let \mathbf{Z} be the row normalized version of \mathbf{M} :

$$\mathbf{Z}_{ij} \stackrel{\Delta}{=} \frac{\mathbf{M}_{ij}}{\sum_k \mathbf{M}_{ik}}, \quad 1 \leq i, j \leq n; \quad (2.4)$$

Since $\bar{\mathbf{Q}}$, like \mathbf{Q} , is irreducible, ν is unique up to a multiplicative constant and ν is a convex linear combination of the rows of \mathbf{Z} , i.e.:

$$\exists \beta^T \in \mathbf{R}_n^+, \quad \beta^T \mathbf{1} = 1 : \quad \nu^T = \beta^T \mathbf{Z}, \quad (2.5)$$

where $\mathbf{1}$ is a column vector of 1, and where \mathbf{Z} is defined by (2.4). An immediate consequence of this convex linear combination is that the component of the equilibrium probability vector ν are bounded by

$$(\nu^{inf})_j \stackrel{\Delta}{=} \min_i \mathbf{Z}_{ij} \leq \nu_j \leq \max_i \mathbf{Z}_{ij} \stackrel{\Delta}{=} (\nu^{sup})_j. \quad (2.6)$$

3. Probabilistic Interpretation.

These bounds may receive a probabilistic interpretation[4]. Consider the absorbing Markov chain associated with matrix \mathbf{Q} , and obtained by replacing \mathbf{G} by one absorbing state:

$$\begin{pmatrix} [\mathbf{Q}]_{n \times n} & [\mathbf{E}\mathbf{1}]_{n \times 1} \\ [\mathbf{0}]_{1 \times n} & [1]_{1 \times 1} \end{pmatrix} \quad (3.1)$$

For any pair (i, j) of states of S , let \mathbf{M}_{ij}^* denote the average number of times that the transient process defined by \mathbf{Q} , started in state i , is in state j before being absorbed by the $(n+1)^{th}$ absorbing state. By definition, these quantities satisfy the following recurrence equations :

$$\mathbf{M}_{ij}^* = \delta_{ij} + \sum_k \mathbf{Q}_{ik} \mathbf{M}_{kj}^* , \quad (3.2)$$

which can be rewritten in matrix form as :

$$\mathbf{M}^* = \mathbf{I} + \mathbf{Q} \mathbf{M}^* = (\mathbf{I} - \mathbf{Q})^{-1} . \quad (3.3)$$

The matrix \mathbf{M}^* is thus identical to the matrix \mathbf{M} defined in (2.3), and the element (i, j) of the matrix \mathbf{Z} , defined in (2.4),

$$\mathbf{Z}_{ij} \stackrel{\Delta}{=} \frac{\mathbf{M}_{ij}}{\sum_k \mathbf{M}_{ik}} = \frac{\mathbf{M}_{ij}}{\sigma_i} , \quad (3.4)$$

is the relative rate of visit to state j , when the process defined by \mathbf{Q} is started in state i . The sum σ_i is the average total number of transitions before absorption when \mathbf{Q} is started in state i ; and the row vector

$$\mathbf{Z}_i \stackrel{\Delta}{=} (\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, \dots, \mathbf{Z}_{in}) . \quad (3.5)$$

is the vector of relative visit rates to the states of S when the process is started in state i .

The bounds (2.6) have therefore the following meaning. The steady state conditional probability of being in a state j , given that j belongs to a subspace S , is comprised between the maximum and the minimum relative visit rate to state j before absorption when every state of S is considered as a possible starting state.

The inequalities (2.6), together with the meaning attached to (3.5), have several other interesting consequences. Assume that in the decomposition (2.1), the matrix \mathbf{F} has only one non-null column, say the i^{th} one. Then, the only state by which subset S can be entered from outside is state i . In this case, the vector \mathbf{Z}_i is precisely equal to the vector ν . More generally : if matrix \mathbf{F} has two non-zero columns, say columns i and j , then the solution ν will be a convex combination of \mathbf{Z}_i and \mathbf{Z}_j ; and if all the columns of matrix \mathbf{F} are non-null then ν will be a convex combination of all the \mathbf{Z}_i , $i=1, \dots, n$. A proof of this property is given in [2].

We have also proved in [2] that the relative visit rate to a state j is maximum when the process is started from that state j . Inequalities (2.6) can therefore be rewritten as :

$$(\nu^{inf})_j = \min_i (\mathbf{Z}_{ij}) \leq \nu_j \leq \max_i (\mathbf{Z}_{ij}) = \mathbf{Z}_{jj} = (\nu^{sup})_j . \quad (3.6)$$

Note that the state i which minimizes the relative visit rate to a state j cannot in general be characterized more precisely; in each case, it will depend on the relative values of the visit numbers \mathbf{M}_{ij} and on the absorption times σ_i . However, as described in the following section, more precise characterizations can sometimes be obtained.

Finally, it is proven in [2,3] that these bounds are the best ones that can be obtained from the submatrix \mathbf{Q} in the sense that a *stochastic* matrix $\bar{\mathbf{Q}}$, $\bar{\mathbf{Q}} \geq \mathbf{Q}$, with an equilibrium state vector which attains these bounds always exists. No conditions on the matrices \mathbf{E} , \mathbf{F}

and \mathbf{G} are required for the existence of the bounds. However, the smaller the elements of \mathbf{E} are, the closer $\bar{\mathbf{Q}}$ is to \mathbf{Q} and, of course, the tighter the bounds are. A detailed analysis of their accuracy can be found in [2]; some essential aspects will be mentioned at the end of the next section.

4. An Example : Two Coupled Finite Queues with Coxian Service Distributions.

A major application of the bounds (2.6) is the analysis, with known accuracy, of the behavior of a subsystem in isolation of the remainder of the system of which it is a part.

4.1. The Model

A simple but rather typical example which can illustrate this type of application is provided by the subsystem shown on Figure 1.

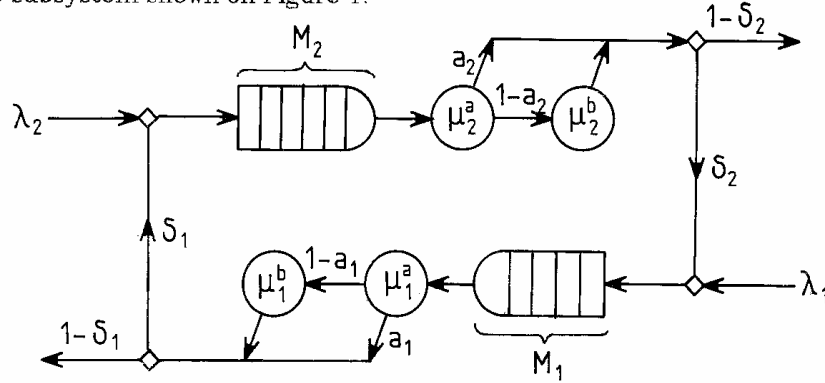


Figure 1.

This subsystem consists of two finite queues, each with a two-stage coxian server. The total number of customers in service or in queue at each server is at most M_1 and M_2 respectively. Customers arrive from the rest of the network with Poisson arrivals λ_1 and λ_2 ; upon service completion, they join the other queue with probability δ_1 or δ_2 , or leave the subsystem with probabilities $(1-\delta_1)$, $(1-\delta_2)$. This model can be viewed as a simplified version of a message switching node with finite capacity buffers, and with two input and two output communication lines.

The state of the subsystem is defined by the quadruple (i_1, s_1, i_2, s_2) , where i_j ($i_j \leq M_j$) is the number of customers at server j , and s_j ($s_j = a$ or b) is the stage of the customer in service at server j , $j = 1, 2$.

If all states with same value (i_2, s_2) for server 2 are grouped together and if these groups of states are arranged in the order $(i_2, s_2) = (0), (1a), (1b), (2a), (2b), \dots, (M_2b)$, the matrix of transition probabilities has the regular structure shown on Figure 2.

The diagonal of \mathbf{D} consists of $(2M_2 + 1)$ blocks each of which comprises all the transitions at server 1 for a given state (i_2, s_2) . These diagonal blocks $\mathbf{D}(0)$, $\mathbf{D}(1a)$, $\mathbf{D}(1b)$, $\mathbf{D}(2a)$, \dots are all identical except for their main diagonal elements (\times) which are the complements to one of the corresponding rowsums of the off-diagonal elements in \mathbf{D} ; each block has size $(2M_1 + 1)$ and an internal structure very similar to that of \mathbf{D} (see Fig. 3).

$D(0)$	U	0				
L_a	D_a	C	U	0		
L_b	0	D_b	0	U		
	L_a	0	D_a	C	U	0
	L_b	0	0	D_b	0	U
			\dots		\dots	\dots
			\dots		\dots	\dots
					L_a	0
					L_b	0
					$D_a(M_2)$	C
					0	$D_b(M_2)$

Figure 2. The System Matrix D .

x	λ_1	0				
z_1^a	x	c_1	λ_1	0		
z_1^b	0	x	0	λ_1		
	z_1^a	0	x	c_1	λ_1	0
	z_1^b	0	0	x	0	λ_1
			\dots		\dots	\dots
			\dots		\dots	\dots
					z_1^a	0
					z_1^b	0
					x	c_1
					0	x

Figure 3. A Diagonal Block $D(0)$, or $D(a)$, or $D(b)$, \dots in D .

The transition rate between the two coxian stages at server 1 is $c_1 = \mu_1^a(1 - a_1)$; $z_1^a = \mu_1^a a_1(1 - \delta_1)$ and $z_1^b = \mu_1^b(1 - \delta_1)$ are the probabilities of a departure from the subsystem at these stages.

The off-diagonal blocks L_a correspond to a service completion at the stage a of server 2, with $z_2^a = \mu_2^a(1 - \delta_2)$ and $h_2^a = \delta_2 \mu_2^a a_2$ (see figure 4); the blocks L_b are identical with z_2^b and h_2^b elements instead.

The blocks U correspond to an arrival at server 2 and are independent of the stage of that server (see figure 5).

The blocks C in the matrix D correspond to transitions from stage a to stage b at server 2; these blocks are diagonal matrices with $c_2 = \mu_2^a(1 - a_2)$ on their diagonal.

4.2. Blocking Probabilities.

Suppose that we are interested in estimating the steady-state probability of a customer being rejected at queue 1 when this queue is full. Without solving the entire matrix D , bounds on this probability can be derived from bounds on marginal probabilities in the following way.

z_2^a	h_2^a	0			
	z_2^a		h_2^a		
		z_2^a		h_2^a	
			z_2^a		h_2^a
			
					...
					z_2^a
					z_2^a

Figure 4. A Lower Block L_m , ($z_2^a = \mu_2^a(1-\delta_2)$, $h_2^a = \delta_2 \mu_2^a a_2$)

λ_2					
h_1^a	λ_2				
h_1^b	0	λ_2			
	h_1^a	0	λ_2	0	
	h_1^b	0	0	λ_2	
			
			
				h_1^a	0
				h_1^b	0
				λ_2	0
				0	λ_2

Figure 5. An Upper Block U , ($h_1^a = \delta_1 \mu_1^a a_1$, $h_1^b = \delta_1 \mu_1^b a_1$)

The probability of rejection at queue 1 is equal to :

$$[h_2^a \times Prob(i_1=M_1, i_2 \neq 0, s_2=a)] + [h_2^b \times Prob(i_1=M_1, i_2 \neq 0, s_2=b)] + [\lambda_1 \times Prob(i_1=M_1)] .$$

We will show how bounds can be obtained on the first of these three terms; the procedure is quite similar for the two other terms.

Let $X_j^{(2)} \stackrel{\Delta}{=} Prob(i_2=j, s_2=a \text{ or } b)$, and let also

$$\nu_1(k, m, n | j) \stackrel{\Delta}{=} Prob(i_1=k, s_1=m, s_2=n | i_2=j), \quad m, n, = a, b ,$$

where the variables s_1, s_2 are missing when $i_1, i_2=0$ respectively.

According to the results presented in Section 2, upper and lower bounds on $\nu_1(k, m, n | j)$ are provided by the maximum and the minimum element of the column corresponding to state $(i_1=k, s_1=m, i_2=j, s_2=n)$ of the normalized inverse of the diagonal block of $(I - D)$ which corresponds to states $(i_2=j)$.

If $\bar{\nu}_1(k, m, n | j)$ and $\underline{\nu}_1(k, m, n | j)$ are these upper and lower bounds, we have

$$\text{Prob}(i_1 = 1, i_2 \neq 0, s_2 = a) \leq \sum_{j=1}^{M_2} X_j^{(2)} \sum_{s_1=a,b} \bar{\nu}_1(M_1, s_1, a | j).$$

For $j = 1, \dots, M-1$, all $\bar{\nu}_1(M_1, s, a | j)$ are equal since the corresponding diagonal blocks of \mathbf{D} are identical. Thus

$$\text{Prob}(i_1 = M_1, i_2 \neq 0, s_2 = a) \leq (1 - X_0^{(2)}) \max_{j=(M_2-1), M_2} \left(\sum_{s_1=a,b} \bar{\nu}_1(M_1, s_1, a | j) \right),$$

while a lower bound on $X_0^{(2)}$ is given by

$$X_0^{(2)} \geq \sum_{j=0}^{M_1} X_j^{(1)} \sum_{s_1=a,b} \underline{\nu}_2(0, s_1 | j) \geq \min_{j=0,1,M_1} \left(\sum_{s_1=a,b} \underline{\nu}_2(0, s_1 | j) \right).$$

In the same way, a lower bound is given by

$$\begin{aligned} \text{Prob}(i_1 = M_1, i_2 \neq 0, s_2 = a) &\geq \sum_{j=1}^{M_2} X_j^{(2)} \sum_{s_1=a,b} \underline{\nu}_1(M_1, s_1, a | j), \\ &\geq (1 - X_0^{(2)}) \min_{j=1, M_1} \left(\sum_{s_1=a,b} \underline{\nu}_1(M_1, s_1, a | j) \right), \end{aligned}$$

with

$$X_0^{(2)} \leq \max_{j=0,1,M_1} \left(\sum_{s_1=a,b} \bar{\nu}_2(0, s_1 | j) \right).$$

Upper and lower bounds on the marginal probabilities ν_i are obtained from the normalized rows of inverses of the form

$$\begin{bmatrix} \mathbf{I} - \mathbf{D}_a & -\mathbf{C} \\ \mathbf{0} & \mathbf{I} - \mathbf{D}_b \end{bmatrix}^{-1}; \quad (4.1)$$

such inverses reduce to

$$\begin{bmatrix} (\mathbf{I} - \mathbf{D}_a)^{-1} & \mathbf{B} \\ \mathbf{0} & (\mathbf{I} - \mathbf{D}_b)^{-1} \end{bmatrix}, \quad (4.2)$$

whith $\mathbf{B} = (\mathbf{I} - \mathbf{D}_a)^{-1} \mathbf{C} (\mathbf{I} - \mathbf{D}_b)^{-1}$, since \mathbf{C} are diagonal matrices. Thus, the computation complexity for bounding the probability of rejection at queue 1 reduces in this example to the computation of a few inverses of order $(M_2 + 1)$ or $(M_1 + 1)$; their total number is actually five here, three of order (M_2+1) and two of order (M_1+1) .

4.3. Accuracy.

The quality of the bounds depends mainly on three factors: (1) the existence of comparatively small values for the probabilities of leaving the subsystem state space, (2) the distribution over the subsystem states of the probabilities of return to the subsystem, and (3) the similarity of the subsystem visit rates to any given state of its subspace. Let us discuss the influence of these three factors in the context of our example.

As stated earlier, the smaller the elements outside the diagonal blocks are, the tighter the bounds will be. In this example the accuracy will thus be better when \mathbf{L}_a , \mathbf{L}_b and \mathbf{U} have comparatively small elements, and especially small non-diagonal elements. Indeed, the diagonal elements in these blocks correspond to transitions which leave the subsystem

under analysis in the same marginal state, and have thereby less influence on the variations of the marginal equilibrium vector from block to block. The accuracy of the bounds will therefore depend primarily in this example on ∂_1 , ∂_2 , μ_2 and λ_2 being small compared to the other parameters.

The second factor affects the tightness of the bounds in this example through the special tridiagonal block structure of the whole matrix \mathbf{D} . Upon leaving a subsystem through a state belonging to a subset \mathbf{D}_a or \mathbf{D}_b , the probability of returning to a state of \mathbf{D}_a is much higher than the probability of returning to a state of \mathbf{D}_b . The computation of the bounds which is based on the assumption that all states of the subsystem may have probability one of first return does not take that special structure of the entire matrix into account. Fortunately such block structures often makes possible to obtain a better lower bound matrix for $\bar{\mathbf{Q}}$ in equation (2.2). A first approximation of $\mathbf{E}(\#I - \mathbf{G})^{-1}\mathbf{F}$ can be easily computed. If the entire matrix has the form

$$\begin{array}{cccccc} \mathbf{L}_{i-1} & \mathbf{Q}_{i-1} & \mathbf{U}_{i-1} & & & \\ & \mathbf{L}_i & \mathbf{Q}_i & \mathbf{U}_i & & \\ & & \mathbf{L}_{i+1} & \mathbf{Q}_{i+1} & \mathbf{U}_{i+1} & \end{array}$$

a lower bound matrix for the exact stochastic matrix $\bar{\mathbf{Q}}_i$ which is defined by (2.2) and which corresponds to \mathbf{Q}_i , is given by :

$$\mathbf{Q}_i + \mathbf{U}_i (\mathbf{I} - \mathbf{Q}_{i+1})^{-1} \mathbf{L}_{i+1} + \mathbf{L}_i (\mathbf{I} - \mathbf{Q}_{i-1})^{-1} \mathbf{U}_{i-1} \quad (4.3)$$

This technique is applicable in the example above, but the normalized inverses do not keep the structure (4.1) and, therefore, cannot be reduced to (4.2).

As for the last factor, it is clear that the bounds on the equilibrium probability to a given state of a subsystem will be far apart if the visit rates to that state before absorption differ much depending on the state of departure. As explained in [2], this is typical of submatrices with a poor separation of the modules of the dominant eigenvalues, like, for instance, reducible or nearly reducible matrices. For the same reason, as we shall see in the next section, tridiagonal matrices may also lead to visit rates which differ substantially from one another.

5. Concluding Remarks

This example shows how bounds on particular performance measures can be obtained at lower computational cost by restricting the analysis to the state subspace of interest.

The discussion in Section (4.3) reveals, however, that matrices representing queueing networks do not have necessarily the best characteristics to guarantee tight bounds. But, as shown by some numerical examples in [5], the accuracy may often be sufficient in practice. Moreover, it should be possible to exploit more thoroughly the possibilities mentioned in Section (4.3) to construct better lower bound matrices for the diagonal blocks of certain typical queueing network matrices.

The bounds are sometimes easy to compute. The case when the submatrix \mathbf{Q} in (2.1) is tridiagonal is of special interest because many queueing submodels involve such matrices. It is possible in this case [5] to obtain a formal definition of the elements of \mathbf{Z} , the normalized version of $(\mathbf{I} - \mathbf{Q})^{-1}$, as well as recurrence relations to compute these elements. Moreover, it is then proved that in every column of \mathbf{Z} , the largest element is always on the diagonal and the smallest one is always either on the first or last row. This result yields a more precise formulation of the bounds since (2.6) become in this case

$$\min(\mathbf{Z}_{1j}, \mathbf{Z}_{nj}) \leq \nu_j \leq \mathbf{Z}_{jj} .$$

The intuitive meaning of this result is clear. In a tridiagonal Markovian process, a state is reachable from another state only by passing through all "intermediary" states. The relative visit rate to a state is thus minimum when starting from the "most distant" state which, in this case, is either state 1 or state n .

When the tridiagonal submatrix reduces to the model of a random walk with two barrier states, it is even possible to express the bounds as explicit functions of the size of the submatrix and of its upper and lower diagonal.

Since the bounds presented here are the tightest ones that can be derived from a given transition submatrix corresponding to a subsystem, and since many subsystems have structural properties that can be exploited to compute these bounds efficiently or even formally, the method is likely to have much practical interest.

6. Acknowledgments

I am grateful to my colleague P. Semal whose remarks and help contributed greatly to improve the presentation of this paper.

7. References

- [1] Berman, A., and Plemmons, R.J., *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, **1979**, 132-142.
- [2] Courtois, P.-J., and Semal, P., *Bounds for the Positive Eigenvectors of Nonnegative Matrices and for Their Approximations by Decomposition*, J. Assoc. Comput. Mach., **31**, **1984**, 804-825.
- [3] Courtois, P.-J., and Semal, P., *On Polyhedra of Perron-Frobenius Eigenvectors*, Linear Algebra and its Applications, **65**, **1985**, 157-170.
- [4] Courtois, P.-J., and Semal, P., *Block Decomposition and Iteration in Stochastic Matrices*, Philips Journal of Research, **39**, **1984**, 178-194.

- [5] Courtois, P.-J., and Semal, P., *Analysis of large Markovian Models by Parts. Error Bounds and Applications*. Manuscript M 109, Philips Research Laboratory, Brussels **May 1985**.
- [6] Kemeny, J.G., and Snell, J.L., *Finite Markov Chain*, D. Van Nostrand Co., Inc., New York, **1960**, 150-152.
- [7] Wunderlich, E.F., Kaufman, L., Gopinath, B., *The Control of Store and Forward Congestion in Packet Switching Networks*, Proc. 5th Int. Conference on Computer Communication, Atlanta, **1980**.