in IEEE Transactions on Computers, 50(5):399-413, June 2001

Bound computation of dependability and performance measures

S. Mahévas
Irisa
Stephanie.Mahevas@ifremer.fr

G. Rubino
Irisa and ENST-Bretagne
rubino@{irisa,rennes.enst-bretagne}.fr

Irisa, Campus de Beaulieu
35042 Rennes Cedex, France

January 31, 2001

Abstract

We propose a new method to obtain bounds of dependability, performance or performability measures concerning complex systems modeled by a large Markov model. Its extends previous published techniques mainly designed to the analysis of dependability measures only, and working under more restrictive conditions. Our approach allows to obtain tight bounds of performance measures on certain cases, and in particular, on models having an infinite state space. We illustrate the method with some analytically intractable open queuing networks, as well as with large dependability models.

Index Terms — Dependability evaluation, performance evaluation, Markov chains, numerical analysis, bounding techniques.

1 Introduction

To derive performance, dependability or performability measures from a model of a complex system, Markov chains, under different forms, are the most widely used mathematical tools. Sometimes, the user directly builds a Markov chain from system specifications. Most often, the model is described in a higher level language such as queues or networks of queues, stochastic Petri nets, etc., and some tool constructs the stochastic process automatically. The usefulness of Markov chains is due to the power of the theory and to the efficient algorithmic technology associated with. However, such a power has a price. There are two major drawbacks when using Markov models. The first one is the fact that, to be able to represent the more and more complex systems built nowadays, the state spaces are larger and larger. In general, models having, say, hundreds or thousands of millions of states are out of scope for exact numerical analysis. The second one is the so-called “rare events” problem, meaning that in many cases (typically in models of highly available systems), the interesting events (typically, the fact that the system is down) have very low probabilities, making problematic the use of Monte Carlo techniques. This is usually related to high numerical values of the ratios between different transition rates of the chain, which leads also to numerical problems in the exact analysis of the models (stiffness). Often, the two problems appear simultaneously. To deal with them, a different approach exists. It consists of computing bounds of the desired measures instead of exact values or statistical estimations [1], [2], [3]. This is the subject of the paper.

Bounding techniques have been mainly developed for cases in which, at the same time, the model is large and stiff. The second aspect is in fact used to help with dealing with the first one, as we see in the paper. The intuition behind the approach is quite simple. When the model is stiff, the stochastic process spends most of its time in a (small) part of the state space. It is then a natural attempt to try to approximate the interesting measures by working basically with that part of the state space. This is done by replacing, in some way, the rest of the state space by “a few” states. All the difficulty is how to handle that large part of the state space of the model in order to have some control on the accuracy of the method.

The system is represented by a continuous time homogeneous and irreducible Markov chain X over the finite state space S with stationary distribution $\boldsymbol{\pi}$ (row vector). We denote $\pi_i = \Pr(X_\infty = i)$ where X_∞ is a stationary version of X. With each state i we associate a reward $r_i \geq 0$; $\mathbf{r}$ is the (row) vector of rewards. This paper deals with the computation of

$$R = \mathrm{E}(r_{X_\infty}) = \sum_{i \in S} r_i \pi_i = \boldsymbol{\pi}\mathbf{r}^{\mathrm{T}}.$$

For instance, in a dependability context, the states are in general called “operational” when they represent a system delivering its service as expected, and “unoperational” otherwise. If a reward equal to 1 is associated with the operational states and equal to 0 with the unoperational ones, then R is the asymptotic availability of the system. In a performance context, suppose that the model is a queuing network and that you are interested in the mean number of customers in queue q. If with each state i we associate a reward equal to the number of customers in station q when the model is in state i, the expectation R is equal to the desired mean number of customers in q. As a second example, if r_i is the number of active processors in some model of a fault-tolerant multi-processor computer when its state is i, then R is a performability measure (the mean *power* of the system).

In this paper we develop an approach that avoids an important drawback of previous works, by extending the class of evaluable pairs (models, measures). The main restriction of previous techniques, which mainly concern dependability models, is that "almost" every state must have at least one transition corresponding to a repair. The approach proposed here works without this condition, but it has a price: in some cases, large linear systems must still be solved. In the paper we illustrate the method with cases where these systems are easy to solve (and with models that are such that previous techniques do not apply). In a performance context, it is frequent to work with infinite state spaces. The techniques we propose can, in some situations, deal with these cases as well. This is also developed and illustrated here.

The paper is organized as follows. Next section states the context, defines some general notation and recall basic previous results on bounds. In Section 3 we recall important fact about state aggregation in Markov chains. The objective of Section 4 is to explain the method of [3]. Section 5 presents our technique together with some needed supplementary results. Section 6 explains how we deal with models having infinite state spaces. In Section 7 we give examples, both in the dependability and in the performance areas, both in the finite and the infinite cases. Section 8 concludes the paper.

2 Preliminaries

Generalities. We are given a finite and irreducible continuous time homogeneous Markov chain X over the state space S, presumed to model some complex system. We denote by A the infinitesimal generator of X. The asymptotic distribution of X is denoted by $\boldsymbol{\pi}$ and it will be seen in the paper as a row vector. We then have $\boldsymbol{\pi} A = \mathbf{0}$.

We are also given a vector of positive reals, $\mathbf{r} = (\ldots r_i \ldots)$ over S. The goal is the evaluation of bounds of $R = \boldsymbol{\pi}\mathbf{r}^{\mathrm{T}}$, *without computing vector* $\boldsymbol{\pi}$. To do this, we assume that we know a lower and an upper bound of the rewards, that is, two reals ϱ_1 and ϱ_2 such that for all $i \in S$,

$$0 \le \varrho_1 \le r_i \le \varrho_2 < \infty.$$

In Section 6 we develop an extension to the case of $|S| = \infty$ and $\varrho_2 = \infty$.

The state space S is assumed to be decomposed (or decomposable) into two disjoint sets, denoted by G and $\bar{G}$. The idea is that the states in G are or include those frequently visited by the chain in equilibrium. In many situations, the user is able to approximatively identify these sets. For instance, in a dependability context, a model corresponding to a repairable system with high reliable components leads to choose G as the set of states having less than some fixed number of failed units. In a queuing model in a light traffic situation, one can associate G with the states where the network has less than some fixed number of customers. The techniques discussed here attempt to give bounds of the asymptotic reward R by working with auxiliary Markov obtained by replacing $\bar{G}$ (and the associated rates) with a "few" states. The good situation is then to have $|G| \ll |S|$.

Some notation associated with a subset of states. For any subset of states $C \subseteq S$, we denote

- $\boldsymbol{\pi}_C$, the restriction of $\boldsymbol{\pi}$ to the set C (row vector with size equal to $|C|$), $\mathbf{r}_C$, the restriction of $\mathbf{r}$ to C, etc.,
- $\pi(C) = \Pr(X_\infty \in C) = \sum_{i \in C} \pi_i = \boldsymbol{\pi}_C \mathbf{1}^{\mathrm{T}}$, where $\mathbf{1}^{\mathrm{T}}$ is the transpose of a row vector having all its entries equal to 1, the dimension being defined by the context,
- $\widehat{\boldsymbol{\pi}}_C$ = distribution of X_∞ conditioned to the event $\{X_\infty \in C\}$, (row) vector with size $|C|$, that is, $\widehat{\boldsymbol{\pi}}_C = \frac{1}{\pi(C)} \boldsymbol{\pi}_C$,
- $\bar{C}$ = the complement $S - C$ of C,
- $in(C) = \{j \in C$ such that there exists $i \in \bar{C}$ with $A_{i,j} > 0\}$,
- $out(C) = \{i \in C$ s.t. there exists $j \in \bar{C}$ with $A_{i,j} > 0\}$,
- A_C, the block of A corresponding to the transitions inside subset C,
- $A_{C,C'}$, the block of A corresponding to the transitions from subset C to subset $C^{'}$.

Forcing the entries in G by a fixed state j. The main idea comes from the basic initial work by Courtois and Semal. It consists of building a family of Markov chains derived from X in the following way. For each state $j \in in(G)$, let us construct the new continuous time homogeneous Markov chain $X^{(j)}$, by forcing the transitions from $\bar{G}$ into G to enter by state j, as illustrated in Figure 1. The infinitesimal generator of $X^{(j)}$ is denoted by $A^{(j)}$:

$$A^{(j)} = \left(\begin{array}{cc} A_G & A_{G\bar{G}} \\ A^{(j)}_{\bar{G}G} & A_{\bar{G}} \end{array} \right).$$

The transition rate from any state $i \in \bar{G}$ to j is equal to $\sum_{l \in in(G)} A_{i,l}$. In other words, $A^{(j)}_{\bar{G}G} = A_{\bar{G}G}\mathbf{1}^{\mathrm{T}}\boldsymbol{e_j}$, where $\boldsymbol{e_j}$ is the jth row vector of the canonical base in $\mathbb{R}^{|G|}$. The other transition rates of $X^{(j)}$, that is inside G, from G to $\bar{G}$ and inside $\bar{G}$, are as in X. First, we prove that $X^{(j)}$ has an unique stationary distribution by means of the following lemma.

Lemma 1 *The Markov chain $X^{(j)}$ has an unique recurrent class; this class includes state j.*

Proof. Denote by S_j the class of states i such that i is reachable from j and j is reachable from i. The claim implies that S_j can be reached from every state $k \in S$. Assume $S_j \neq S$ and let $k \in S - S_j$. If $k \in out(\bar{G})$, by definition of $out()$, k is connected to j. If $k \in \bar{G} - out(\bar{G})$ then there is necessarily a path from k to some $l \in out(\bar{G})$ (since X is irreducible), which is completely included in $\bar{G}$, and by definition of $X^{(j)}$, l is connected to j. It remains the case of a state $k \in G - \{j\}$. If, in X, there is a path from k to j completely included in G, we are done. If not, since X is irreducible, there is at least a path from k to j passing through $\bar{G}$, thus entering for the first time $\bar{G}$ by some state l and then, we are in the first discussed case. ■

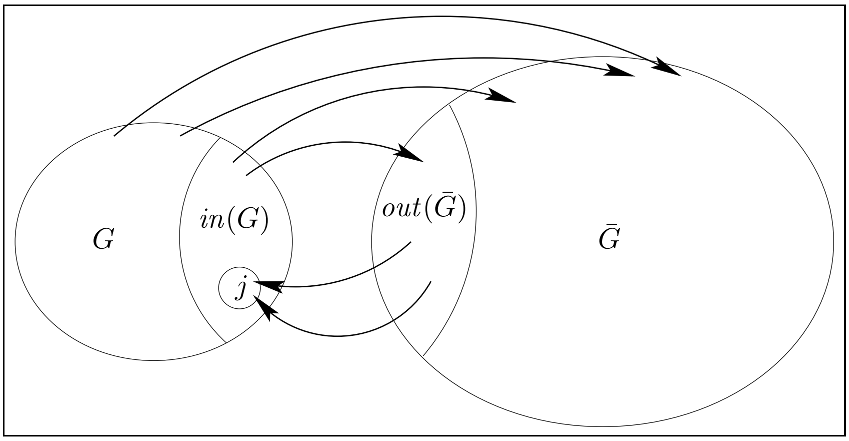

Figure 1: The topology of $X^{(j)}$

It is easy to see that there can be transient states in $X^{(j)}$. For instance, think of this simple situation: some state $i \in G$ has only one transition out to j and one transition in from $k \in \bar{G}$. In $X^{(j)}$, such a state is transient.

Since $X^{(j)}$ has an unique recurrent class, it has an unique stationary distribution which we denote by $\boldsymbol{\pi}^{(j)}$. Then, for each state $j \in in(G)$, we have the following result, which is a slight extension of results in [1]:

Theorem 1 *There exists a positive vector $\boldsymbol{\beta}$ with support $in(G)$, that is, satisfying $\beta_j = 0$ if $j \notin in(G)$, such that $\boldsymbol{\beta}\mathbf{1}^{\mathrm{T}} = 1$ and*

$$\sum_{j \in in(G)} \beta_j \boldsymbol{\pi}^{(j)} = \boldsymbol{\pi}.$$

Proof. See Appendix A.

In [1], the authors show this result under the assumption that the chains are irreducible. In the Appendix we show that the only necessary assumption is the existence of an unique stationary distribution for $X^{(j)}$. We also show that the result is still valid in the infinite state space case (used in Section 8).

From this theorem, we derive the two following immediate results, put together as a corollary.

Corollary 1 *If we denote $\pi^{(j)}(G) = P(X_\infty^{(j)} \in G)$, we have*

$$\min_{j \in in(G)} \pi^{(j)}(G) \leq \pi(G), \tag{1}$$

and if $R^{(j)} = \boldsymbol{\pi}^{(j)}\mathbf{r}^{\mathrm{T}}$, we have

$$\min_{j \in in(G)} R^{(j)} \leq R \leq \max_{j \in in(G)} R^{(j)}. \tag{2}$$

Proof. The proof is in [3], for the particular case of the asymptotic availability measure. The extension to the general asymptotic reward measure is straightforward. ■

Corollary 1 gives the relationships that will be used to derive bounds of R. In the sequel, we will develop a general approach to build a lower bound of $\min_{j \in in(G)} R^{(j)}$ and an upper bound of $\max_{j \in in(G)} R^{(j)}$.

3 Aggregation of states

To go further, we suppose that we are given a partition $\{C_I, I = 0, 1, \ldots, M\}$ of S and an integer K with $0 < K < M$, and that G is defined as

$$G = \bigcup_{I=0}^{K-1} C_I.$$

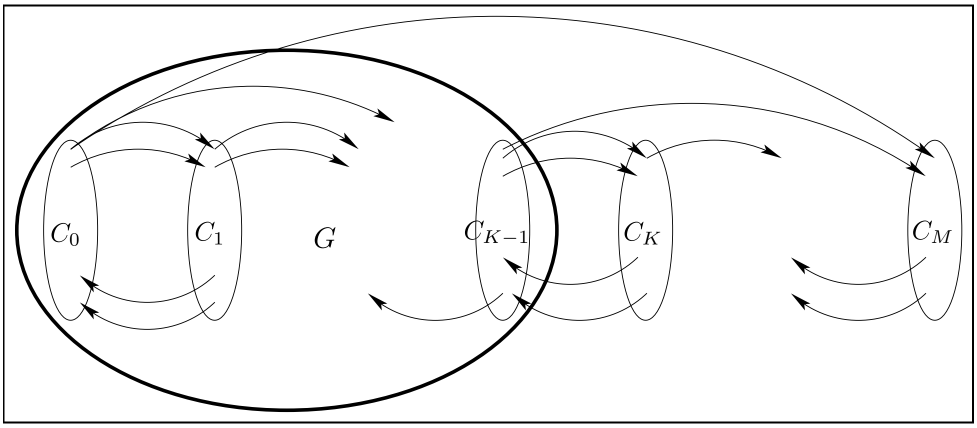

Figure 2: Chain X and the partition $\{C_I, I = 0, 1, \ldots, M\}$ of S

In a performance context, assuming that we deal with something like a queuing network or a stochastic Petri net, C_I can be, for instance, the set of states where the system, or some of its subsystems, has I customers or tokens. In a dependability case, if we work with a model of some fault-tolerant multi-component system, C_I can be, for instance, the set of states corresponding to I operational components. The good property for such a partition is that the higher the index I, the lower the probability that X_∞ belongs to the set C_I.

In what follows, we assume that transitions from class C_I to class C_J are not allowed if $J \leq I - 2$, that is, that the following condition holds:

Condition 1 *For any two integer indices I and J such that $J \leq I - 2$, for any states $i \in C_I$ and $j \in C_J$ we have $A_{i,j} = 0$.*

Observe that, given the irreducibility of X, this implies that for all index $I > 0$ there are at least two states $i \in C_I$ and $j \in C_{I-1}$ such that $A_{i,j} > 0$.

Following with the examples used a few lines before, in the case of a queuing model or a Petri net, this means, for instance, that simultaneous departures are not allowed. In the dependability example, the condition says that simultaneous repairs are not allowed.

For each state $j \in in(G)$, we will consider now the following aggregation of $X^{(j)}$. We define a continuous time homogeneous Markov chain $X^{(j)\text{agg}}$ which is constructed from $X^{(j)}$ by collapsing the classes $C_K, C_{K+1}, \ldots, C_M$ of the partition into single states $c_K, c_{K+1}, \ldots, c_M$.

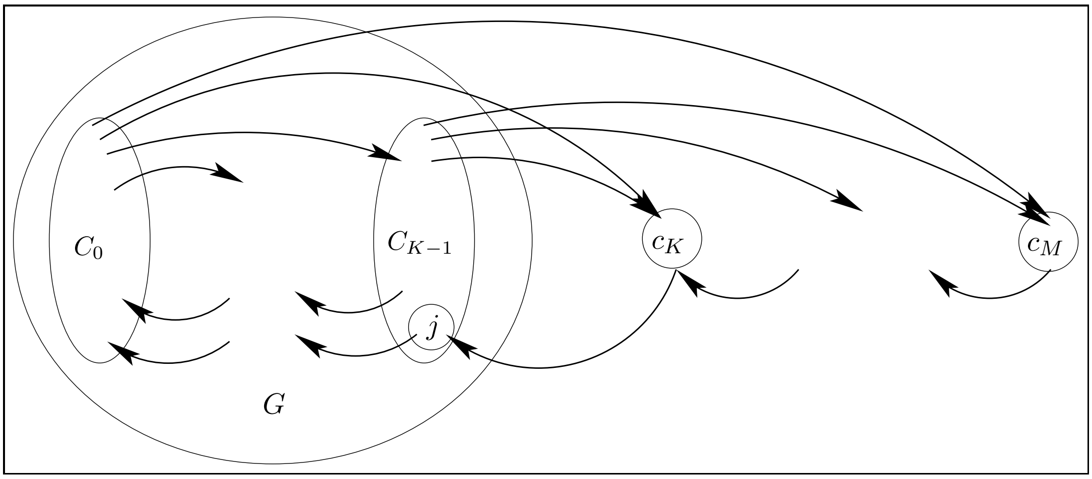

Figure 3: The topology of $X^{(j)\text{agg}}$

If we denote by $A^{(j)\text{agg}}$ the infinitesimal generator of $X^{(j)\text{agg}}$, recalling that $\pi^{(j)}(C_I)$ is the probability that $X^{(j)}_\infty$ belongs to C_I, the transition rates of $X^{(j)\text{agg}}$ are given by the following expressions:

- for all $h \in G$ (or $h \in out(G)$), and for all $I \geq K$,
$$A^{(j)\text{agg}}_{h,c_I} = \sum_{i \in C_I} A_{h,i}, \tag{3}$$
- for any $I > K$,
$$A^{(j)\text{agg}}_{c_I,c_{I-1}} = \mu^{(j)}_I = \frac{\sum_{i \in C_I} \pi^{(j)}_i \sum_{l \in C_{I-1}} A_{i,l}}{\pi^{(j)}(C_I)}, \tag{4}$$
- $$A^{(j)\text{agg}}_{c_K,j} = \mu^{(j)}_K = \frac{\sum_{i \in C_K} \pi^{(j)}_i \sum_{l \in C_{K-1}} A_{i,l}}{\pi^{(j)}(C_K)}, \tag{5}$$
- for any $I \geq K$ and $J > I$,
$$A^{(j)\text{agg}}_{c_I,c_J} = \lambda^{(j)}_{I,J} = \frac{\sum_{i \in C_I} \pi^{(j)}_i \sum_{l \in C_J} A_{i,l}}{\pi^{(j)}(C_I)}. \tag{6}$$

We denote by $S^{\text{agg}} = G \cup \{c_K, c_{K+1}, \ldots, c_M\}$ the state space of $X^{(j)\text{agg}}$. Since $X^{(j)}$ has an unique recurrent class (Lemma 1), it is immediate to see that $X^{(j)\text{agg}}$ has also an unique recurrent class including j. Let us denote $\pi_i^{(j)\mathrm{agg}} = \Pr(X_\infty^{(j)\mathrm{agg}} = i)$, $i \in S^{\mathrm{agg}}$, where $X_\infty^{(j)\mathrm{agg}}$ denotes a stationary version of $X^{(j)\mathrm{agg}}$. We then have the following well known result on aggregation:

$$\text{for any } g \in G,\ \pi_g^{(j)\mathrm{agg}} = \pi_g^{(j)}, \text{ and for any } I \geq K,\ \pi_{c_I}^{(j)\mathrm{agg}} = \pi^{(j)}(C_I). \tag{7}$$

The chain $X^{(j)\mathrm{agg}}$ is called in [1] "the exact aggregation of $X^{(j)}$ with respect to the given partition". We adopt here this terminology (in [4], it is called "the pseudo-aggregation of $X^{(j)}$ w.r.t. the given partition"). Of course, to build it we need the stationary distribution $\boldsymbol{\pi}^{(j)}$ of $X^{(j)}$, which is unknown. We define in the next subsection another chain which "bounds", in some way, chain $X^{(j)\mathrm{agg}}$, and from which the desired bounds of R will be computed.

4 "Bounding" the Markov chain $X^{(j)\mathrm{agg}}$

For each state $j \in in(G)$, let us define a homogeneous and irreducible Markov chain $Y^{(j)}$ over S^{agg} with the same topology as the aggregated chain $X^{(j)\mathrm{agg}}$, in the following way. We keep the same transition rates inside the subset G and from G to the c_I's, which are computable without the knowledge of the stationary distribution of $X^{(j)}$. The transitions inside the set of states $\{c_K, c_{K+1}, \ldots, c_M\}$ and from c_K to j are changed as follows: we replace the (unknown) exact aggregated rates from c_I to c_J, $I < J$, by some $\lambda_{I,J}^{+}$; we replace the (unknown) exact aggregated rates from c_I to c_{I-1}, $I > K$, by some μ_I^{-}, and the (unknown) exact aggregated rate from c_K to j, by μ_K^{-}. These modifications must satisfy

$$\text{for } K \leq I > J \leq M, \quad \lambda_{I,J}^{+} \geq \lambda_{I,J}^{(j)}, \tag{8}$$

$$\text{for } I \geq K, \quad 0 < \mu_I^{-} \leq \mu_I^{(j)}. \tag{9}$$

Then, between the three chains $X^{(j)}$, $X^{(j)\mathrm{agg}}$ and $Y^{(j)}$, the following relation holds.

Theorem 2 *If we denote by $\mathbf{y}^{(j)}$ the stationary distribution of $Y^{(j)}$, we have*

$$\widehat{\boldsymbol{\pi}}_G^{(j)} = \widehat{\boldsymbol{\pi}}_G^{(j)agg} = \widehat{\mathbf{y}}_G^{(j)}. \tag{10}$$

Proof. Let us denote by $P^{(j)}$ (respectively by $P'^{(j)}$) the stochastic matrix of $X^{(j)}$ (respectively of $Y^{(j)}$). We have:

$$P^{(j)} = \begin{pmatrix} P_G & P_{G\bar{G}} \\ P_{\bar{G}G}^{(j)} & P_{\bar{G}} \end{pmatrix}, \quad P'^{(j)} = \begin{pmatrix} P_G & P'_{G\bar{G}} \\ P'^{(j)}_{\bar{G}G} & P'_{\bar{G}} \end{pmatrix},$$

where the matrix $P'_{G\bar{G}}$ is equal to $(P_{GC_K}\mathbf{1}^{\mathrm{T}}, \ldots, P_{GC_M}\mathbf{1}^{\mathrm{T}})$.

Given that $\boldsymbol{\pi}^{(j)}P^{(j)} = \boldsymbol{\pi}^{(j)}$ and $\mathbf{y}^{(j)}P'^{(j)} = \mathbf{y}^{(j)}$, we have:

$$\begin{cases} \widehat{\boldsymbol{\pi}}_G^{(j)}\left(P_G - P_{G\bar{G}}P_{\bar{G}}^{-1}P_{\bar{G}G}^{(j)}\right) = \widehat{\boldsymbol{\pi}}_G^{(j)} \\ \widehat{\mathbf{y}}_G^{(j)}\left(P_G - P'_{G\bar{G}}(P'_{\bar{G}})^{-1}P'^{(j)}_{\bar{G}G}\right) = \widehat{\mathbf{y}}_G^{(j)} \end{cases}$$

From [1, theorem 8], it follows that $\widehat{\boldsymbol{\pi}}_G^{(j)}$ and $\widehat{\mathbf{y}}_G^{(j)}$ belong to the polyhedron $\mathcal{P}((I - P_G)^{-1})$. From Appendix A, Lemma 8, $\widehat{\boldsymbol{\pi}}_G^{(j)}$ and $\widehat{\mathbf{y}}_G^{(j)}$ are both equal to the jth vertex of $\mathcal{P}((I - P_G)^{-1})$. From (7), we have directly the first equality between $\widehat{\boldsymbol{\pi}}_G^{(j)}$ and $\widehat{\boldsymbol{\pi}}_G^{(j)agg}$ ■

Theorem 3 *Between $y^{(j)}(G) = \Pr(Y_\infty^{(j)} \in G)$ where $Y_\infty^{(j)}$ is a stationary version of $Y^{(j)}$, and $\pi^{(j)\mathrm{agg}}(G) = \pi^{(j)}(G)$, we have the relation*

$$y^{(j)}(G) \le \pi^{(j)}(G). \tag{11}$$

If one of the inequalities (8) or (9) is strict, then $y^{(j)}(G) < \pi^{(j)}(G)$.

Proof. See Appendix B. ■

Define over S^{agg} the two reward vectors $\mathbf{r_1}$ and $\mathbf{r_2}$ obtained by completing vector $\mathbf{r}_G$ with rewards on the aggregated states $c_K, c_{K+1}, \ldots, c_M$ equal to ϱ_1 in $\mathbf{r_1}$ and equal to ϱ_2 in $\mathbf{r_2}$.

Theorem 4

$$\min_{j \in in(G)} \mathbf{y}^{(j)}\mathbf{r_1}^{\mathrm{T}} \le R, \tag{12}$$

$$\max_{j \in in(G)} \mathbf{y}^{(j)}\mathbf{r_2}^{\mathrm{T}} \ge R. \tag{13}$$

Proof. Let us consider the expression of $\mathbf{y}^{(j)}\mathbf{r_1}^{\mathrm{T}}$:

$$\begin{aligned} \mathbf{y}^{(j)}\mathbf{r_1}^{\mathrm{T}} &= \mathbf{y}_G^{(j)}\mathbf{r}_G^{\mathrm{T}} + \varrho_1 y^{(j)}(\bar{G}) \\ &= y^{(j)}(G)\left(\widehat{\mathbf{y}}_G^{(j)}\mathbf{r}_G^{\mathrm{T}} - \varrho_1\right) + \varrho_1. \end{aligned}$$

Given that $\pi^{(j)}(G) = \pi^{(j)\mathrm{agg}}(G)$ from (7), that $\widehat{\mathbf{y}}_G^{(j)}\mathbf{r}_G^{\mathrm{T}} \ge \varrho_1$ (since $r_i \ge \varrho_1$ for all state i) and using Theorem 2, we have the following inequality:

$$\begin{aligned} \mathbf{y}^{(j)}\mathbf{r_1}^{\mathrm{T}} &\le \pi^{(j)}(G)\left(\widehat{\boldsymbol{\pi}}_G^{(j)}\mathbf{r}_G^{\mathrm{T}} - \varrho_1\right) + \varrho_1 \\ &= \boldsymbol{\pi}^{(j)}\mathbf{r_1}^{\mathrm{T}} = R_1^{(j)} \le R^{(j)}. \end{aligned}$$

From Corollary 1, we have:

$$\min_{j \in in(G)} \mathbf{y}^{(j)} \mathbf{r}_{\mathbf{1}}^{\mathrm{T}} \le \min_{j \in in(G)} R_1^{(j)} \le \min_{j \in in(G)} R^{(j)} \le R. \tag{14}$$

In the same way, denoting $R_2^{(j)} = \boldsymbol{\pi}^{(j)} \mathbf{r}_{\mathbf{2}}^{\mathrm{T}}$ and writing

$$\mathbf{y}^{(j)} \mathbf{r}_{\mathbf{2}}^{\mathrm{T}} = y^{(j)}(G) \left(\widehat{\mathbf{y}}_G^{(j)} \mathbf{r}_G^{\mathrm{T}} - \varrho_2 \right) + \varrho_2$$

and observing that $\widehat{\mathbf{y}}_G^{(j)} \mathbf{r}_G^{\mathrm{T}} - \varrho_2 \le 0$, we obtain

$$\mathbf{y}^{(j)} \mathbf{r}_{\mathbf{2}}^{\mathrm{T}} \ge R_2^{(j)} \ge R^{(j)},$$

and thus

$$\max_{j \in in(G)} \mathbf{y}^{(j)} \mathbf{r}_{\mathbf{2}}^{\mathrm{T}} \ge \max_{j \in in(G)} R_2^{(j)} \ge \max_{j \in in(G)} R^{(j)} \ge R. \tag{15}$$

■

Resuming, the bounds of R are obtained using (14) and (15). To do this, we must be able to build chain $Y^{(j)}$, that is, to build bounds $\lambda_{I,J}^+$ and μ_I^- of the corresponding (unknown) transition rates in $X^{(j)\mathrm{agg}}$ (relations (8) and (9)). We describe now the way this is done in [3]. Next section describes our technique, which has the same goal.

The approach of [3]. Recall that we want to compute $\lambda_{I,J}^+$ and μ_I^- *without* the knowledge of $\boldsymbol{\pi}^{(j)}$, the (unknown) stationary distribution of $X^{(j)}$ for each $j \in in(G)$, it would be nice to use

$$\forall I, J \text{ s.t. } K \le I < J \le M, \quad \lambda_{I,J}^+ = \max_{i \in C_I} \sum_{l \in C_J} A_{i,l}, \tag{16}$$

$$\forall I \text{ s.t. } K \le I \le M, \quad \mu_I^- = \min_{i \in C_I} \sum_{l \in C_{I-1}} A_{i,l}, \tag{17}$$

This is the idea followed in [3]. The use of relation (16) immediately implies that $\lambda_{I,J}^+ > 0$. Let us examine now the bounds μ_I^-. In order to have $\mu_I^- > 0$ for any value of K, we need a supplementary condition to be satisfied by X:

Condition 2 *For any index $I \neq 0$, for all state $i \in C_I$ there exists at least a state $j \in C_{I-1}$ such that $A_{i,j} > 0$.*

This can be quite restrictive as we will illustrate later, but the interest relies in the fact that it allows to obtain direct lower bounds of the $\mu_I^{(j)}$'s. In Section 5, we develop a new approach that does not need this assumption, allowing to work with much more general models.

Other related works. Before presenting the method that we propose, let us briefly describe other related papers in the area. First note the approach of [5] who construct from the original model two new models which respectively lower and upper bound the measure using particular job-local-balance equations. However this technique doesn't give tight bounds and becomes more complicated to apply with complex systems [5]. The starting work from which papers like this one are built are [1] and [2]. In [6], these results are improved, following the same research lines. A different improvement is [3] and we follow here this approach to obtain a more powerful bounding technique. Briefly, in [6] the author derives a general iterative bounding technique having the following main differences with [3] and with our work: it can be applied without restrictions (while ours or the method of [3] needs some conditions to work) but is more expensive. The final form of the approach of [6] is quite different, however: it is presented as an iterative process where one bounds the conditional distributions the $\widehat{\boldsymbol{\pi}}_{C_I}$'s, several times if necessary, and of the probabilities $\pi(C(I))$, in order to derive bounds on the total vector π and then on R. The key point in the complexity comparison is that [6] basically needs the inversion of a matrix to obtain each necessary bound (for vectors $\widehat{\boldsymbol{\pi}}_{C_I}$'s and for probabilities $\pi(C(I))$'s). The technique in [3] exploits the strong Condition 2 to obtain a less expensive process. Our technique is more expensive than this one, but it needs much less restrictive conditions. This last feature allows us to obtain tight bound for some performance models on infinite state spaces, as shown in our paper.

To reduce the computational cost of the bounds obtained by the first algorithm of [3], a technique of "duplication of states" is proposed in [3], [7] and [8]. The main objective in [3] is to reduce the number of linear systems to solve. In [7], the authors propose a "multi-step bounding algorithm" to improve the bounds by increasing the threshold K without restarting the work from the beginning. That is, the results for level $K + 1$ can use those corresponding to level K. But the spread between the bounds has a non-zero limiting value [8]. This problem is handled in [8] using a technique called "bound spread reduction" to reduce the error introduced at each step of the previous iterative procedure. A simple heuristic to choose between the "multi-step bounding algorithm" and the "bound spread reduction" is developed and illustrated in [8]. The method that we propose here can also use the same technique to improve its efficiency (however, we do not develop this point in the paper).

The same authors [9] have chosen another approach to bound mean response times in heterogeneous multi-server queuing systems. From the original model, they construct two new models, one to obtain a lower bound of the considered measure and another one to obtain an upper bound. Another line of research is [10] which, improving previous work by the same author, gives better bounds when additional information ("distance to failure") is available. In that paper the number of linear systems to be solved is also reduced.

5 The proposed method

This is the main part of the paper. Its goal is to derive a method to avoid Condition 2 and still be able to bound the measure R. We start by introducing the idea informally. Then, we recall some results of [4], where the authors analyze the sojourn times of a Markov chain X in a part of its state space, and the asymptotic behavior of these (in general dependent) random variables, and we derive a new one (Lemma 3), which we need to formalize the method.

5.1 The idea underlying our method

Consider a birth and death process and denote λ_i (respectively μ_i) the birth rate (respectively death rate) associated with state i. The mean sojourn time in state i (or mean holding time) is $h_i = 1/(\lambda_i + \mu_i)$ and the probability that after visiting i the next state is $i-1$ is $p_i = \mu_i/(\lambda_i + \mu_i)$. Observe then that

$$\mu_i = \frac{p_i}{h_i}. \tag{18}$$

The intuitive idea leading to our bounding technique is to write the (exact) aggregated rate in $X^{(j)}$ from class C_I from C_{I-1}, that is, the transition rate from c_I to c_{I-1} in $X^{(j)\mathrm{agg}}$, $\mu_I^{(j)}$, in a similar form than (18). In the next subsection we write it as (relation (24))

$$\mu_I^{(j)} = \frac{p_I^{(j)}}{h_I^{(j)}},$$

and we derive useful expressions of $p_I^{(j)}$ and $h_I^{(j)}$ allowing us to obtain a bound of $\mu_I^{(j)}$.

5.2 Sojourn times and aggregation of states

Let us denote by $H_{I,n}^{(j)}$ the length of the nth sojourn of $X_\infty^{(j)}$ in class C_I. The first visited state of C_I during this sojourn is denoted by $V_{I,n}^{(j)}$ ($V_{I,n}^{(j)} \in in(C_I)$) and after leaving C_I, the next visited state is denoted by $W_{I,n}^{(j)}$ ($W_{I,n}^{(j)} \in in(\bar{C}_I)$).

The distribution of $V_{I,n}^{(j)}$, as a row vector $\boldsymbol{v_I^{(j)}}(n)$ defined over C_I, is given by the following expression [4]:

$$\boldsymbol{v_I^{(j)}}(n) = \boldsymbol{v_I^{(j)}}(1)(B_{C_I}^{(j)})^{n-1},$$

where $B_{C_I}^{(j)}$ is the stochastic matrix $A_{C_I}^{-1} A_{C_I,\bar{C}_I}^{(j)} A_{\bar{C}_I}^{-1} A_{\bar{C}_I,C_I}^{(j)}$ and $\boldsymbol{v_I^{(j)}}(1) = \boldsymbol{\pi}_{C_I}^{(j)} - \boldsymbol{\pi}_{\bar{C}_I}^{(j)} A_{\bar{C}_I}^{-1} A_{\bar{C}_I,C_I}^{(j)}$. Of course for all $n > 1$, $\boldsymbol{v_I^{(j)}}(n)$ has non-zero entries only on states i belonging to $in(C_I)$.

The distribution of $H_{I,n}^{(j)}$ is given by [4]

$$\Pr(H_{I,n}^{(j)} > t) = \boldsymbol{v_I^{(j)}}(n) \exp(A_{C_I} t) \mathbf{1}^{\mathrm{T}}$$

and its mean is

$$\mathrm{E}(H_{I,n}^{(j)}) = -\boldsymbol{v_I^{(j)}}(n) A_{C_I}^{-1} \mathbf{1}^{\mathrm{T}}.$$

In [4] it is in particular shown that vector

$$\boldsymbol{v_I^{(j)}} = \frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I}}{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I} \mathbf{1}^{\mathrm{T}}} \tag{19}$$

is the stationary distribution of the Markov chain $(V_{I,n}^{(j)})_n$, that is, if $\boldsymbol{v_I^{(j)}}(1) = \boldsymbol{v_I^{(j)}}$ then $\boldsymbol{v_I^{(j)}}(n) = \boldsymbol{v_I^{(j)}}$ for all $n \geq 1$. We denote by $v_{I,i}^{(j)}$ the component of $\boldsymbol{v_I^{(j)}}$ corresponding to state $i \in C_I$. We should also note that $\boldsymbol{v_G^{(j)}} = \boldsymbol{e_j}$, where $\boldsymbol{e_j}$ is the jth row vector of the canonical base in $\mathbb{R}^{|G|}$.

Let us consider now some relationships between chains $X^{(j)}$ and $X^{(j)\mathrm{agg}}$ from the sojourn time point of view. We denote by $h_I^{(j)}$ the mean holding time of $X^{(j)\mathrm{agg}}$ in state c_I, $K \leq I \leq M$, that is,

$$\text{for } I \geq K, \quad h_I^{(j)} = \frac{1}{\mu_I^{(j)} + \sum_{J>I} \lambda_{I,J}^{(j)}}. \tag{20}$$

A result needed here is given in the following lemma:

Lemma 2

$$\textit{For any } I \geq K, \quad h_I^{(j)} = -\boldsymbol{v_I^{(j)}} A_{C_I}^{-1} \mathbf{1}^{\mathrm{T}}. \tag{21}$$

For the proof, see [4], basically Corollary 4.6. Lemma 2 says that the mean holding time (i.e. the mean sojourn time) of $X^{(j)\mathrm{agg}}$ in c_I is equal to the mean sojourn time of $X^{(j)}$ in C_I when it enters C_I by state i with probability $v_{I,i}^{(j)}$ (for instance, think of the first sojourn in C_I of a version of $X^{(j)}$ having as initial distribution the vector $\boldsymbol{\alpha^{(j)}}$ such that $\boldsymbol{\alpha_{C_I}^{(j)}} = \boldsymbol{v_I^{(j)}}$).

Let us denote by $\widehat{h}_{i,I}$ the mean sojourn time of $X^{(j)}$ in C_I conditioned to the fact that the process enters the set C_I by state i. Observe that, for all $j \in in(G)$,

$$\widehat{h}_{i,I} = \mathrm{E}(H_{I,n}^{(j)} \mid V_{I,n}^{(j)} = i) \text{ for all } n \geq 1. \tag{22}$$

From Relations (21) and (22), we can write

$$h_I^{(j)} = \sum_{i \in in(C_I)} v_{I,i}^{(j)} \, \widehat{h}_{i,I}. \tag{23}$$

Now, for the purposes of this paper, we have to consider the event “when the nth sojourn of $X^{(j)}$ in C_I ends, the next visited state belongs to C_{I-1}”, that is, $\{W_{I,n}^{(j)} \in C_{I-1}\}$. It is straightforward to verify that the probability of this event is

$$\text{for } I > K, \quad \Pr(W_{I,n}^{(j)} \in C_{I-1}) = -\boldsymbol{v_I^{(j)}}(n) A_{C_I}^{-1} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}.$$

When $I = K$, we also have

$$\Pr(W_{K,n}^{(j)} = j \in in(G)) = -\boldsymbol{v_K^{(j)}}(n) A_{C_K}^{-1} A_{C_K,C_{K-1}}^{(j)} \mathbf{1}^{\mathrm{T}},$$

where $A_{C_K,C_{K-1}}^{(j)}$ is the matrix equal to $A_{C_K,C_{K-1}} \mathbf{1}^{\mathrm{T}} \boldsymbol{e_j}$, and $\boldsymbol{e_j}$ is the jth row vector of the canonical base in $\mathbb{R}^{|C_{K-1}|}$.

The event similar to $\{W_{I,n}^{(j)} \in C_{I-1}\}$ in the aggregated chain $X^{(j)\mathrm{agg}}$ is “when leaving state c_I, the chain jumps to c_{I-1}”. Its probability is

$$p_I^{(j)} = \frac{\mu_I^{(j)}}{\mu_I^{(j)} + \sum_{J>I} \lambda_{I,J}^{(j)}}.$$

Observe that the transition rate $\mu_I^{(j)}$ from c_I to c_{I-1} in $X^{(j)\mathrm{agg}}$ is

$$\mu_I^{(j)} = \frac{p_I^{(j)}}{h_I^{(j)}}. \tag{24}$$

The following result (similar to Lemma 2) holds:

Lemma 3

$$\textit{For any } I > K, \;\; p_I^{(j)} = -\boldsymbol{v_I^{(j)}} A_{C_I}^{-1} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}} \tag{25}$$

and

$$p_K^{(j)} = -\boldsymbol{v_K^{(j)}} A_{C_K}^{-1} A_{C_K,C_{K-1}}^{(j)} \mathbf{1}^{\mathrm{T}}. \tag{26}$$

Proof.

Let $I > K$. From (6) and from (4), we can write

$$\lambda_{I,J}^{(j)} = \frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I,C_J} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} \mathbf{1}^{\mathrm{T}}}, \quad \mu_I^{(j)} = \frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} \mathbf{1}^{\mathrm{T}}}.$$

Then,

$$\begin{aligned}\mu_I^{(j)} + \sum_{J>I} \lambda_{I,J}^{(j)} &= \frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I,\bar{C}_I} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} \mathbf{1}^{\mathrm{T}}} \\ &= -\frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} \mathbf{1}^{\mathrm{T}}} \\ &\quad \text{(since } A_{C_I,\bar{C}_I}\mathbf{1}^{\mathrm{T}} = -A_{C_I}\mathbf{1}^{\mathrm{T}}).\end{aligned}$$

This leads to

$$\begin{aligned}p_I^{(j)} &= \frac{\mu_I^{(j)}}{\mu_I^{(j)} + \sum_{J>I} \lambda_{I,J}^{(j)}} \\ &= \frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} \mathbf{1}^{\mathrm{T}}} \left(-\frac{\boldsymbol{\pi}_{C_I}^{(j)} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I} \mathbf{1}^{\mathrm{T}}} \right) \\ &= -\frac{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\pi}_{C_I}^{(j)} A_{C_I} \mathbf{1}^{\mathrm{T}}}.\end{aligned}$$

It remains to check that the last expression is equal to $-\boldsymbol{v}_{\boldsymbol{I}}^{(\boldsymbol{j})} A_{C_I}^{-1} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}$. The case of class C_K is proven in the same way. ■

As for Lemma 2, Lemma 3 says that the probability that $X^{(j)\mathrm{agg}}$ will jump to c_{I-1} when leaving c_I is the same as the conditional probability for $X^{(j)}$ to jump to C_{I-1} when leaving C_I, given that $X^{(j)}$ enters C_I by state i with probability $v_{I,i}^{(j)}$. This implies that, if we denote by $\widehat{p}_{i,I}$ the conditional probability that $X^{(j)}$ jumps to C_{I-1} when leaving C_I, given that the sojourn started in state $i \in C_I$, we have

$$p_I^{(j)} = \sum_{i \in in(C_I)} v_{I,i}^{(j)} \, \widehat{p}_{i,I}. \tag{27}$$

5.3 The bounding algorithm

Let us assume that Condition 2 is not satisfied. To obtain the bounds of R given in (12) and (13), we proceed as in [3]. The problem is the computation of lower bounds of the $\mu_I^{(j)}$'s.

First let us consider a new subset of states of C_I, $in(C_I)^*$, which is the set of entry points i of C_I such that if $X^{(j)}$ enters C_I by i, there is a non-null probability that the next visited class is C_{I-1}:

$$in(C_I)^* = \{i \in in(C_I) \mid \widehat{p}_{i,I} > 0\}.$$

Our methods needs then that the two sets $in(C_I)^*$ and $in(C_I)$ are equal. Let us put it explicitly.

Condition 3 *For all $I > 1$ and all $i \in in(C_I)$, the probability to jump from C_I to C_{I-1} when the sojourn in C_I starts in i, is not null (that is, $in(C_I)^* = in(C_I)$).*

This condition is obviously much less restrictive than Condition 2. We did not find any realistic model where it does not hold. Under Condition 3, lower bounds of the $\mu_I^{(j)}$'s are given in the following result.

Theorem 5 *For all $I \geq K$, for all $j \in in(G)$,*

$$\mu_I^* = \min_{i \in in(C_I)} \frac{\widehat{p}_{i,I}}{\widehat{h}_{i,I}} \leq \mu_I^{(j)}. \tag{28}$$

Proof. The result simply follows from relations (23) and (27), writing that, for any $I \geq K$,

$$\mu_I^{(j)} = \frac{p_I^{(j)}}{h_I^{(j)}} = \frac{\sum_{i \in in(C_I)} v_{I,i}^{(j)} \, \widehat{p}_{i,I}}{\sum_{i \in in(C_I)} v_{I,i}^{(j)} \, \widehat{h}_{i,I}},$$

and then using the fact that $\sum_{i \in C_I} v_{I,i}^{(j)} = 1$. ■

Let us resume the algorithm. The input data are the partition and a given level K. The steps to be followed are the following:

- Once the partition and the threshold K fixed, compute the starred bounds given by (28). To do this,
  - for each class C_I, $I \geq K$, compute the $\widehat{p}_{i,I}$'s and the $\widehat{h}_{i,I}$'s ; alternatively, lower bounds of the $\widehat{p}_{i,I}$'s and upper bounds of the $\widehat{h}_{i,I}$'s can be used (we will see in next section that the infinite models are analyzed this way)
  - compute μ_I^* using (28).
- Generate G and then, for any $j \in in(G)$, find the stationary distribution $\mathbf{y}^{(j)}$ of the chain $Y^{(j)}$ with the choice $\mu_I^- = \mu_I^*$. Possibly, use the techniques in [8] or in [10] to reduce the number of linear systems to solve.
- Compute the lower and upper bounds of R using (12) and (13).

The main drawback of this algorithm is that the computation of the $\widehat{p}_{i,I}$'s and the $\widehat{h}_{i,I}$'s may be numerically intractable due to the possible size of class C_I. Moreover, even if they can be calculated, the induced cost may be too high for the user. A possible way to handle this problem is to try to obtain new bounds on these numbers. We are following this direction in our current research work. In this paper, we want only to illustrate the use of our approach in cases where deriving the bounds can be done analytically. However we can note that if Condition 2 holds, the bounds obtained by the new algorithm are better than those of [3], as stated in next result.

Lemma 4 *If Condition 2 holds (which implies that Condition 3 holds as well), then for all $I \geq K$, $\mu_I^- \leq \mu_I^* \leq \mu_I$.*

Proof. Let us denote by $\boldsymbol{e_i}$ the vector whose entries are 0 except the ith one, which is equal to 1. From Lemmas 2 and 3, we have:

$$\widehat{h}_{i,I} = -\boldsymbol{e_i} A_{C_I}^{-1} \mathbf{1}^{\mathrm{T}}, \tag{29}$$

$$\widehat{p}_{i,I} = -\boldsymbol{e_i} A_{C_I}^{-1} A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}. \tag{30}$$

Let us consider the vector $A_{C_I,C_{I-1}} \mathbf{1}^{\mathrm{T}}$. Each one of its components are greater than $\min_{i \in C_I} \sum_{j \in C_{I-1}} A_{i,j}$, that is, greater than μ_I^-. From the relations above, we have

$$\widehat{p}_{i,I} \geq \mu_I^- \widehat{h}_{i,I}$$

which ends the proof. ■

6 Bounding R in infinite models

In this section, we adapt the method described before to the case of an infinite state space S and finite classes c_I's (implying that $M = \infty$). We assume that X is ergodic. We also assume, of course, that $R < \infty$. Given that each class is finite, the cardinality of $in(G)$ is also finite. Observe first that when $|S| = \infty$, the arguments used in Lemma 1 remain valid, proving here that from any state i there is a finite path to j in $X^{(j)}$. Now, since the infinitesimal generator of $X^{(j)}$ is the same as the infinitesimal generator of X except for a finite part of S, necessary $X^{(j)}$ is also ergodic, and therefore, with a single positive recurrent class containing j. In Appendix C, we show that the conclusions of Theorem 1 are still valid in the infinite state space case. Moreover, since $R < \infty$, we necessarily have $R^{(j)} < \infty$ as well. To see this, observe that if for some $j_0 \in in(G)$ we have $R^{(j_0)} = \boldsymbol{\pi}^{(j_0)} \mathbf{r}^{\mathrm{T}} = \infty$, then from the expression

$$R = \boldsymbol{\pi} \mathbf{r}^{\mathrm{T}} = \sum_{j \in in(G)} \beta_j \boldsymbol{\pi}^{(j)} \mathbf{r}^{\mathrm{T}}$$

we obtain $R = \infty$, in contradiction with our starting assumption.

To simplify the analysis, we add the assumption that the transitions from C_I to C_J are null if $I - J > 1$. If this condition is not verified, the corresponding relations are more complex, but the method still applies.

Assume that Conditions 1 and 3 hold. We can compute the starred bounds of previous subsection.

For each $j \in in(G)$, let us define a new chain $Z^{(j)}$ over the state space $G \cup \{c\}$ with the same transition rates than X inside G. From any $g \in G$ to c the transition rate is equal to $A^{(j)\mathrm{agg}}_{g,c_K}$ as in $X^{(j)\mathrm{agg}}$. From c to G, there exists a single non-null transition rate which is from c to j, denoted by ν_j, and defined by

$$\nu_j = \frac{\mu^{(j)}_K}{\sum_{I=K}^{\infty} \theta^{(j)}_I} \tag{31}$$

where $\theta^{(j)}_K = 1$ and, for $I > K$,

$$\theta^{(j)}_I = \frac{\lambda^{(j)}_{K,K+1} \cdots \lambda^{(j)}_{I-1,I}}{\mu^{(j)}_{K+1} \cdots \mu^{(j)}_I}. \tag{32}$$

Observe that since X is assumed to be ergodic, $\sum_I \theta^{(j)}_I < \infty$.

Denoting by $\boldsymbol{z}^{(j)}$ the stationary distribution of $Z^{(j)}$, we have

Lemma 5

$$z^{(j)}_c = \sum_{I=K}^{\infty} \pi^{(j)\mathrm{agg}}_{c_I}$$

and for all $g \in G$,

$$z^{(j)}_g = \pi^{(j)\mathrm{agg}}_g.$$

Proof. The proof is immediate by writing the equilibrium equations for $X^{(j)\mathrm{agg}}_{\infty}$

$$\pi^{(j)\mathrm{agg}}_{c_I} \lambda^{(j)}_{I,I+1} = \pi^{(j)\mathrm{agg}}_{c_{I+1}} \mu^{(j)}_{I+1},$$

which imply that

$$\pi^{(j)\mathrm{agg}}_{c_I} = \theta^{(j)}_I \pi^{(j)\mathrm{agg}}_{c_K}.$$

In order to have the equality $z^{(j)}_c = \sum_{I=K}^{\infty} \pi^{(j)\mathrm{agg}}_{c_I}$ we need

$$\nu_j = \frac{\pi^{(j)\mathrm{agg}}_{c_K} \mu^{(j)}_K}{\sum_{I=K}^{\infty} \pi^{(j)\mathrm{agg}}_{c_I}}.$$

The result follows by writing $\pi^{(j)\mathrm{agg}}_{c_I}$ as a function of $\pi^{(j)\mathrm{agg}}_{c_K}$ in this last relation. ■

In the same way, let us define another irreducible Markov chain $Z^{'(j)}$ and its stationary distribution $\boldsymbol{z}^{'(j)}$ over $G \cup \{c\}$, with the same rates than for $Z^{(j)}$ inside G and from any $g \in G$ to c. The only entry in G from c is j and the transition rate from c to j, denoted by $\nu^{'}_{j}$, is defined as follows:

$$\nu^{'}_{j} = \frac{\mu^*_K}{\sum_{I=K}^{\infty} \theta^*_I} \tag{33}$$

where

$$\forall I > K, \ \ \theta^*_I = \frac{\lambda^+_{K,K+1} \cdots \lambda^+_{I-1,I}}{\mu^*_{K+1} \cdots \mu^*_I}, \ \ \theta^*_K = 1. \tag{34}$$

Then, we have the following results:

Lemma 6 *When $Y^{(j)}$ is built using rates $\lambda^+_{I,I+1}$ and μ^*_I, $I \geq K$, we have*

$$z^{'(j)}_c = \sum_{I=K}^{\infty} y^{(j)}_{c_I}$$

and for all $g \in G$,

$$z^{'(j)}_g = y^{(j)}_g.$$

Proof. The proof is as for Lemma 5. ■

At this point, you should note that $\nu^{'}_{j} \leq \nu_j$, since

$$\mu^*_K \sum_{I=K}^{\infty} \theta_I \leq \mu_K \sum_{I=K}^{\infty} \theta^*_I,$$

and that $\nu^{'}_{j}$ is in fact independent of j because we use an upper bound of $\lambda^{(j)}_I$ which is independent of j itself.

Let us come back to our bounds. Let us define over $G \cup \{c\}$, the two reward vectors $\mathbf{r}^{'}_{\mathbf{1}}$ and $\mathbf{r}^{'}_{\mathbf{2}}$ obtained by completing vector $\mathbf{r}_G$ with a reward on the aggregated state c equal to ϱ_1 in $\mathbf{r}^{'}_{\mathbf{1}}$ and denoted by r_c in $\mathbf{r}^{'}_{\mathbf{2}}$. Since $0 \leq \varrho_1 < \infty$, to obtain a lower bound on R we proceed exactly as in the previous section. The problem can arise in the case of $\varrho_2 = \infty$. First, let us consider the case when ϱ_2 is finite.

Theorem 6 *If $\varrho_2 < \infty$, let r_c be equal to ϱ_2. Then, we have the following bounds of R:*

$$\min_{j \in in(G)} \boldsymbol{z}^{'(j)} \mathbf{r}^{'\mathrm{T}}_{\mathbf{1}} \leq R, \tag{35}$$

$$\max_{j \in in(G)} \boldsymbol{z}^{'(j)} \mathbf{r}^{'\mathrm{T}}_{\mathbf{2}} \geq R. \tag{36}$$

Proof. From Lemma 6, we know that

$$\boldsymbol{z}'^{(j)}\mathbf{r}'^{\mathrm{T}}_{\mathbf{1}} = \mathbf{y}^{(j)}_{G}\mathbf{r}^{\mathrm{T}}_{G} + (1 - y^{(j)}(G))\varrho_1,$$

and

$$\boldsymbol{z}'^{(j)}\mathbf{r}'^{\mathrm{T}}_{\mathbf{2}} = \mathbf{y}^{(j)}_{G}\mathbf{r}^{\mathrm{T}}_{G} + (1 - y^{(j)}(G))\varrho_2.$$

Then, the proof is identical to the proof of Theorem 4. ■

Assume now that $\varrho_2 = \infty$ and denote by r_{c_I} an upper bound of the rewards on C_I. Then we have the following preliminary result:

Lemma 7 *If $\varrho_2 = \infty$, under the condition*

$$\sum_{I=K}^{\infty} \theta^{(j)}_I r_{c_I} < \infty,$$

letting

$$r_c = \max\left\{ \max_{j\in in(G)} \frac{\sum_{I=K}^{\infty}\theta^{(j)}_I r_{c_I}}{\sum_{I=K}^{\infty}\theta^{(j)}_I}, \max_{i\in G} r_i \right\}, \tag{37}$$

an upper bound of the expected reward R is

$$\max_{j\in in(G)} \left(\sum_{i\in G} z^{(j)}_i r_i + r_c z^{(j)}_c \right) \geq R. \tag{38}$$

Proof. For any $j \in in(G)$,

$$\begin{aligned} \boldsymbol{\pi}^{(j)}\mathbf{r}^{\mathrm{T}} &= \boldsymbol{\pi}^{(j)}_{G}\mathbf{r}^{\mathrm{T}}_{G} + \boldsymbol{\pi}^{(j)}_{\bar{G}}\mathbf{r}^{\mathrm{T}}_{\bar{G}} \\ &= \boldsymbol{\pi}^{(j)\mathrm{agg}}_{G}\mathbf{r}^{\mathrm{T}}_{G} + \boldsymbol{\pi}^{(j)}_{\bar{G}}\mathbf{r}^{\mathrm{T}}_{\bar{G}} \\ &\leq \boldsymbol{\pi}^{(j)\mathrm{agg}}_{G}\mathbf{r}^{\mathrm{T}}_{G} + \sum_{I=K}^{\infty} \pi^{(j)\mathrm{agg}}_{c_I} r_{c_I}. \end{aligned}$$

From the proof of Lemma 5, $\pi^{(j)\mathrm{agg}}_{c_I} = \theta^{(j)}_I \pi^{(j)\mathrm{agg}}_{c_K}$, so, given that $\sum_{I=K}^{\infty}\theta^{(j)}_I r_{c_I} < \infty$,

$$\boldsymbol{\pi}^{(j)}\mathbf{r}^{\mathrm{T}} \leq \boldsymbol{z}^{(j)}_{G}\mathbf{r}^{\mathrm{T}}_{G} + \pi^{(j)\mathrm{agg}}_{c_K} \sum_{I=K}^{\infty} \theta^{(j)}_I r_{c_I}.$$

From the definition of r_c, we have

$$\boldsymbol{\pi}^{(j)}\mathbf{r}^{\mathrm{T}} \leq \boldsymbol{z}^{(j)}_{G}\mathbf{r}^{\mathrm{T}}_{G} + z^{(j)}_c r_c.$$

■

Thus, obtaining an upper bound of r_c allows us to derive an upper bound of R, as stated in the next theorem:

Theorem 7 *If $r_c^* \geq r_c$, then an upper bound of the expected reward R is*

$$\max_{j \in in(G)} \left(\sum_{i \in G} z_i^{'(j)} r_i + r_c^* z_c^{'(j)} \right) \geq R. \tag{39}$$

Proof. Consider the expression of $\boldsymbol{z}^{'(j)} \mathbf{r_2}^{'\mathrm{T}}$

$$\boldsymbol{z}^{'(j)} \mathbf{r_2}^{'\mathrm{T}} = \boldsymbol{z}_G^{'(j)} \mathbf{r}_G^{\mathrm{T}} + z_c^{'(j)} r_c.$$

From Theorem 2, we have

$$\widehat{\boldsymbol{z}}_G^{'(j)} = \widehat{\boldsymbol{z}}_G^{(j)}.$$

Using the remark above, $\nu_j^{'} \leq \nu_j$, and Theorem 3, we also have $z^{'(j)}(G) \leq z^{(j)}(G)$, and so,

$$z_c^{'(j)} \geq z_c^{(j)}.$$

Then given that $r_c \geq r_i, \ \forall i \in in(G)$,

$$\boldsymbol{z}^{(j)} \mathbf{r_2}^{'\mathrm{T}} \leq \boldsymbol{z}^{'(j)} \mathbf{r_2}^{'\mathrm{T}}.$$

It follows that

$$\sum_{i \in G} z_i^{(j)} r_i + r_c z_c^{(j)} \leq \sum_{i \in G} z_i^{'(j)} r_i + r_c z_c^{'(j)}.$$

Then if $r_c^* \geq r_c$, we obtain (39). ■

This tells us that if we can compute an upper bound of r_c, we have in the right hand side of (39) an upper bound of R.

A particular case

Let us analyze what happens when $\lambda_{I,J}^{+}$ and μ_I^* are constant and respectively equal to λ and μ. A condition for the stability of the model is given by $\lambda < \mu$. Denoting $\varrho = \lambda/\mu$, we have

$$\theta_I^* = \varrho^{I-K}, \ \ \nu_j^{'} = \mu(1-\varrho).$$

In the examples that follow (7.2 and 7.3), we are in this particular case. We just have to bound r_c with

$$r_c^* = \max\left(\max_{i \in G} r_i, \frac{\sum_{I=K}^{\infty} \theta_I^* r_{c_I}}{\sum_{I=K}^{\infty} \theta_I^*} \right) = \max\left(\max_{i \in G} r_i, (1-\varrho) \sum_{I=K}^{\infty} \varrho^{I-K} r_{c_I} \right).$$

This is the technique that we will use to obtain bounds in the queuing models of next section.

7 Illustrations

This section illustrates the efficiency of the bounding method proposed in the paper. First, we use a standard dependability model, a “Machine Repairman Model”, which leads to a large finite Markov chain that can not be handled by the technique published in [3]. The second example is an open queuing network composed by two queues in series, leading to an infinite Markov chain with no known analytical solution. In this case, we bound the mean number of customers in each node. We can observe that this model can not be handled by matrix-geometric techniques. The third example is another open tandem of queues. Here, there are blocking mechanisms since we consider finite buffers in all the nodes except the first one, and we bound blocking probabilities. These two open examples can be transformed into closed versions by limiting the total allowed number of customers and in this case, the conditions necessary to use the method of [3] do not hold neither, as in the first example.

7.1 Bounding the asymptotic availability of a MRM

Our first example is a standard multi-component system subject to failures and repairs. There are two types of components. The number of components of type k is denoted by N_k and their time to failure is exponentially distributed with parameter λ_k, $k = 1, 2$. Think for instance of a communication network where the components are nodes or lines. In such a system we can easily find a large number of components leading in turn to models with huge state spaces.

After a failure, the components enter a repair facility with one server and repair time distributed according to a d_k-stage Coxian distribution for type k machines, with mean m_k. Type 1 components are served with higher priority than type 2, and the priority is non preemptive. We assume that type 2 units are put immediately in operation when repaired, but that type 1 ones need a delay exponentially distributed (with parameter μ) to come back to operation. Thus, in the model, type 1 customers go to a second infinite server queue. This allows us to illustrate the method when the repair subsystem is more complex than a single queue.

To define the state space, we use n_1 (respectively n_2) to represent the number of machines of type 1 (respectively 2) in the repair queue; we denote by k the type of the machine being repaired (with value 0 if the repair station is empty) and by d be the phase of the current service with $1 \leq d \leq d_k$ ($d = 0$ if the repair station is empty). Denote by n_3 the number of machines of type 1 in transit (that is, in the delay station). We are interested in bounding the asymptotic availability of the system. Let

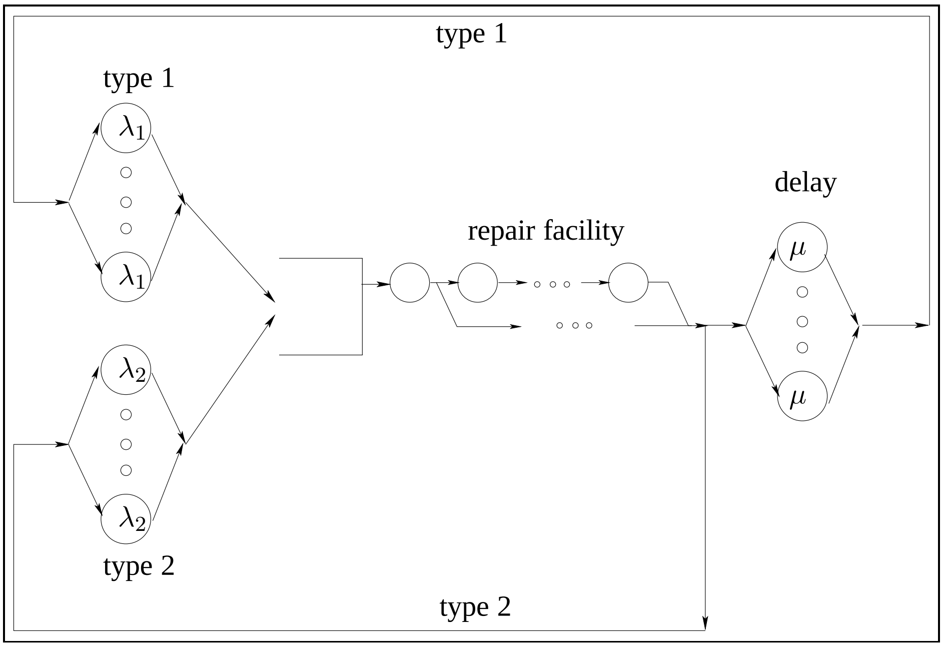

Figure 4: A Machine Repair Model

us assume that the system is operational as soon as there are at least $nMin_1$ machines of type 1 and at least $nMin_2$ machines of type 2 operating.

On states $s = (n_1, n_2, d, k, n_3)$ we have a Markov chain on which we consider the subsets of states C_I defined by

$$C_I = \{s \mid n_1 + n_2 + n_3 = I\}$$

with $0 \le I \le N_1 + N_2$. They define a partition of the state space. Observe that Condition 2 is not satisfied and thus that the method in [3] can not be applied. On the contrary, Condition 3 holds.

Let us consider the following parameter values:

- $N_1 = 80$, $N_2 = 120$, $nMin_1 = 79$, $nMin_2 = 115$, $\lambda_1 = 0.00004$, $\lambda_2 = 0.00003$, $m_1 = m_2 = 1.0$, $d_1 = 6$, $d_2 = 5$ and $\mu = 3.0$. The size of the whole state space is $|S| = 4344921$. Using two small values of K, we obtain the following numerical results:

K	$\lvert S^{\text{agg}} \rvert$	Lower bound	Upper bound
5	226	0.9997597121	0.9997597466
10	1826	0.9997597349	0.9997597349

- If we change the definition of operational system allowing 77 type 1 units as the threshold, that is, if we change $nMin_1$ to $nMin_1 = 77$, we have again a state space S with cardinality

$|S| = 4344921$ and for the same values of K we obtain

K	$\lvert S^{\text{agg}} \rvert$	Lower bound	Upper bound
5	226	0.9999999698	0.9999999852
10	1826	0.9999999841	0.9999999841

leading to a significant improvement of the availability of such a system.

As a technique to check the used software, let us consider the following situation. Let us keep the previous example with the values $N_1 = 80$ and $N_2 = 120$. We consider Coxian distributions for the repair times with 2 phases or stages (that is, $d_1 = d_2 = 2$) but we choose their parameters in such a way that they are equivalent to exponential service times: if $\nu_{k,d}$ is the parameter of the dth stage for a type k component and if l_d is the probability that phase d is the last one ($l_{d_k} = 1$), then for all phase d we have $\nu_{k,d} l_d = 1/m_k$ (technically, we put ourselves in a *strong lumpability* situation). Moreover, if the scheduling of the repair facility is changed to preemptive priorities, then type 2 units are invisible to type 1 ones, and with respect to type 1 components we have a product form queueing network. Standard algorithms can then be used to compute, for instance, the mean number of type 1 machines in the repair subsystem which we denote by $\bar{N}_1$. Using the QNAP2 product of Simulog, we obtain for $\lambda_1 = 0.00004$, $m_1 = 0.2$ and $\mu = 3.0$, the value $\bar{N}_1 = 0.0006404$. Using our algorithm with $K = 3$ we obtain $0.0006403 < \bar{N}_1 < 0.0006413$.

7.2 Bounding the mean number of customers in a two-node tandem

Consider the following simple open queuing network with 2 FIFO nodes (Figure 5). In this example, customers arrive from outside according to a Poisson process with rate λ. Each queue has infinite capacity and the service times at both nodes have the same Erlang distribution with 2-stages and expectation equal to $2/\nu$.

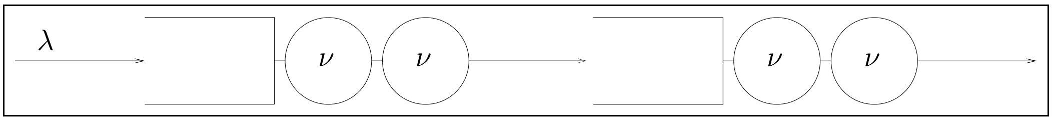

Figure 5: A two-node open queuing network

We consider the usual Markov representation of the state s of this queuing system, $s = (n_1, d_1, n_2, d_2)$, with n_i customers in node i and phase d_i in the server of node i, $i = 1, 2$, with the convention that

$d_i = 0$ if $n_i = 0$. Let us define the partition $(C_I)_{I \geq 0}$ of the state space with

$$C_I = \{(n_1, d_1, n_2, d_2) \mid n_1 + n_2 = I\}.$$

This model does not possess a known closed form solution and it is not possible to solve it directly due to its infinite state space cardinality. Moreover, given that Condition 2 is not satisfied, the method in [3] can not be applied. The matrix-geometric approach [11] can not be used neither, since the state space has a 2-dimensional structure and both dimensions are unbounded.

Let us apply our method to bound the mean number of customers in each node. In both cases, the a priori bound ϱ_2 on the rewards is infinite. First, we should note that λ is an upper bound of $\lambda^+_{I,I+1}$ for all I. Concerning the needed lower bound on the μ_I's, we use the regular structure of the C_I's. It is a matter of standard Markov analysis to verify that that for each C_I, $I \geq 1$, the value of μ^*_I is obtained for state $(I, 1, 0, 0)$ and that this value is the same for every I; it is given by

$$\mu^*_I = \mu^* = \frac{\lambda\nu^4}{(\lambda+\nu)^4 - \nu^4}.$$

If we set $\varrho = \lambda/\mu^*$, we have

$$\theta^*_I = \varrho^{I-K}, \text{ and } \nu^{'}_j = (1-\varrho)\mu^*_K.$$

Observe that this model is stable for $\lambda < \nu/2$.

For instance, if we set $\nu = 1.0$ and $\lambda = 0.18 < \mu^* \approx 0.1892$, we obtain that the mean number of customers in the network is equal to 1 with an error less than 0.01, using less than 200 generated states.

7.3 Bounding blocking probabilities in a three-node tandem

Let us consider the three-node open queuing network shown in Figure 6. Excepting node 1 which has an infinite capacity, the two other nodes have a finite capacity of respective sizes H_2 and H_3. Customers arrive at the first node according to a Poisson process with rate λ. All the services are exponentially distributed and the service rates are respectively μ_1, μ_2 and μ_3. We also assume that all the nodes implement a FIFO service discipline, and that there is a blocking-after-service behavior in the first and second nodes. This means that when the second node is saturated, the first one blocks its server after the end of the current service and until the departure of the customer being serviced in the second one. Then, if the latter is not saturated, simultaneously a customer leaves the second node to the third node and another customer passes from node 1 to node 2. On the other hand, if the last node is saturated, the second node as the first one blocks its server after the end of the current service and until the customer serviced in the third node leaves the system.

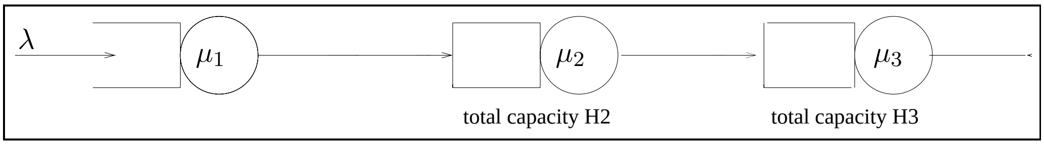

Figure 6: A three-node open queuing network with blocking-after-service

Once a Markov process is build in the usual way, we partition the state space as before, defining C_I as the set of states corresponding to I customers in the tandem. As in the previous example, there is no closed form solution [12] and the state space cardinality is infinite. Given the fact that queues 2 and 3 are bounded, it is easily verified that for $I \geq H_2 + H_3 + 1$, $|C_I|$ is constant. This allows us to simplify the analysis by choosing any $K > H_2 + H_3 + 1$ (note that if we had chosen some $K \leq H_2 + H_3 + 1$, we still could have done the same by collapsing in c all the aggregated states c_I with $I > H_2 + H_3 + 1$ and keeping states $c_K, \ldots, c_{H_2+H_3+1}$). The stability condition associated with the Markov chain used to bound the asymptotic measures is $\sum_{J=1}^{\infty} \theta_J^* < \infty$. Given that $K > H_2 + H_3 + 1$, for all $I \geq K$, $\mu_I^* = \mu_K^*$ and $\lambda_{I,I+1}^+ = \lambda$. So, letting $\varrho = \lambda/\mu_K^*$, we have

$$\theta_I^* = \left(\frac{\lambda}{\mu_K^*}\right)^{I-K} = \varrho^{I-K}, \qquad \nu_j^{'} = (1-\varrho)\mu_K^*.$$

To illustrate the technique, we bound the following measures: (i) the mean number of customers in the first node (this leads to a case with $\varrho_2 = \infty$), and (ii) the blocking probabilities in the first and second nodes (for these measures, we have $\varrho_2 < \infty$).

Let us consider the following parameter values: $\lambda = 0.2$, $\mu_1 = 0.7$, $\mu_2 = 1.5$, $\mu_3 = 0.2$, $H_2 = 18$ and $H_3 = 10$. After generating 3839 states, we obtain the value of the mean number of customers in the first node with an absolute error less than 10^{-10}, that is, the difference between the computed upper and lower bound is less than 10^{-10}. The mean number of customers in node 1 is 0.4000000000.

In the same way, we show respectively in Figure 7 and in Figure 8 the asymptotic probability of having servers 1 and 2 blocked. As above, given that we are interesting in performance measures with an absolute error less than 10^{-10}, we only plot the average between both computed bounds. The probability of blocking of node 1 is plotted as a function of the service rate of the corresponding server, μ_2, with $(H_2, H_3, \lambda, \mu_1, \mu_3) = (10, 8, 0.1224, 1.5, 0.5)$ and the probability of blocking of node 2 as a function of μ_3, with $(H1, H2, \lambda, \mu_1, \mu_2) = (18, 10, 0.1224, 1.5, 0.7)$.

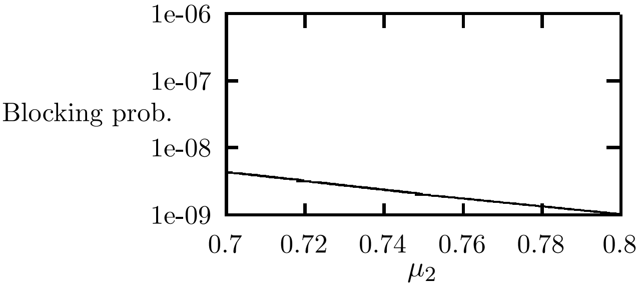

Figure 7: Probability of blocking of node 1 as a function of μ_2

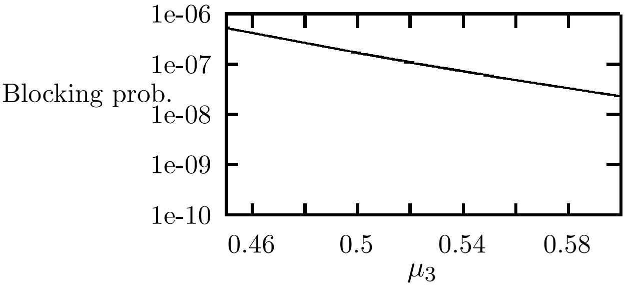

Figure 8: Probability of blocking of node 2 as a function of μ_3

8 Conclusions

This paper proposes a new way of obtaining upper and lower bounds of asymptotic performability measures, from finite or infinite Markov models. The asymptotic performability includes as particular cases the asymptotic availability in a dependability context, or standard asymptotic performance measures such as mean number of customers, blocking probabilities, loss probabilities, etc. To be applied, the method, as presented here, needs enough knowledge of the structure of the model in order to be able to derive analytically or to evaluate numerically certain values which are necessary to obtain the bounds. This is not always possible and current research aims to deal with this situation. In any case, there are many models similar to the type of infinite queuing networks used in this paper to illustrate the method, with no known closed solution and where, to the best of our knowledge, any other available bounding techniques do not apply.

Acknowledgments

The authors would like to thank J. Carrasco and B. Sericola for helpful discussions on different technical points.

References

[1] P.-J. Courtois and P. Semal. Bounds for the positive eigenvectors of nonnegative matrices and for their approximations by decomposition. *Journal of the Association for Computing Machinery*, 31(4):804–825, october 1984.

[2] P.-J. Courtois and P. Semal. Computable bounds for conditional steady-state probabilities in large markov chains and queueing models. *IEEE on Selected Areas in Communications*, 4(6):926–936, september 1986.

[3] R.R. Muntz, E. De Souza e Silva, and A. Goyal. Bounding availibility of repairable computer systems. *IEEE Transactions on Computers*, 38(12), december 1989.

[4] G. Rubino and B. Sericola. Sojourn times in finite markov process. *Journal on Applied Probability*, 27:744–756, July 1989.

[5] N.M. van Dijk. A simple bounding methodology for non-product form finite capacity queueing systems. In *Proc. of the 1st Int Workshop on Queueing Networks with Blocking*, May 1988.

[6] P. Semal. Refinable bounds for large markov chains. *IEEE Transactions on Computers*, 44(10):1216–1222, october 1995.

[7] R.R. Muntz and J.C.S. Lui. Evaluating bounds on steady-state availibility of repairable systems from markov models. In *Numerical solution of Markov chains*, pages 435–454, 1991.

[8] R.R. Muntz and J.C.S. Lui. Computing bounds on steady state availibility of repairable computer systems. *Journal of Association for Computing Machinery*, 41(4):676–707, july 1994.

[9] R.R. Muntz and J.C.S. Lui. Bounding the response time of a minimum expected delay routing system. *IEEE Transactions on Computers*, 44(5):1371–1382, december 1995.

[10] J. Carrasco. Improving availability bounds using the failure distance concept. In *Procs. of the DCCA–4 Int. Conf.*, january 1994.

[11] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models. An Algorithmic Approach.* The Johns Hopkins University Press, 1981.

[12] H.G. Perros. *Queueing Networks with Blocking*. Oxford University Press, 1994.

[13] P. Semal. *Analysis of large Markov models, bounding techniques and applications*. PhD thesis, University of Louvain, Belgium, 1992.

[14] P. Semal. Two bounding schemes for the steady-state solution of markov chains. *Numerical Solution of Markov Chains*, 1995.

[15] E. De Souza e Silva and P. Mejia. State space exploration in markov models. In *Sigmetrics/Performance*, pages 152–166, 1992.

[16] A. Graham. *Nonnegatige matrices and applicable topics in linear Algebra*. Ellis Horwood Limited, 1987.

[17] A. Berman and R.J. Plemmons. *Nonnegatige matrices in the Mathematical Sciences*. Classics in Appiled Mathematics, 1994.

[18] E. Cinlar. *Introduction to Stochastic Processes.* Prentice-Hall, Inc., 1975.

[19] T.G. Robertazzi. *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. Springer-Verlag, 1994.

[20] T.G. Robertazzi. Fast recursive solution of equilibrium state probabilities for three tandem queues with limited buffer space. *In Proc. of the 1994 Conference on Information Sciences and Systems*, pages 790–793, March 1994.

A Proof of Theorem 1

Let us consider the irreducible and aperiodic stochastic matrix P, obtained by uniformization of A with respect to the uniformization rate $\Lambda \geq \sup_i |A_{ii}|$, that is, matrix $P = I + A/\Lambda$. We have $\boldsymbol{\pi} = \boldsymbol{\pi} P$. Then, let us construct the matrix Q with the same size as P, such that $Q_{i,j} = 0$ for all $i \in C_K$ and $j \in in(G)$, and $Q_{i,j} = P_{i,j}$ in the other cases. Given that for all $i, j \in S$ we have $Q_{i,j} \leq P_{i,j}$, Q is a lower bound of P, that is a sub-stochastic matrix. We should note that Q is a strict lower bound of P because P is irreducible. Thus matrix $(I - Q)$ is invertible [16, Chapter 1, p. 66].

Let us denote by $\mathcal{P}(M)$ the polyhedron given by the set of convex combinations of the normalized rows of the square matrix M. Observe that if P is a stochastic matrix and if $\boldsymbol{\pi} = \boldsymbol{\pi} P$, then $\boldsymbol{\pi} \in \mathcal{P}(P)$. Consider

$$\mathcal{P}((I-Q)^{-1}) = \{\boldsymbol{v} \in \mathbb{R}_{1\times|S|} \mid \exists \boldsymbol{\beta} \in \mathbb{R}_{1\times|S|}, \boldsymbol{\beta}\mathbf{1}^{\mathrm{T}} = 1, \boldsymbol{v} = \boldsymbol{\beta}\Sigma^{-1}(I-Q)^{-1}\}, \tag{40}$$

where $\Sigma^{-1} = Diag((I-Q)^{-1}\mathbf{1}^{\mathrm{T}})^{-1} = Diag(\sigma_k)$ is the normalization matrix. We will denote by $\boldsymbol{z}^{(\mathbf{1})}, \ldots, \boldsymbol{z}^{(|\boldsymbol{S}|)}$ the vertices of $\mathcal{P}((I-Q)^{-1})$.

Proof of Theorem 1 As for P, we consider the uniformization of $A^{(j)}$, denoted by $P^{(j)}$. Because of the irreducibility of A, there exists an unique normalized vector $\boldsymbol{\pi}^{(j)}$ (Lemma 1), such that $\boldsymbol{\pi}^{(j)} A^{(j)} = 0$ and $\boldsymbol{\pi}^{(j)} P^{(j)} = \boldsymbol{\pi}^{(j)}$.

From [1, theorem 8], it follows that $\boldsymbol{\pi}^{(j)}$ belongs to $\mathcal{P}((I-Q)^{-1})$. We show that $\boldsymbol{\pi}^{(j)}$ is equal to the jth vertex of $\mathcal{P}((I-Q)^{-1})$.

Lemma 8 *The jth vertex of $\mathcal{P}((I-Q)^{-1})$ is $\boldsymbol{\pi}^{(j)}$.*

Proof. We prove first the existence of a stochastic matrix such that the jth vertex of $\mathcal{P}((I-Q)^{-1})$, denoted by $\boldsymbol{z}^{(\boldsymbol{j})}$, is its stationary distribution. Then, we show that this stochastic matrix is $P^{(j)}$. To start, let us introduce a new matrix $\Omega = Diag(\mathbf{1}(I-Q)^{-1})$. By definition of $\boldsymbol{z}^{(\boldsymbol{j})}$, there exists a normalized and positive vector $\boldsymbol{\beta}^{(\boldsymbol{j})}$ such that $\boldsymbol{z}^{(\boldsymbol{j})} = \boldsymbol{\beta}^{(\boldsymbol{j})}\Sigma^{-1}(I-Q)^{-1}$. Vector $\boldsymbol{\beta}^{(\boldsymbol{j})}$ is defined by $\beta_i^{(j)} = 1$ if $i = j$ and $\beta_i^{(j)} = 0$ in the other cases. Indeed $(I-Q)^{-1}$ has full rank. That means that each row of the matrix belongs to the base of the polyhedron. Let us consider the matrix $C^{(j)}$ equal to $\frac{1}{c^{(j)}}\Omega^{-1}\boldsymbol{\nu}^{\mathbf{T}}\boldsymbol{\beta}^{(\boldsymbol{j})}\Sigma^{-1}$, where $c^{(j)}$ is a constant equal to $\boldsymbol{\beta}^{(\boldsymbol{j})}\Sigma^{-1}(I-Q)^{-1}\Omega^{-1}\boldsymbol{\nu}^{\mathbf{T}}$ and $\boldsymbol{\nu}$ is a positive and normalized row vector.

Then we have the following relations:

$$\boldsymbol{z}^{(\boldsymbol{j})}(Q + C^{(j)}) = \boldsymbol{\beta}^{(\boldsymbol{j})}\Sigma^{-1}(I-Q)^{-1}(Q + C^{(j)}) = \boldsymbol{z}^{(\boldsymbol{j})}, \tag{41}$$

and if we denote $\boldsymbol{w}^{\mathbf{T}} = (I-Q)^{-1}\Omega^{-1}\boldsymbol{\nu}^{\mathbf{T}}$,

$$(Q + C^{(j)})\boldsymbol{w}^{\mathbf{T}} = (Q + C^{(j)})(I-Q)^{-1}\Omega^{-1}\boldsymbol{\nu}^{\mathbf{T}} = \boldsymbol{w}^{\mathbf{T}}. \tag{42}$$

From [17, Theorem 5.4], matrix $\frac{(Q+C^{(j)})}{\varrho(Q+C^{(j)})}$ is similar to a stochastic matrix. That means there exists a matrix T such that $\frac{(Q+C^{(j)})}{\varrho(Q+C^{(j)})} = T^{-1}BT$ with $B\mathbf{1}^{\mathrm{T}} = \mathbf{1}^{\mathrm{T}}$.

Using the same arguments as in the proof of theorem [1, 8], we have $Q + C^{(j)} = P^{(j)}$.

■

At this point, from lemma above we know that $\boldsymbol{\pi}^{(j)} = \boldsymbol{z}^{(\boldsymbol{j})} \in \mathcal{P}((I-Q)^{-1})$. To prove Theorem 1, we only have to show that $\boldsymbol{\pi}$ belongs to $\mathcal{P}((I-Q)^{-1})$. As for $\boldsymbol{\pi}^{(j)}$ (in the proof of [1, theorem 8]) and given that $\boldsymbol{\pi}(I-P)=0$ and $(P-Q)\geq 0$, there exists a normalized and positive vector $\boldsymbol{\beta}$ such that $\boldsymbol{\pi}=\boldsymbol{\beta}\Sigma^{-1}(I-Q)^{-1}$. It follows that $\boldsymbol{\pi}$ belongs to $\mathcal{P}((I-Q)^{-1})$ and $\boldsymbol{\pi}=\sum_{j\in S}\beta^{(j)}\boldsymbol{z}^{(j)}$. Moreover from the expression of $\boldsymbol{\pi}$ above, we have the following equality for $\boldsymbol{\beta}$: $\boldsymbol{\beta}=\boldsymbol{\pi}(P-Q)\Sigma$. Matrix $(P-Q)$ is equal to a matrix whose columns are null excepted those associated with the entry points in G. This means that the only non null elements of $\boldsymbol{\beta}$ are β_i for $i\in in(G)$. Then from Lemma 8, it follows that $\boldsymbol{\pi}=\sum_{j\in in(G)}\beta^{(j)}\boldsymbol{\pi}^{(j)}$. ■

B Proof of Theorem 3

Let us first consider the exact aggregated Markov chain constructed from X, by collapsing the subset G into a single state g, and each subset C_I for $I\geq K$ into a single state c_I. We denote the aggregated transition rate from g to any c_I, by $\lambda_{g,I}$, from any c_I to c_J (for $J>I$) by $\lambda_{I,J}$, from c_I to c_{I-1} by μ_I ($I>K$) and from c_K to g by μ_K. Now consider a second Markov chain having the same topology the first one, but such that the transition rates from c_I to c_J when $J\geq I$, denoted by $\lambda^{'}_{I,J}$, are upper bounds of the corresponding rates in the first aggregation, and lower bounds in case of transitions from c_I to c_{I-1} or from c_K to g (the respective values are denoted by $\mu^{'}_I$, $I\geq K$).

Let us denote by $\boldsymbol{v}$ (respectively by $\boldsymbol{v}'$) the stationary distribution of the first chain (respectively of the second). Then we have:

Lemma 9 $\boldsymbol{v}^{'}_g\leq\boldsymbol{v}_g$.

Proof. Suppose that $\boldsymbol{v}^{'}_g>\boldsymbol{v}_g$. From the equilibrium equations, we can write that:

$$\boldsymbol{v}_{c_K}=\frac{(\sum_{J\geq K}\lambda_{g,J})\boldsymbol{v}_g}{\mu_K},$$

$$\boldsymbol{v}^{'}_{c_K}=\frac{(\sum_{J\geq K}\lambda^{'}_{g,J})\boldsymbol{v}^{'}_g}{\mu^{'}_K}.$$

By definition, we know that $\lambda_{g,J}\leq\lambda^{'}_{g,J}$ and $\mu_K\leq\mu^{'}_K$. Thus it follows:

$$\boldsymbol{v}^{'}_{c_K}>\boldsymbol{v}_{c_K}.$$

Recursively, if we write for each $I>K$ the equilibrium equations, similar results are obtained. That is:

$$\forall I>K,\ \boldsymbol{v}^{'}_{c_I}>\boldsymbol{v}_{c_I}.$$

Then given that $\boldsymbol{v}$ is a stationary distribution, we have:

$$\boldsymbol{v}^{'}_g + \sum_{I \geq K} \boldsymbol{v}^{'}_{c_I} > 1$$

This means the first assumption is false. ■

In the same way, let us consider the two exact aggregated Markov chains constructed, as above, from $X^{(j)\mathrm{agg}}$ and from $Y^{(j)}$, by collapsing the subset G in a single state g. Given that there is a single point to enter in G from $\bar{G}$, the aggregated rate from c_K to g in former chain (respectively in the latter chain), is equal to the transition rate from c_k to j in $X^{(j)\mathrm{agg}}$ (respectively in $Y^{(j)}$). The transition rates from g to any c_I in the two new chains are given by the following expressions:

$$\lambda_{g,I} = \sum_{i \in G} \widehat{\pi}^{(j)agg}_{G,i} A^{(j)\mathrm{agg}}_{i,c_I},$$

$$\lambda^{'}_{g,I} = \sum_{i \in G} \widehat{y}^{(j)}_{G,i} A^{(j)\mathrm{agg}}_{i,c_I}.$$

By construction of $X^{(j)\mathrm{agg}}$ and $Y^{(j)}$, the restriction of their respective infinitesimal generators to the subset G are identical. From Theorem 2, we have the equality between the two conditional vectors $\widehat{\boldsymbol{\pi}}^{(j)agg}_G$, $\widehat{\mathbf{y}}^{(j)}_G$. That means that

$$\lambda_{g,I} = \lambda^{'}_{g,I}.$$

Moreover the transition rates between the other aggregated states are the same as in $X^{(j)\mathrm{agg}}$ and $Y^{(j)}$. Let us denote by $\pi^{(j)\mathrm{agg}}_g$ (respectively by $y^{(j)}_g$), the stationary probability of being in state g for the exact aggregated matrix obtained from $X^{(j)\mathrm{agg}}$ (respectively from $Y^{(j)}$). Thus, the conditions required to apply Lemma 9 are verified. It follows:

$$\pi^{(j)\mathrm{agg}}_g \geq y^{(j)}_g.$$

The fact that $\pi^{(j)\mathrm{agg}}_g = \pi^{(j)\mathrm{agg}}(G)$ and $y^{(j)}_g = y^{(j)}(G)$ (same results as (7)), ends the proof.

C The existence of $(I - Q)^{-1}$

To apply Theorem 1 in the infinite state space case, we assume first that the process is uniformizable. Next, observe that $X^{(j)}$ is also uniformizable and ergodic (recall that G is always finite). It only remains to show that matrix $(I-Q)$ is invertible. To do this, let us first recall the block structure of Q:

$$Q = \begin{pmatrix} P_G & P_{G\bar{G}} \\ 0 & P_{\bar{G}} \end{pmatrix}.$$

One way to obtain the existence of inverse of $(I-Q)$ is to consider each block separately. Given that the only infinite blocks in Q are $P_{G\bar{G}}$ and $P_{\bar{G}}$, we just have to show that $I - P_{\bar{G}}$ is invertible. This can be done by means of standard results as presented in [18]. Since $P_{\bar{G}}$ is a sub-stochastic matrix, the sum $I + P_{\bar{G}} + P_{\bar{G}}^2 + \cdots$ is finite and sub-stochastic. Using [18, 6.4.5], if $\boldsymbol{f}^{\mathbf{T}}$ denotes the vector defined by

$$f_i = \lim_{n\to\infty} \sum_{j\in\bar{G}} (P_{\bar{G}}^n)_{i,j},$$

we have that $\boldsymbol{f}^{\mathbf{T}}$ is the maximal solution of the linear system $\boldsymbol{x}^{\mathbf{T}} = P_{\bar{G}}\boldsymbol{x}^{\mathbf{T}}$, with $0 \le x_i \le 1$, and either $f_i = 0$ for all i, or $\sup_{i\in\bar{G}} f_i = 1$. Since X is irreducible and have only positive recurrent states, the system of linear equations above, have a unique solution verifying $f_i = 0$ for all i [18, 5.3.29]. This means that the matrix $P_{\bar{G}}$ verifies:

$$\lim_{n\to\infty} P_{\bar{G}}^n \mathbf{1}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}.$$

Given that all the elements of $P_{\bar{G}}$ are positive, it follows that

$$\lim_{n\to\infty} P_{\bar{G}}^n = 0,$$

which is a sufficient condition for the convergence of the series.