Available online at www.sciencedirect.com
ScienceDirect
Performance Evaluation 00 (2010) 1–22
Performance Evaluation

Approximating Passage Time Distributions in Queueing Models by Bayesian Expansion

Giuliano Casale

Imperial College London
Department of Computing
180 Queens Gate
London, SW7 2AZ

Abstract

We introduce Bayesian Expansion (BE), an approximate numerical technique for passage time distribution analysis in queueing networks. BE uses a class of Bayesian networks to approximate the exact joint probability density of the model by a product of conditional marginal probabilities that scales efficiently with the model size. We show that this naturally leads to decomposing a queueing network into a set of Markov processes that jointly approximate the dynamics of the model and from which passage times are easily computed.

Approximation accuracy of BE depends on the specific Bayesian network used to decompose the joint probability density. Hence, we propose a selection algorithm based on the Kullback-Leibler divergence to search for the Bayesian network that provides the most accurate results. Random models and case studies of increasing complexity show the significant accuracy gain of distribution estimates returned by BE compared to Markov and Chebyshev inequalities that are frequently used for percentile estimation in queueing networks.

Keywords: Passage time analysis, queueing network models, Bayesian networks

1. Introduction

Passage time distribution analysis is a fundamental tool for computing percentiles of response times and throughputs in queueing models. Performance metric percentiles are widely used in specifying quality-of-service requirements and service level agreements for IT services [1]. Existing methods for passage time analysis in general networks focus on numerical approximation and simulation because closed-form analytical expressions for passage time distributions exist only in special cases [4, 14]. Simulation-based estimates apply to general models, but they require many samples and repeated experiments to estimate distributions accurately. This is a

Email address: `g.casale@imperial.ac.uk` (Giuliano Casale)

limiting factor in what-if analysis, sensitivity studies, and in sizing studies based on constrained optimization that often require evaluating hundreds of thousands of possible system configurations [1]. Analytical percentile bounds based on Markov or Chebyshev inequalities are usually cheaper to evaluate than simulation [17], however they do not offer comparable accuracy for the distribution body, while they can be effective for tail estimates [11]. Exact theoretical formulas are accurate and computationally efficient [10, 12, 16, 5, 8, 19, 14, 15], but they apply only to special networks, often cyclic or tree-like topologies for which Laplace transform expressions of response time distributions are available.

To address the above limitations we introduce Bayesian Expansion (BE), a new approximate method for passage time distribution analysis in queueing networks with finite state space. The distinguishing feature of BE is that it approximates state probabilities driven by a class of Bayesian networks [18]. Bayesian networks have been recently applied in measurement-driven performance evaluation [23], however the present paper is (to the best of the author's knowledge) the first work that applies these models to the theoretical issues underlying queueing networks rather than to inference and learning from data sets. We show that a Bayesian network naturally defines an approximation of both transient and equilibrium state probability densities of a queueing network in terms of simpler marginal probabilities. This approximation has two main features: first, it tackles state space explosions issues; next, it naturally defines a decomposition of the queueing network into a set of aggregated models that can be used for passage time analysis. The idea of decomposing a performance model into a set of Markov processes has been considered in very few works in the literature [2, 7]; the fundamental innovation of BE is to introduce a Bayesian network that strives to maximize prediction accuracy based on information-theoretic techniques. This effectively approximates state probabilities and achieves low computational costs for passage time analysis.

The remainder of this paper is organized as follows. After introducing required definitions in Section 2, we describe BE in Section 3 and show that its accuracy can depend significantly on the choice of the Bayesian network used to approximate the joint probability density. We then develop in Section 4 a Bayesian network selection strategy based on the Kullback-Leibler divergence. Application of the BE technique to general queueing networks is discussed in Section 5. Finally, using case studies and random models of tractable size we show in Section 6 that BE approximates passage time distributions with accuracy that is often very close to exact results and significantly more accurate than Markov and Chebyshev inequalities that are widely used for percentile estimation.

2. Background

For simplicity of exposition, we focus on a class of closed queueing models composed by a set of M stations connected with an arbitrary topology network, however the BE methodology applies with minor changes also to other queueing networks with finite state space, e.g., an open network of queues with finite buffers. The closed queueing network is assumed to be populated by a finite set of N jobs. Upon completion from queue i, a job is routed to queue j with probability $p_{i,j}$. We illustrate the methodology on models where stations serve jobs according to a processor sharing (PS) or infinite server (IS) policy, we discuss generalization to other scheduling disciplines in Section 5. Service times are assumed to be exponentially distributed, extension to other service time distributions is also discussed in Section 5. We indicate with μ_i the mean service rate of station i; for load-dependent stations this is allowed to be a function $\mu_i(n_i)$ of

the number of jobs n_i currently in station i. Special cases are $\mu_i(n_i) = n_i\mu_i(1)$, where the load-dependent station is an infinite server, and $\mu_i(n_i) = \min(c_i, n_i)\mu_i(1)$, which is a multi-server station with $c_i \geq 1$ servers. Based on these assumptions, the Markov process underlying the queueing network has state space

$$\boldsymbol{S} = \{\boldsymbol{n} \,|\, \boldsymbol{n} \equiv (n_1, n_2, \ldots, n_M), \textstyle\sum_{i=1}^{M} n_i = N, n_i \geq 0\} \tag{1}$$

and infinitesimal generator Q specifying the rates of transition between pairs of states in $\boldsymbol{S} \times \boldsymbol{S}$. Define $\pi(\boldsymbol{n}) \equiv \pi(n_1, n_2, \ldots, n_M)$ to be the joint probability density that characterizes the equilibrium of the Markov process, and let $\boldsymbol{\pi} = \{\pi(n_1, n_2, \ldots, n_M)\}$ be the equilibrium distribution vector such that $\boldsymbol{\pi} = \boldsymbol{\pi}Q$, $\boldsymbol{\pi}\mathbf{1} = 1$, where $\mathbf{1}$ is a column vector of ones. The model enjoys the product-form solution

$$\pi(n_1, n_2, \ldots, n_M) = \frac{1}{G(N)} \prod_{i=1}^{M} \beta_i(n_i), \qquad \boldsymbol{n} \in \boldsymbol{S}, \tag{2}$$

where $\beta_i(n_i) = 1/\prod_{k=1}^{n_i} \mu_i(k)$, and $G(N)$ is a normalizing constant that can be computed efficiently using the convolution algorithm [3]. Marginal distributions are also easily computed by

$$\pi(n_i) = \beta(n_i)\frac{G_i(N - n_i)}{G(N)}, \qquad \pi(n_i, n_j) = \beta(n_i)\beta(n_j)\frac{G_{i,j}(N - n_i - n_j)}{G(N)} \tag{3}$$

where $G_i(N - n_i)$ is the normalizing constant of a queueing network having station i removed and a population of $N - n_i$ jobs, and similarly $G_{i,j}$ refers to a model with both stations i and j removed. From the properties of convolution it follows that the entire set of values in (3) can be computed efficiently in $O(N^2)$ for fixed number of queues, being N the total job population which is the main driver of computational costs in queueing networks. This is relevant for the application of (3) to BE discussed in Section 4.3.

2.1. Established Techniques for Passage Time Distribution Analysis

Given a Markov process initialized according to a probability distribution $\boldsymbol{\alpha}$ over a state space Σ, passage time analysis studies the distribution $F(t)$ of the time to first reach a subset of states $\Sigma^0 \subset \Sigma$ after the initial position. This problem models a number of common questions that arise in queueing theory such as computing job inter-arrival time percentiles or determining response time distributions. We here consider generalized passage time problems in which only events due to activation of a *subset* of rates of jump to Σ^0 is considered in the analysis. This allows to evaluate, for instance, inter-arrival times of jobs following a given route. In this case $F(t)$ is phase-type (PH-type) distributed with representation $(\boldsymbol{\alpha}, T)$, where T is obtained from the infinitesimal generator Q of the Markov process by setting to zero all off-diagonal rates in the considered subset that jump to states in Σ^0. We refer to T as a *subgenerator* of the Markov process, which is also called in the literature a PH-generator. For a PH-type distribution, the cumulative distribution function is readily computed as

$$F(t) = 1 - \boldsymbol{\alpha}e^{Tt}\mathbf{1},$$

where e^{Tt} is the matrix exponential function applied to Tt. Recall that the transient state $\boldsymbol{\pi}(t)$ of an absorbing Markov process with subgenerator T is computed by the Kolmogorov forward equations as

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)e^{Tt}.$$

It is then easy to see that passage time distribution analysis is equivalent to computing

$$F(t) = 1 - \boldsymbol{\pi}(t)\mathbf{1}, \qquad \boldsymbol{\pi}(0) = \boldsymbol{\alpha}, \tag{4}$$

hence it is a special case of transient analysis. From basic properties of the matrix exponential, it is also

$$\boldsymbol{\pi}(t) = \lim_{n\to\infty} \boldsymbol{\pi}(0)e^{T\Delta t_1}e^{T\Delta t_2}\cdots e^{T\Delta t_n}, \tag{5}$$

where $\Delta t_l = t_l - t_{l-1} = t/n$, $t_0 = 0$, and $t_n = t$. Thus, for a sufficiently large n one may accurately approximate the passage time solution by solving transient analysis problems over a sequence of small intervals. This approach is well-investigated in Markov process theory [21, 13], where it is often exploited that for $\Delta t_l \to 0$ it is

$$e^{T\Delta t_l} \approx I + T\Delta t_l + o({\Delta t_l}^2).$$

Therefore, for small Δt_l one may sequentially evaluate

$$\boldsymbol{\pi}(t_l) = \boldsymbol{\pi}(t_{l-1})(I + T\Delta t_l), \quad l = 1, \ldots, n, \tag{6}$$

in order to approximately integrate the Kolmogorov forward equations up to time t_n. Alternatively, one may use the uniformization technique on each interval Δt_l to account for higher-order terms in the expansion of the matrix exponential, thus achieving better accuracy at moderately increased costs [3].

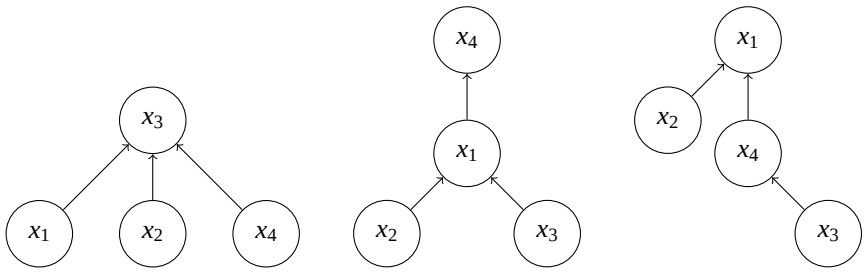
Finally, we observe that in queueing network models the matrix T is often prohibitively large, since its order is the same of the infinitesimal generator of the underlying Markov process. This order grows combinatorially with the number of queues and jobs in the network. Thus, the above formulas can be applied directly for passage time analysis only on small queueing network models. For our reference model with IS and PS servers, the order of T is small with respect to other types of queueing networks, nevertheless it is often hard to consider passage time problems with more than a few tens of jobs. In the next section, we develop the BE technique which tackles these limitations.

3. Bayesian Expansion

We now develop the proposed application of Bayesian networks to queueing network approximation. A Bayesian network is a probabilistic model used to characterize conditional dependencies between random variables. Given a set of random variables $n_1, n_2, \ldots, n_M$, conditional dependencies are expressed using a directed acyclic graph (DAG) such that each random variable n_i has one or more parent variables $\boldsymbol{n}_{\mathsf{par}(i)} = \{n_j, n_k, \ldots\}$, where $\mathsf{par}(i)$ maps i into the indexes of the parent variables for n_i. Thanks to the acyclic structure of the DAG, the joint equilibrium distribution of a Bayesian network admits a product-form expression

$$\pi(n_1, n_2, \ldots, n_M) = \prod_{i=1}^{M} \pi(n_i|\boldsymbol{n}_{\mathsf{par}(i)}), \tag{7}$$

where $\pi(n_i|\boldsymbol{n}_{\mathsf{par}(i)})$ is replaced by $\pi(n_i)$ if n_i is the *root node* of the DAG for which it is conventionally assumed $\mathsf{par}(i) = i$. Bayesian networks are often used in machine learning to approximate an empirical distribution $\pi(n_1, n_2, \ldots, n_M)$ that is too expensive to compute explicitly from a data



(a) BTa (b) BTb (c) BTc

Figure 1: Three possible Bayesian trees (BTs) for a joint probability density $\pi(x_1, x_2, x_3, x_4)$

set or too large to store in memory. The function $\mathsf{par}(\cdot)$ fully characterizes the Bayesian network and in the rest of the paper is referred to as its *dependence structure*.

Bayesian networks are here applied to the approximation of a queueing network equilibrium distribution $\pi(n_1, n_2, \ldots, n_M)$ with the aim of addressing state space explosion issues. We focus on a class of Bayesian networks with tree-like dependence structure, referred to as Bayesian Trees (BTs), where all random variables n_i have a single parent. A BT readily defines a *second-order product approximation*

$$\pi(n_1, n_2, \ldots, n_M) \approx \prod_{i=1}^{M} \pi(n_i|n_{\mathsf{par}(i)}) = \prod_{i=1}^{M} \frac{\pi(n_i, n_{\mathsf{par}(i)})}{\pi(n_{\mathsf{par}(i)})}, \qquad \pi(n_{\mathsf{par}(i)}) = \textstyle\sum_{n_i=0}^{N-n_{\mathsf{par}(i)}} \pi(n_i, n_{\mathsf{par}(i)}), \quad (8)$$

where for the root node we define $\pi(n_i|n_{\mathsf{par}(i)}) = \pi(n_i) = \sum_{n_j=0}^{N-n_i} \pi(n_j, n_i)$, for any $j \neq i$ such that $\mathsf{par}(j) = i$. Approximation (8) is computed using the marginal probabilities $\pi(n_i, n_{\mathsf{par}(i)})$ alone, thus it is computationally efficient since a joint state distribution with N^M values is approximated using only MN^2 marginal values. Techniques to define the dependence structure of the BN are discussed later in Section 4. Figure 1 gives examples of BTs approximating a joint probability density $\pi(x_1, x_2, x_3, x_4)$. The dependence structure of the BTs is as follows

	$par(\cdot)$ structure			
BT	x_1	x_2	x_3	x_4
a	3	3	3	3
b	4	1	1	4
c	1	1	4	1

Thus, e.g., the joint probability density is approximated by BTb as the second-order expression

$$\pi(x_1, x_2, x_3, x_4) \approx \pi(x_1|x_4)\pi(x_2|x_1)\pi(x_3|x_1)\pi(x_4). \quad (9)$$

3.1. Bayesian Expansion

We initially consider a queueing network observed in steady-state with equilibrium distribution $\pi(n_1, n_2, \ldots, n_M)$ and introduce Bayesian expansion (BE) by the following definition.

Definition 1 (Bayesian Expansion)**.** *A BT for the joint probability density $\pi(n_1, n_2, \ldots, n_M)$ of a queueing network defines a* Bayesian expansion *of the model, which is a set of M aggregated*

Markov processes, one for each station in the network. The aggregated process for station i has infinitesimal generator Q_i defined on the state space

$$\boldsymbol{S}_i = \{\boldsymbol{n} \,|\, \boldsymbol{n} \equiv (n_i, n_{\mathsf{par}(i)}, n_{\mathsf{rem}(i)}), n_{\mathsf{rem}(i)} = N - n_i - n_{\mathsf{par}(i)}\},$$

where $n_{\mathsf{par}(i)}$ is the parent variable of n_i in the BT. The aggregated process of the root node of the BT is defined to be identical to the aggregated Markov process of any of its child nodes.

The above definition states that the second-order product approximation (8) maps the Markov process underlying a queueing network into a family of simpler Markov processes with state space $(n_i, n_{\mathsf{par}(i)}, n_{\mathsf{rem}(i)}) \equiv (n_i, n_{\mathsf{par}(i)})$, where $n_{\mathsf{rem}(i)}$ is *dependent* on the other variables. Each aggregated process may be seen as a queueing network composed by station i, station $\mathsf{par}(i)$, and the aggregate server $\mathsf{rem}(i) = \{1, 2, \ldots, M\} \setminus \{i, \mathsf{par}(i)\}$ that describes the remainder of the network. Note that the $\mathsf{rem}(i)$ server is not necessarily a flow equivalent server since BE is later applied to the transient case where the Chandy-Herzog-Woo theorem does not hold [3]. Note also that BE differs from the methods in [2, 7] since it considers, thanks to the BT, just M aggregated processes instead of M^2. The key property of the BE is that the state of any queue in $\mathsf{rem}(i)$ can be estimated using a specialization of (8) as discussed in the next section.

3.2. Approximate Infinitesimal Generators

We now define the infinitesimal generators Q_i, $i = 1, \ldots, M$, of the aggregated Markov processes generated by BE. These infinitesimal generators are fundamental for passage time distribution analysis as described later in Section 3.3. Since the aggregated process for station i has state space $\boldsymbol{S}_i$, its underlying generator Q_i is immediately defined by the rates $q_i(\boldsymbol{n}, \boldsymbol{n}')$ connecting pairs of states $(\boldsymbol{n}, \boldsymbol{n}') \in \boldsymbol{S}_i \times \boldsymbol{S}_i$. The rates $q_i(\boldsymbol{n}, \boldsymbol{n}')$ are readily obtained by summing in each state $\boldsymbol{n}$ the departure rates from stations i, $\mathsf{par}(i)$, and $\mathsf{rem}(i)$ leading to state $\boldsymbol{n}'$. The departure rates from stations i and $\mathsf{par}(i)$ when they are busy are readily computed as service rates scaled by routing probabilities. Conversely, the departure rate from $\mathsf{rem}(i)$ is approximated by the aggregate throughput flowing out from the $\mathsf{rem}(i)$ subnetwork into a given station $k = 1, \ldots, M$. Conditioning on state $\boldsymbol{n} = (n_i, n_{\mathsf{par}(i)})$, the aggregate rate of departure from $\mathsf{rem}(i)$ to k may be written as[1]

$$\mu_{\mathsf{rem}(i),k}(n_i, n_{\mathsf{par}(i)}) = \sum_{j \in \mathsf{rem}(i)} \sum_{n_j=1}^{n_{\mathsf{rem}(i)}} \pi(n_j \,|\, n_i, n_{\mathsf{par}(i)}) \mu_j(n_j) p_{j,k} \tag{10}$$

where it is implicitly $n_{\mathsf{rem}(i)} = N - n_i - n_{\mathsf{par}(i)}$. The issue associated with (10) is that it cannot be computed directly from the BE. This is because $\pi(n_j \,|\, n_i, n_{\mathsf{par}(i)})$ involves three independent random variables, but BE considers only second-order marginal probabilities of the type $\pi(n_i, n_{\mathsf{par}(i)})$. To overcome this issue we introduce an approximation to estimate the departure rate from $\mathsf{rem}(i)$. We assume that when $p_{j,k} > 0$, $j \in \mathsf{rem}(i)$, the BT imposes at least one between

$$\mathsf{par}(j) = i \qquad \text{or} \qquad \mathsf{par}(j) = \mathsf{par}(i) \qquad \text{or} \qquad \mathsf{par}(\mathsf{par}(i)) = j \tag{11}$$

for all $i = 1, \ldots, M$, $i \neq j \neq \mathsf{par}(i)$. To justify the above conditions, let us observe first that then when the rates (10) are considered in the global balance equations of the aggregated Markov processes they define a summation over marginal probabilities $\pi(n_j, n_i, n_{\mathsf{par}(i)}) = \pi(n_j \,|\, n_i, n_{\mathsf{par}(i)}) \pi(n_i, n_{\mathsf{par}(i)})$.

[1] This expression is exact at equilibrium for the class of product-form networks assumed in Section 2. For more general models it may be considered as an approximation or reformulated with higher-order conditional probabilities involving more than three random variables. In the latter case also (12) should be extended to approximate such probabilities in the spirit of (8).

Hence, if the departure rate of $j \in \mathsf{rem}(i)$ is evaluated for some destination k, (11) assures that (8) can be applied to (10) in a computationally efficient manner as

$$\pi(n_j, n_i, n_{\mathsf{par}(i)}) \approx \pi(n_j \,|\, n_{\mathsf{par}(j)})\pi(n_i, n_{\mathsf{par}(i)})$$

using a single conditional probability term[2]. This is obvious if $\mathsf{par}(j) = i$ or $\mathsf{par}(j) = \mathsf{par}(i)$, while for the third condition in (11) one needs first to apply Bayes formula to $\pi(n_{\mathsf{par}(i)} \,|\, n_j)$ in order to compute $\pi(n_j \,|\, n_{\mathsf{par}(i)})$ and then use the latter in place of $\pi(n_j \,|\, n_{\mathsf{par}(j)})$. To avoid a complex notation, we limit to describe the approximation based on $\pi(n_j \,|\, n_{\mathsf{par}(j)})$.

Based on the above discussion, if (11) holds the departure rate for jobs leaving $\mathsf{rem}(i)$ may now be approximated as

$$\mu_{\mathsf{rem}(i),k}(n_i, n_{\mathsf{par}(i)}) \approx \sum_{j \in \mathsf{rem}(i)} \sum_{n_j=1}^{n_{\mathsf{rem}(i)}} \pi(n_j \,|\, n_{\mathsf{par}(j)}, n_j \leq n_{\mathsf{rem}(i)})\mu_j(n_j)p_{j,k}, \tag{12}$$

where

$$\pi(n_j \,|\, n_{\mathsf{par}(j)}, n_j \leq n_{\mathsf{rem}(i)}) = \frac{\pi(n_j, n_{\mathsf{par}(j)})}{\sum_{n_j=0}^{n_{\mathsf{rem}(i)}} \pi(n_j, n_{\mathsf{par}(j)})}. \tag{13}$$

Expression (13) leverages on our knowledge of $n_{\mathsf{rem}(i)}$ to exclude from the normalizing constant originally used in the conditioning of $\pi(n_j \,|\, n_{\mathsf{par}(j)})$ all events $n_j > n_{\mathsf{rem(i)}}$ which clearly cannot occur since $j \in \mathsf{rem(i)}$. This provides a tighter approximation of the departure rates compared to normalizing over all possible values of the n_j random variable. More importantly, equation (12) uses (13) to approximate the rates of the stations $\mathsf{rem}(i)$ using *only* marginal probabilities that are available in the BE which addresses the main limitation of (10). Summarizing, by assuming to use a BT that satisfies (11) we are now able to define a set of approximate generators $\widetilde{Q}_i$ based on (12).

3.3. BT-based Passage Time Distribution Analysis

The application of the definitions given in the previous section to passage time analysis requires to consider the transient state of the queueing network. Although BE extends in a straightforward way to transient joint densities $\pi(n_1, n_2, \ldots, n_M; t)$, i.e., by expressing the dependence on the time t in each of the marginal probabilities in (8), these joint densities are not usually known in advance, which makes it difficult to establish criteria for BT specification. From now on, we therefore assume to define the BT based on the equilibrium distribution $\boldsymbol{\pi}$ only.

For the sake of brevity, we also limit our discussion to state spaces of the type (1), however the approach extends with minor changes also to related state spaces, such as those obtained by extending (1) with the tagged job approach in order to determine response time distributions [16], see Section 5.

Using the equilibrium BT, we model the passage times in the queueing network using a set of non-homogeneous Markov processes with generators $\widetilde{Q}_i(t)$ defined similarly to the generators $\widetilde{Q}_i$ introduced in Section 3.2. Let us assume that a passage time distribution is iteratively evaluated over a discrete set of instants $t_0 = 0 < t_1 < \ldots < t_n$. As such, marginal probabilities at time

[2]For instance, if $\mathsf{par}(j) = w \neq i \neq j$, then one would need an extra summation on all possible values of the random variable n_w which increases significantly computational costs for defining the aggregated generators Q_i.

Algorithm 1
input: BT
input: initial probability distributions $\boldsymbol{\pi}_i(0)$, $1 \le i \le M$
input: observation instants $t_0 = 0 < t_1 < t_2 < \ldots < t_n$
input: reference station f
for $l = 1, \ldots, n$
  define the subgenerators $T_i(t_l)$, $1 \le i \le M$, from $\boldsymbol{\pi}_i(t_{l-1})$ using (14) and (15)
  determine $\boldsymbol{\pi}_i(t_l)$, $1 \le i \le M$, using (16) or uniformization
  determine $F(t_l) = 1 - \boldsymbol{\pi}_f(t_l)\mathbf{1}$
end
output: $F(t_l)$ for $l = 1, \ldots, n$

Figure 2: *Algorithm for BT-based Passage Time Distribution Analysis*

t_{l-1} have been all computed when starting the evaluation at time $t_l > t_{l-1}$. Defining the non-homogeneous generators at each of these instants involves two steps. First, we extend (12) to include the dependence on the current instant of time t_l as follows

$$\mu_{\mathsf{rem}(i),k}(n_i, n_{\mathsf{par}(i)}, t_l) \approx \sum_{j \in \mathsf{rem}(i)} \sum_{n_j=1}^{n_{\mathsf{rem}(i)}} \pi(n_j \,|\, n_{\mathsf{par}(j)}, n_j \le n_{\mathsf{rem}(i)}, t_{l-1}) \mu_j(n_j) p_{j,k}, \tag{14}$$

where all marginal probabilities are available from the solution for $t_{l-1} < t_l$ and

$$\pi(n_j \,|\, n_{\mathsf{par}(j)}, n_j \le n_{\mathsf{rem}(i)}, t_{l-1}) = \frac{\pi(n_j, n_{\mathsf{par}(j)}, t_{l-1})}{\sum_{n_j=0}^{n_{\mathsf{rem}(i)}} \pi(n_j, n_{\mathsf{par}(j)}, t_{l-1})} \tag{15}$$

being $\pi(n_j, n_{\mathsf{par}(j)}, t_{l-1})$ the estimated marginal joint density of queues j and $\mathsf{par}(j)$ at time t_{l-1}. Next, we define the non-homogeneous generators $\widetilde{Q}_i(t_l)$, $1 \le i \le M$, by replacing $\mu_{\mathsf{rem}(i),k}(n_i, n_{\mathsf{par}(i)})$ in $\widetilde{Q}_i$ with $\mu_{\mathsf{rem}(i),k}(n_i, n_{\mathsf{par}(i)}, t_l)$. Note that a similar approach may be used to extend (10) and specify a set of non-homogeneous generators $Q_i(t)$ that do not use approximation (14); thus, $\widetilde{Q}_i(t)$ may be seen as an approximation of such generators.

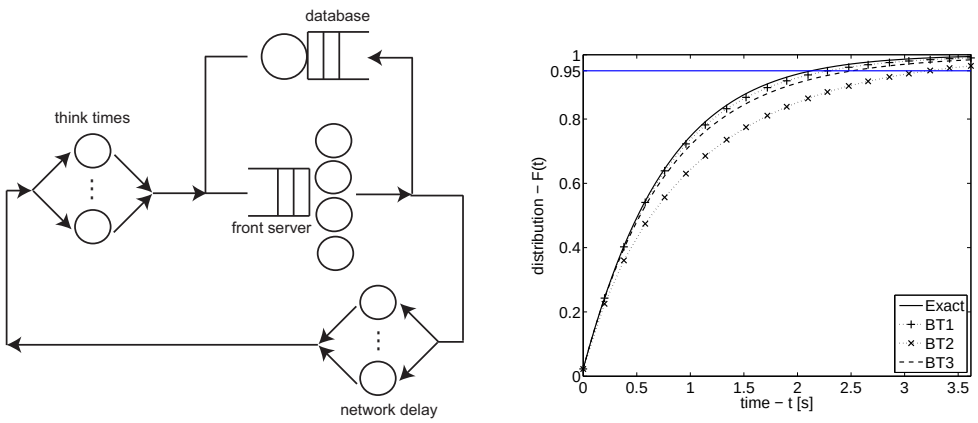
Finally, let the subgenerator $\widetilde{T}_i(t_l)$ be obtained from $\widetilde{Q}_i(t_l)$ by setting to zero the off-diagonal jump rates of interest to states in set Σ^0. We generalize (6) to the non-homogeneous case as

$$\boldsymbol{\pi}_i(t_l) = \boldsymbol{\pi}_i(t_{l-1})(I + \widetilde{T}_i(t_l)\Delta t_l), \tag{16}$$

for $i = 1, \ldots, M$ and $l = 1, \ldots, n$, where $\Delta t_l = t_l - t_{l-1}$. For example, given a reference station f, $1 \le f \le M$, used to compute throughput or response time distributions, (16) allows to compute the probability mass absorbed by states in Σ^0 at time t_l as $F(t_l) = 1 - \boldsymbol{\pi}_f(t_l)\mathbf{1}$. Alternatively, one may choose to average the probability mass absorbed over time in all aggregated Markov processes; numerical results of this approach are usually slightly worse than the ones obtained with the reference station technique, however averaging may be conceptually easier to define when studying passage times that involve the joint output of several stations. A pseudo-code summarizing the proposed passage time analysis algorithm is given in Figure 2. Similarly to standard passage time analysis described in Section 2.1, the uniformization technique can be used in place of (16) to obtain better accuracy at increased costs.

3.4. Illustrating Example

The example in this subsection illustrates a typical case where selecting different BTs to define the BE returns very different accuracy levels in the passage time distribution estimates.



(a) (b)

Figure 3: BE approximation for a multi-tier application model

This suggests that the accuracy of BE is conditional on the selection of a proper BT. To address this issue, we develop in the next section a greedy strategy that selects a BT to be used in BE with the aim of maximizing accuracy while satisfying computational cost constraints. When applied to the example in this section, such technique is able to select the BT that returns the most accurate approximation.

We consider a queueing network model of a multi-tier application. Figure 3(a) illustrates the model, which includes three load dependent stations, modeling think times (tt), network delays (nd), and a quad-core front server (fs). The front server is connected to a database back-end (db) modeled by a single server queue. Service rates are $\mu_{tt} = 0.1s^{-1}$, $\mu_{nd} = 1s^{-1}$, $\mu_{db} = 30s^{-1}$, and $\mu_{fs} = 20s^{-1}$. Routing probabilities at the departure from the front server are $p_{fs,nd} = 0.8$, $p_{fs,db} = 0.2$; the job population is $N = 15$.

	$par(\cdot)$ structure			
BT	n_{fs}	n_{nd}	n_{db}	n_{tt}
1	nd	nd	nd	nd
2	tt	tt	tt	tt
3	nd	nd	fs	nd

Table 1: Dependence structures in Figure 3. Rows indicate the parent of each variable in a BT.

The distribution $F(t)$ of a passage time variable t (in seconds) is observed at the output of the nd station for all jobs and estimated using time instant differing for a step of $\Delta t_l = 18ms$. Figure 3(b) compares approximate and exact results for the different choices of the BT given in Table 1 using the same tabular representation of par$(\cdot)$ introduced in Section 3. BT1 and BT2 are simple trees with a single level below the root; BT3 has instead two levels and a structure similar to the routing matrix of the network. Time for evaluation of a single point by BE is close to constant and on average equal to $158ms$ in all three cases. For BT1 the differences in the exact and approximate solutions are small, with minor deviations observed in approximation of the tail; these arise due to small errors in capturing the exact decay rate of the passage time density.

Conversely, although the dependence structure of BT2 is similar to BT1, a different choice of the root leads in this example to significantly underestimate the passage time distribution. The vertical gap of BT2 from the exact distribution is more than 10% of the probability mass, thus inevitably leading to significant deviations in percentile estimates on the horizontal axis. Specifically, the relative error in estimation of the 95th percentiles is 6.72% for BT1, 17.65% for BT2, and 50.42% for BT3, thus supporting our claim that the BT selection is of paramount importance for the approximation accuracy of BE. This motivates the investigation in the next section.

4. Bayesian Tree Specification

Stemming from the observation in Section 3.4 that different BTs provide different approximation accuracies for passage time distribution analysis, we focus in this section on the definition of a BT that can provide accurate results at the lowest possible computational costs. Specifically, the stated goal of this section is to explain how to specify a BT in order to satisfy the following set of requirements:

- the approximate subgenerators $\widetilde{T}_i(t)$ can be defined in $O(N^2)$ as the total population N grows. This is the minimum achievable complexity for a second-order product approximation. It is also a stronger requirement than asking for an aggregate state space that grows as $O(N^2)$ since the former includes the cost of computing the approximate transition rates in (14).
- in order to maximize accuracy, the approximate subgenerators $\widetilde{T}_i(t)$ should be as close as possible, with respect to some distance metric, to the subgenerators $T_i(t)$ that are defined from the aggregated Markov processes $Q_i(t)$.

The main findings and contributions of this section to achieve such goals are the following:

- we obtain in Section 4.1 a recursive expression to efficiently compute the approximate transition rates in (14); based on this result, we argue in Section 4.1 that it is always possible to define all subgenerators $\widetilde{T}_i(t)$ in $O(N^2)$;
- we then develop an automatic BT specification technique based on binary linear programming (BLP). The BLP expresses computational cost requirements as linear binary constraints (Section 4.1), whereas it models approximation accuracy using a binary linear objective function (Section 4.2);
- we argue in Section 4.3 that canonical BLP objective functions used in the literature for BT specification [9] do not account for the properties of BE. We therefore provide a new criterion for BT specification based on the conditional entropy metric [18].

4.1. Efficient Computation of Subgenerators

Let us define station $j \in \mathsf{rem}(i)$ as $O(1)$-*approximable* within the subgenerator $\widetilde{T}_i(t_l)$ if the ratio between the cost of estimating the departure rate from j to k in (14) and the cost of generating the aggregated state space is $O(1)$ under increasing populations. Thus, for a $O(1)$-approximable station, (14) imposes a fixed computational overhead for defining the subgenerators. Let us denote this departure rate by

$$\mu_{j,k}(n_{\mathsf{par}(j)}, n_{\mathsf{rem}(i)}, t_l) = \sum_{n_j=1}^{n_{\mathsf{rem}(i)}} \pi(n_j \mid n_{\mathsf{par}(j)}, n_j \leq n_{\mathsf{rem}(i)}, t_{l-1}) \mu_j(n_j) p_{j,k}, \tag{17}$$

such that from (14) it is

$$\mu_{\mathsf{rem}(i),k}(n_i, n_{\mathsf{par}(i)}, t_l) = \sum_{j \in \mathsf{rem}(i)} \mu_{j,k}(n_{\mathsf{par}(j)}, N - n_i - n_{\mathsf{par}(i)}, t_l).$$

Note that $O(1)$-approximability is non-trivial, since a naive implementation of (14) leads to a $O(N)$ overhead instead of $O(1)$. We first provide sufficient conditions for a station to be $O(1)$-approximable.

Proposition 1. *In a BT satisfying (11), all stations are* $O(1)$-approximable *within all subgenerators where they belong to* $\mathsf{rem}(i)$ *since the service rates admit the following recursive expression*

$$\mu_{j,k}(n_{\mathsf{par}(j)}, n_{\mathsf{rem}(i)}, t_l) = \gamma_j(n_{\mathsf{rem}(i)})\mu_{j,k}(n_{\mathsf{par}(j)}, n_{\mathsf{rem}(i)} - 1, t_l) + \pi(n_j = n_{\mathsf{rem(i)}} | n_{\mathsf{par(j)}}, t_{l-1})\mu_j(n_{\mathsf{rem(i)}})p_{j,k}$$

with termination condition $\mu_{j,k}(n_{\mathsf{par}(j)}, 0, t_l) = 0$*, and where*

$$\gamma_j(n_{\mathsf{rem}(i)}) = \frac{\sum_{n'_j=0}^{n_{\mathsf{rem}(i)}-1} \pi(n'_j, n_{\mathsf{par}(j)}, t_{l-1})}{\sum_{n'_j=0}^{n_{\mathsf{rem}(i)}} \pi(n'_j, n_{\mathsf{par}(j)}, t_{l-1})}.$$

This recursive expression allows to compute all rates in $O(N^2)$ prior *to defining the subgenerators* $\widetilde{T}_i$ *and thus keeps the cost for their definition to* $O(N^2)$*.*

Proof. We rewrite (17) as

$$\mu_{j,k}(n_{\mathsf{par}(j)}, n_{\mathsf{rem}(i)}, t_l) = \sum_{n_j=1}^{n_{\mathsf{rem(i)}}-1} \pi(n_j | n_{\mathsf{par(j)}}, n_j \le n_{\mathsf{rem(i)}}, t_{l-1})\mu_j(n_j)p_{j,k} + \pi(n_j = n_{\mathsf{rem(i)}} | n_{\mathsf{par(j)}}, t_{l-1})\mu_j(n_{\mathsf{rem(i)}})p_{j,k},$$

and the result follows immediately by observing that

$$\begin{aligned}\sum_{n_j=1}^{n_{\mathsf{rem(i)}}-1} \pi(n_j | n_{\mathsf{par(j)}}, n_j \le n_{\mathsf{rem(i)}}, t_{l-1})\mu_j(n_j)p_{j,k} &= \sum_{n_j=1}^{n_{\mathsf{rem(i)}}-1} \frac{\pi(n_j, n_{\mathsf{par(j)}}, t_{l-1})}{\sum_{n'_j=0}^{n_{\mathsf{rem}(i)}} \pi(n'_j, n_{\mathsf{par}(j)}, t_{l-1})}\mu_j(n_j)p_{j,k} \\ &= \gamma_j(n_{\mathsf{rem}(i)}) \sum_{n_j=1}^{n_{\mathsf{rem(i)}}-1} \frac{\pi(n_j, n_{\mathsf{par(j)}}, t_{l-1})}{\sum_{n'_j=0}^{n_{\mathsf{rem}(i)}-1} \pi(n'_j, n_{\mathsf{par}(j)}, t_{l-1})}\mu_j(n_j)p_{j,k} \\ &= \gamma_j(n_{\mathsf{rem}(i)})\mu_{j,k}(n_{\mathsf{par}(j)}, n_{\mathsf{rem}(i)} - 1, t_l)\mu_j(n_j)p_{j,k}.\end{aligned}$$

□

The above theorem shows that, for a BT satisfying (11) such that (14) holds, it is always possible to define all subgenerators $\widetilde{T}_i(t_k)$ in $O(N^2)$. We now show that a BT satisfying (11) always exists.

Proposition 2. *There exist at least M BTs such that the corresponding BE has stations that are* $O(1)$*-approximable in all subgenerators where they belong to* $\mathsf{rem}(i)$*.*

Proof. Consider a BT where there exists a random variable n_w that acts as root node for all other random variables, i.e., $\mathsf{par}(i) = w$ for $1 \le i \le M$. Clearly, there exist M of such BTs depending on the choice of $w = 1, \ldots, M$. It readily follows that, for each of these, (11) is always satisfied because $\mathsf{par}(j) = \mathsf{par}(i) = w$ in all aggregated Markov processes. □

The last result guarantees the existence of at least M BTs satisfying (11) and this allows an efficient definition of the subgenerators by Proposition 1. However, the set of BTs satisfying (11) often includes also other BTs in addition to these M. In fact, a BT which defines a BE where all stations are $O(1)$-approximable satisfies the following set of constraints

$$\sum_{i=1}^{M} e_{i,i} = 1, \tag{18}$$

$$\sum_{j=1}^{M} e_{i,j} = 1, \qquad i = 1, \ldots, M; \tag{19}$$

$$p_{j,k} e_{i,w} \leq e_{j,i} + e_{j,w} + e_{w,j}, \qquad i, j, k, w = 1, \ldots, M; i \neq j \neq w; \tag{20}$$

where the binary variables $e_{i,j} \in \{0, 1\}$ uniquely define the dependence structure of the BT as

$$e_{i,j} = \begin{cases} 1, & \text{if } \mathsf{par}(i) = j, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

for $i, j = 1, \ldots, M$. The first two conditions guarantee that the BT is feasible. In fact, constraint (18) assures that there is a single root node. Note that if the root is selected among identical stations the outcome is a random decision of the BLP solver. Instead, (19) guarantees that each random variable has a single parent node. The last condition (20) characterizes the class of BTs where all stations are $O(1)$-approximable. Note that $e_{w,i}$ does not appear in it since the constraint is always satisfied if $e_{i,w} = 0$, otherwise $e_{i,w} = 1$ implies trivially $e_{w,i} = 0$. These conditions formalize (11), where $w \equiv \mathsf{par}(i)$, which enforces that a single conditional probability is needed for approximation (14); this in turn assures $O(1)$-approximability by Proposition 1. Note that the destination variable k ranges across $1, \ldots, M$ instead of $\{i, \mathsf{par}(i)\}$ because in some passage time analysis problems, noticeably in response time distribution analysis, the approximation requires to detail also inner transitions within $\mathsf{rem}(\mathrm{i})$, see Section 5. Whenever such transitions are not needed, one may instead limit to be either i or $w \equiv \mathsf{par}(i)$.

The above characterization uniquely identifies the set of BTs where all stations are $O(1)$-approximable. The next subsection explains how to use the above characterization to select a BT within this set that gives accurate approximation results.

4.2. Approximation Accuracy

Consider a set of weights $w_{i,j}$ associated to the variables $e_{i,j}$ that define the $O(1)$-approximability constraints such that the objective function $f = \sum_{i,j} w_{i,j} e_{i,j}$ quantifies the reward of choosing a specific BT for BE. We can express the problem of selecting the best BT structure as the problem of maximizing f; the challenge is to define a set of weights that may be representative of BE approximation accuracy. Based on these approach, we readily obtain the BT by solving, a BLP subject to the $O(1)$-approximability constraints and where the objective function f quantifies the relative merit of choosing a particular BT. Note that since BLP is NP-hard in the general case, approximate methods may be used for its solution on large instances. We first discuss the critical definition of the weights $w_{i,j}$ in the BLP objective function. Define $I(n_i, n_j)$ to be the *mutual information* [18] of random variables n_i and n_j, i.e.,

$$I(n_i, n_j) = \sum_{n_i=0}^{N} \sum_{n_j=0}^{N-n_i} \pi(n_i, n_j) \log \frac{\pi(n_i, n_j)}{\pi(n_i)\pi(n_j)} \geq 0. \tag{22}$$

A classic result obtained by Chow and Liu in [18] is that if $w_{i,j} = I(n_i, n_j)$ then the BT with maximum weight $f_{max} = \max \sum_{i,j} w_{i,j} e_{i,j}$ has equilibrium (8) which gives the closest approximation of the original distribution $\pi(n_1, n_2, \ldots, n_M)$. A Chow-Liu BT can be easily determined by computing the maximum weight spanning tree defined over the complete graph with edges weighted by $w_{i,j}$, which is computed in $O(M^2 \log M)$ by Kruskal's algorithm. In the classic result of Chow and Liu, distance from the optimal solution uses the Kullback-Leibler divergence [18]. For a probability distribution $P(n_1, n_2, \ldots, n_M)$ and an approximation model $P^a(n_1, n_2, \ldots, n_M)$, this is defined as

$$\mathcal{D}(P \,\|\, P^a) = \sum_{(n_1, n_2, \ldots, n_M)} P(n_1, n_2, \ldots, n_M) \log \frac{P(n_1, n_2, \ldots, n_M)}{P^a(n_1, n_2, \ldots, n_M)}, \tag{23}$$

which may be interpreted as the average surprise, in an information-theoretic sense, resulting from comparing an empirical distribution with its approximation.

Based on the optimality result of Chow and Liu, one would expect a Chow-Liu BT to be optimal also for queueing network approximation. However, this is not often the case, for instance BT2 in Figure 3 is an example of Chow-Liu BT that is clearly suboptimal compared to the other BTs. Indeed, since passage time analysis with BE involves several approximations not necessarily limited to the specification of the BT, it is not easy to assess the role of each of these (and possibly their mutual interactions) in determining the final approximation accuracy. The interpretation we propose to explain the reduced effectiveness of Chow-Liu BTs in BE is that the effect of (8) at the level of the aggregated Markov processes is reflected mainly in the error of (14) relatively to the transient version of (10), rather than at the level of the joint probability distribution. That is, while Chow and Liu minimize the divergence

$$\mathcal{D}^{cl} = \mathcal{D}(\pi(n_1, n_2, \ldots, n_M) \,\|\, \textstyle\prod_{i=1}^{M} \pi(n_i | n_{\mathsf{par}(i)})), \tag{24}$$

our interpretation considers the following divergence more relevant for the accurate definition of the subgenerators in BE

$$\mathcal{D}^{be} = \sum_{i=1}^{M} \sum_{j \in \mathsf{rem}(i)} \mathcal{D}^{be}_{i,j}, \qquad \mathcal{D}^{be}_{i,j} = \mathcal{D}(\pi(n_i, n_{\mathsf{par}(i)}, n_j) \,\|\, \pi(n_j | n_{\mathsf{par}(j)}) \pi(n_i, n_{\mathsf{par}(i)})), \tag{25}$$

where we consider the equilibrium distributions since we have assumed in Section 3.3 to specify equilibrium BTs only. This interpretation is clearly a simplification, for example it ignores the approximation error due to assuming that (10) holds also in a transient regime, which is the implicit assumption driving the definition of (14). Unfortunately, the lack of analytic results for transient analysis of queueing networks makes it hard to verify exactly the validity of such interpretations, thus we limit to show in Section 6 the increased accuracy of the BTs defined with the proposed approach.

4.3. Greedy Selection Strategy

We are now in condition to propose a weighting scheme for the selection of the BT that drives the queueing network approximation in BE based on (25). Let us begin by defining the *joint entropy* of a set of V random variables

$$H(x_1, \ldots, x_V) = -\sum_{\boldsymbol{x}} \pi(x_1, \ldots, x_V) \log \pi(x_1, \ldots, x_V), \tag{26}$$

Algorithm 2
input: number of queues M
input: marginal distributions $\pi(n_i, n_j)$ for all pair of stations (i, j)
for $i = 1, \ldots, M$
  for $j = 1, \ldots, M$
    compute $w_{i,j} = H(n_j|n_i)$
  end
end
solve $f_{be} = \min \sum_{i,j} w_{i,j} e_{i,j}$ subject to (18)-(20) and $w_{i,j} = H(n_j|n_i)$.
define BT based on $e_{i,j}$ values according to (21)
output: BT

Figure 4: *Bayesian tree selection algorithm*

which for $V = 1$ is the usual entropy $H(x) = -\sum_x \pi(x) \log \pi(x)$. We can write the Kullback-Leibler divergence $\mathcal{D}^{be}_{i,j}$ as

$$\mathcal{D}^{be}_{i,j} = \sum_{(n_i, n_{\mathsf{par}(i)}, n_j)} \pi(n_i, n_{\mathsf{par}(i)}, n_j) \log \frac{\pi(n_i, n_{\mathsf{par}(i)}, n_j)\pi(n_{\mathsf{par}(j)})}{\pi(n_i, n_{\mathsf{par}(i)})\pi(n_j, n_{\mathsf{par}(j)})} \tag{27}$$

which yields

$$\begin{aligned}\mathcal{D}^{be}_{i,j} = &\sum_{(n_i, n_{\mathsf{par}(i)}, n_j)} \pi(n_i, n_{\mathsf{par}(i)}, n_j) \log \pi(n_i, n_{\mathsf{par}(i)}, n_j) + \sum_{(n_i, n_{\mathsf{par}(i)}, n_j)} \pi(n_i, n_{\mathsf{par}(i)}, n_j) \log \pi(n_{\mathsf{par}(j)}) \\ &- \sum_{(n_i, n_{\mathsf{par}(i)}, n_j)} \pi(n_i, n_{\mathsf{par}(i)}, n_j) \log \pi(n_i, n_{\mathsf{par}(i)}) - \sum_{(n_i, n_{\mathsf{par}(i)}, n_j)} \pi(n_i, n_{\mathsf{par}(i)}, n_j) \log \pi(n_j, n_{\mathsf{par}(j)}).\end{aligned} \tag{28}$$

Since in the last three terms the logarithm is affected just by a subset of variables of $(n_i, n_{\mathsf{par}(i)}, n_j)$, we can sum the inner $\pi(n_i, n_{\mathsf{par}(i)}, n_j)$ terms on the other variables and rewrite the expression as

$$\mathcal{D}^{be}_{i,j} = -H(n_i, n_{\mathsf{par}(i)}, n_j) - H(n_{\mathsf{par}(j)}) + H(n_i, n_{\mathsf{par}(i)}) + H(n_j, n_{\mathsf{par}(j)}), \tag{29}$$

where all terms in (28) are now interpreted as joint entropies. We can then observe that the evaluation of this divergence would require a computational cost for $H(n_i, n_{\mathsf{par}(i)}, n_j)$ that grows cubically with the population, hence approximations are needed to study $\mathcal{D}^{be}_{i,j}$ in order to keep the complexity quadratic.

We propose a BT selection strategy where we evaluate the effect of adding j with parent $\mathsf{par}(j)$ to the BT by assuming that the parent of previously added nodes is fixed. This is equivalent to assigning j and $\mathsf{par}(j)$ with a greedy algorithm that tries to perform the best decision for the current set of choices assuming that past decisions are optimal. Under this greedy approach, $H(n_i, n_{\mathsf{par}(i)}, n_j)$ and $H(n_i, n_{\mathsf{par}(i)})$ are unaffected by the next decision on $\mathsf{par}(j)$, because the latter entropy is fixed, while the former appears whenever $p_{j,i} > 0$ and does not depend on the choice of $\mathsf{par}(j)$. Therefore, it follows that the greedy decision affects only the difference

$$\Delta(\mathsf{par}(j)) = H(n_j, n_{\mathsf{par}(j)}) - H(n_{\mathsf{par}(j)}). \tag{30}$$

Thus, by minimizing $\Delta(\mathsf{par}(j))$ we obtain an approximate algorithm for minimization of $\mathcal{D}^{be}$. However, $\Delta(\mathsf{par}(j))$ is easily shown to be the *conditional entropy*

$$H(n_j \,|\, n_{\mathsf{par}(j)}) = -\sum_{n_j=0}^{N} \sum_{n_{\mathsf{par}(j)}=0}^{N-n_j} \pi(n_j, n_{\mathsf{par}(j)}) \log \pi(n_j|n_{\mathsf{par}(j)}). \tag{31}$$

Based on this result, we propose to define the BT that drives the queueing network approximation by the dependence tree that *minimizes* the sum of weights $w_{i,j} = H(n_j \,|\, n_i)$ subject to the $O(1)$-approximability constraints, see the pseudo-code in Figure 4. An important aspect connected to this definition is that the conditional entropy can be computed efficiently in $O(N^2)$ by (3), thus this greedy approach is compatible with the computational costs targeted by BE. In this approach, $w_{i,j}$ is interpreted as the change in the divergence $\mathcal{D}^{be}_{i,j}$ if n_j is added to the BT with variable n_i as parent. Experimental results provided in Section 6 show the large improvement of approximation accuracy of the conditional entropy selection compared to the Chow-Liu approach.

5. Generalization of Bayesian Expansion

We now discuss the ability of BE to generalize to models which do not satisfy some of the assumptions taken in Section 2.

5.1. Tagged Job Approach and Response Time Distribution Analysis

Another standard approach for passage time analysis, alternative to the calculation of $F(t)$ by (4), is the use of state spaces defined with the tagged job. Such spaces include an auxiliary random variable k_{tag}, $1 \leq k_{tag} \leq M$, which describes the current position of a test customer that cycles into the network of queues [16]. In such models, the equilibrium probability for the underlying Markov process has therefore the form $\pi(n_1, n_2, \ldots, n_M, k_{tag})$. The tagged job approach is useful to compute response times at the level of individual queues (sojourn times) and at the level of the network (cycle time) using passage time analysis. Cycle times are obtained by initializing the process in the equilibrium state seen upon arrival by a job to a reference queue f and then defining Σ_0 as the set of states where the tagged job re-enters f following its departure from a queue $i \neq f$. Sojourn times have a similar initialization, but Σ_0 is defined as the set of states reached immediately after the tagged job leaves f.

BE can be readily generalized to support the tagged job approach by the approximation

$$\pi(n_1, n_2, \ldots, n_M, k_{tag}) \approx \prod_{i=1}^{M} \pi(n_i | n_{\mathsf{par}(i)}, k_{tag}), \tag{32}$$

such that each aggregated Markov process includes the auxiliary variable k_{tag} that provides the joint probability of observing the tagged job in a certain position while the active state in the aggregated process is $(n_i, n_{par(i)})$. This generalization increases moderately the state space sizes of the aggregated processes by a factor of M. Nonetheless, this cannot be avoided, at least without increasing approximation errors, because replication of the k_{tag} variable is needed to determine the rate of transition of the tagged job from $k_{tag} \in \mathsf{rem}(i)$ into stations i or $\mathsf{par}(i)$ for each state $(n_i, n_{par(i)})$. Note also that the movements of the tagged job *within* $\mathsf{rem}(i)$ that change the k_{tag} value should be accounted for in the generators; such rates are easily computed with an expression equivalent to (14) but defined using the transient probabilities of the aggregated process for station k_{tag}.

5.2. Scheduling Disciplines

Auxiliary random variables may also be needed to represent internal details of scheduling disciplines different from IS or PS. For example, in order to compute cycle times in a queueing network with one or more FCFS stations, one needs to track explicitly the relative position of the tagged job during its residence time in a FCFS buffer. Thus, the equilibrium probability in these

models takes the form $\pi(n_1, n_2, \ldots, n_M, k_{tag}, k_{fcfs})$, where the buffer position k_{fcfs}, $0 \leq k_{fcfs} \leq N$, is zero when k_{tag} is not a FCFS station. Similarly to (32), the auxiliary variable k_{fcfs} should be replicated in the state spaces of the aggregated Markov processes defined by BE.

Indeed, for models with FCFS stations the application of BE would require a greater computational effort than for the IS/PS case, since the state space size would be up to N times larger. However, this appears to be an intrinsic limitation common to numerical techniques that evaluate directly the Markov process underlying the queueing network, rather than originating from specific aspects of the BE approach. We expect similar issues to arise also with other scheduling disciplines different from FCFS depending on the range of variability of the auxiliary random variables introduced by these disciplines.

5.3. Non-Product-Form Models

It appears possible to generalize the BE technique at least to some classes of non-product-form models. While this is still an open subject for investigation, a number of general considerations can be already drawn on the feasibility of this extension.

The product-form assumptions are used in BE approximation as follows: (i) for the efficient evaluation of the conditional entropies in the pseudo-code in Figure 4 by means of (3); (ii) for the definition of the initial probability vector $\boldsymbol{\pi}_i(0)$, which we assume to be the equilibrium distribution of Q_i seen immediately after activation of a state into the Σ_0 set assuming that the process is not stopped upon entering this set. For closed product-form models, when this event is associated to a job arrival or departure the distribution $\boldsymbol{\pi}_i(0)$ is known to be the equilibrium state distribution $\boldsymbol{\pi}$ of an equivalent model having a population with a job less [20].

Extensions of BE to non-product-form models require the definition of suitable replacements for the probabilities in (i) and (ii) . However, due to the lack of exact closed-form expressions these quantities should be computed by solving the Markov process underlying the queueing network. Therefore, these additional computational costs should be considered prior to starting an analysis with BE.

Furthermore, non-product-form models should be distinguished into two classes. A first class includes non-product-form networks which can be studied with reasonable effort directly at the Markov process level, such as queueing networks supporting RS-RD blocking, limited forms of state-dependence, or including queues with PH-type or MAP service processes [7]. BE is expected to generalize to these models, although the maximum job population that can be analyzed could be smaller than for the product-form case on some instances. A second class is instead formed by models for which the BE state space aggregation would still lead to extremely large state spaces, also accounting for the auxiliary variables. For instance, networks with multiple queues supporting BAS blocking are often intractable at the Markov process level due to the rapid combinatorial growth of precedence conditions for unblocking. Such models would be probably difficult to study also using the BE approach.

6. Numerical Validation

This section reports experiments on random models and case studies of increasing complexity that illustrate the accuracy of the BE approach to passage time distribution analysis. BE is compared with Markov and Chebyshev inequalities that are often used in the literature for percentile estimation [17]. For a passage time metric Y having mean $E[Y]$ and variance $Var[Y]$,

	model		Cond. Ent. BT		Chow-Liu BT		Markov ineq.		Chebyshev	
metric	N	M	X	R	X	R	X	R	X	R
Δ_{cdf}	5	4	1.2	3.0	1.5	4.0	30.1	30.0	20.1	20.0
Δ_{cdf}	10	4	0.9	1.8	0.8	3.7	29.8	29.5	20.1	20.0
Δ_{pct}	5	4	8.4	15.8	14.5	20.1	565.4	565.8	79.0	78.7
Δ_{pct}	10	4	3.0	8.6	11.7	19.1	558.5	557.4	79.4	79.3

Table 2: Error analysis of 95th percentile estimation. Experimental results on random models with N jobs and M queues (1.0 ≡ 1% error). X are inter-arrival times, R are network cycle times.

Markov inequality provides the bound

$$F_Y(t) \geq 1 - \frac{E[Y]}{t} \tag{33}$$

whereas Chebyshev inequality is given by [11]

$$F_Y(t) \geq 1 - \frac{Var[Y]}{(t - E[Y])^2}. \tag{34}$$

Throughout this section, we consider two passage time metrics: X is a random variable representing the job inter-arrival times at a reference station, R represents network cycle times. Exact values of X and R are obtained using exact numerical methods for small and medium-sized models and estimated with long-run simulations (10 million samples) for large models.

6.1. Random Models of Tractable Size

Due to the high costs of exactly computing passage times by numerical techniques, for the random model evaluation we consider models with tractable state space size. These models have $M = 4$ queues, $N = \{5, 10\}$ jobs, and random topology. With these parameters, the exact evaluation of passage times in a single model requires between 5 and 15 minutes on an Intel Core Duo 2.16 GHz machine, whereas BE executes in a few seconds using our prototype MATLAB implementation. Passage times are measured at the output of a randomly-chosen station of the network, both for inter-arrival times (random variable X) and cycle times (random variable R). BE is executed using the equilibrium BT defined by the conditional entropy method and for comparison also with the Chow-Liu BT. BT selection by BLP takes no more than a few seconds at the beginning of each run. We quantify accuracy error of an approximation $\widetilde{F}_Y(t)$ using two metrics:

- the mean absolute relative error Δ_{pct} of the 95th percentile position

$$\Delta_{pct} = \left| \frac{\widetilde{F}_Y^{-1}(0.95) - F_Y^{-1}(0.95)}{F_Y^{-1}(0.95)} \right| \tag{35}$$

- the mean absolute relative error Δ_{cdf} of the mass corresponding to the 95th percentile

$$\Delta_{cdf} = \left| \frac{\widetilde{F}_Y(F_Y^{-1}(0.95)) - 0.95}{0.95} \right| \tag{36}$$

where $F_Y^{-1}(0.95)$ is the exact value of the 95th percentile of Y. The two error metrics may be interpreted graphically as the relative errors of the approximation $\widetilde{F}_Y(t)$ with respect to the exact distribution $F_Y(t)$ on the horizontal axis (Δ_{pct}) and on the vertical axis (Δ_{cdf}). Since distributions bend horizontally around the 95th percentile, a small probability mass error Δ_{cdf} is strongly amplified in the corresponding percentile error Δ_{pct} which is therefore a challenging metric for validation.

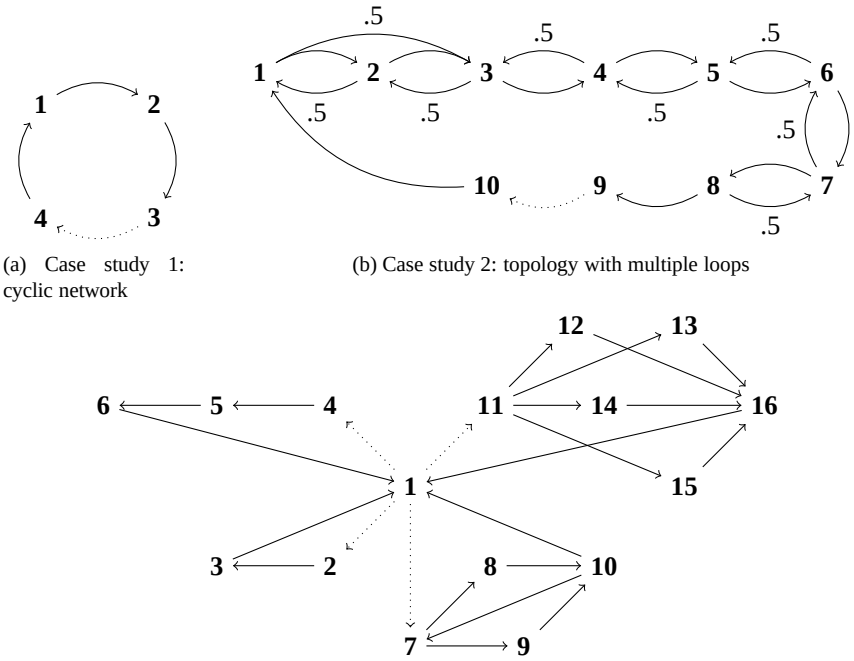
Table 2 gives results of the experiments on models with random topologies and service rates. Markov and Chebyshev inequalities offer poor accuracy with errors up to 557%. Markov inequality errors are enormous, yet also Chebyshev inequality is clearly unsatisfactory with errors up to 79%. Conversely, BE with the conditional entropy BT offers the best average error. For the Δ_{cdf} metric, we see that the error of BE is always less than 3.0%. The stricter Δ_{pct} metric shows that BE has already good performance for small population ($N = 5$), but accuracy increases as the total population grows. This effect can be explained by noting that, for low population values, the conditional probabilities in approximation (12) are computed by summing a small number of terms, hence estimation errors in a single term are heavily reflected on the entire summation. Furthermore, as the population grows, our reference model approaches a Jackson network hence queues tend to be mutually independent and the error of (12) with respect to (10) decreases. We also observe that BE based on Chow-Liu BTs performs effectively as well, however the average performance is significantly worse than for the conditional entropy approach. Indeed there exist some models where the Chow-Liu BT performs better than the conditional entropy BT, unfortunately the lack of exact solutions for transient analysis makes it difficult to gain more insights on the underlying reasons.

With respect to computational times, in the experiments with $N = 5$ the mean time for computation of a single point for the throughput with BE is $6ms$, and it grows to $20ms$ for $N = 10$; for response time these are $6ms$ and $21ms$, respectively. Usually, at least some tens or hundreds of points are needed for an accurate approximation of passage time distributions. This provides good intuition on the high efficiency of the BE method compared to the several minutes required for the exact evaluation of a single model by numerical techniques.

6.2. Case study 1: cyclic queueing network

Throughout the next subsections, we report case studies of increased complexity that prove that the BE technique can provide very good accuracy in models with structured topologies. Indeed, as shown in Table 2, there exist also cases where the approximation error exceeds 15% (e.g., for cycle time analysis), however in all our experiment we observed BE accuracy to grow systematically as the number of jobs in the network increases.

We consider a cyclic network composed by $M = 4$ queues having rates $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_4 = 4$, and population of $N = 5, 10, 20$ jobs. For this class of networks there exist several exact and efficient analysis techniques [10, 12, 16, 5, 8, 19, 14], thus BE is neither needed nor recommended to analyze this class of models and we provide this example just for illustration purposes. We are here interested in studying the inter-arrival time distribution of jobs observed at the output of station 3. This can be formulated as a passage time analysis problem with initial distribution $\boldsymbol{\pi}(0)$ identical to the equilibrium distribution of a model with $N - 1$ jobs relatively to the states where queue 4 is busy; elsewhere the initial probability is set to zero. For instance, the probability in $\boldsymbol{\pi}(0)$ relatively to state $\boldsymbol{n} = (2, 1, 1, 1)$ is the same probability of state $(2, 1, 1, 0)$ in a model with a job less; the initial probability for state $(1, 2, 2, 0)$ is instead zero. The states with nonzero probability identify the set Σ_0 which is reached upon activation of a departure transition from queue 3.



(a) Case study 1: cyclic network

(b) Case study 2: topology with multiple loops

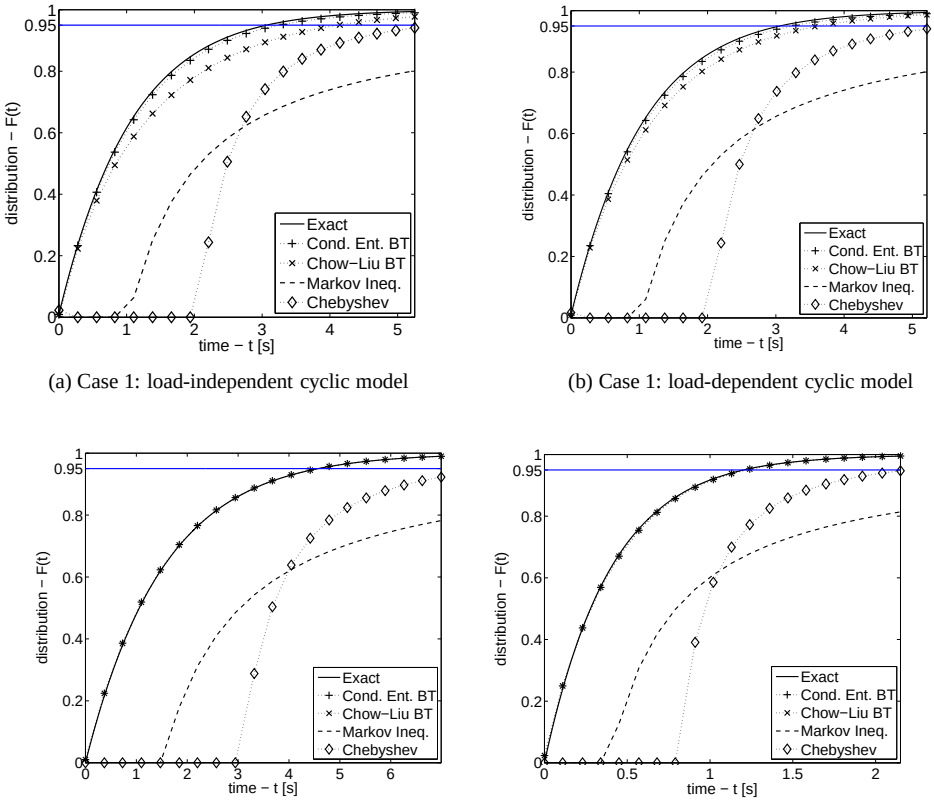
(c) Case study 3: complex network. See case description for routing probabilities.

Figure 5: Routing topologies used in the case studies. Dotted edges indicate the subset of transitions to Λ_0 that are considered in the generalized passage time analysis, see Section 2.1.

On this model, the times for evaluation of a single point by BE are $5ms$ for $N = 5$, $7ms$ for $N = 10$, and $30ms$ for $N = 20$. Results for $N = 5$ are shown in Figure 6(a) proving the good agreement between exact solution and BE approximation based on the conditional entropy BT. Small discrepancies are noted for $F(t) > 0.7$, however these differences disappear on models with $N = 10$ and $N = 20$ (not shown in the figure) where BE further improves its accuracy. Figure 6(a) also shows the inaccuracies of Markov and Chebyshev inequalities and provides another case where the Chow-Liu BT yields worse results compared to the conditional entropy BT. The dependence structure of the BTs used in Figure 6(b) is reported below.

BT	$par(\cdot)$ structure			
	n_1	n_2	n_3	n_4
Cond. Ent.	4	4	4	4
Chow-Liu	1	1	1	1

We also considers the same model discussed above in the case where station 3 has $c_2 = 2$ servers, thus its load-dependent rate is $\mu_3(n) = 3\min(2, n)$. Results in Figure 6(b) show that BE performance is quite insensitive to the presence of load-dependence rates and actually the performance of the Chow-Liu BT is even improved. Note that the BT selection strategy returns in this case the following BTs:



(a) Case 1: load-independent cyclic model

(b) Case 1: load-dependent cyclic model

(c) Case 2: topology with multiple loops

(d) Case 3: complex network

Figure 6: Distribution of inter-arrival times (X) in experimental case studies

BT	$par(\cdot)$ structure			
	n_1	n_2	n_3	n_4
Cond. Ent.	4	4	4	4
Chow-Liu	2	2	1	1

6.3. *Case study 2: topology with multiple loops*

We now evaluate the performance of BE on a medium-scale model with $M = 10$ queues having multiple loops between stations, a case that cannot be addressed by exact analytic methods [19]. Service rates are set to $\mu_i = 10 - i$, $i = 1, \ldots, 10$; routing probabilities are shown in the topology diagram in Figure 5(b). The time for evaluation of a single point with BE is $2.5s$ for $N = 25$ and $9.6s$ for $N = 50$. The integration step is $\Delta t_l = 0.0155$. Results for $N = 50$ are qualitatively similar to the other cases and shown in Figure 6(c) in comparison with a long-run simulation with 10 million samples. We do not observe any significant deviation between approximation and exact results. Furthermore, this reports a case where the Chow-Liu BT performs equally well of the conditional entropy BT. The dependence structure of the two BTs is reported below.

BT	*par*(·) structure									
	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
Cond. Ent.	9	9	9	9	9	9	9	9	9	9
Chow-Liu	1	8	8	8	8	8	8	1	8	8

6.4. Case study 3: complex network

This is a large-scale model that we have investigated to prove the scalability of BE with the model size. The population is $N = 50$ jobs which yields a prohibitively large state space with 10^{14} states that are reduced to only $13,270$ by BE. The model has $M = 16$ stations, where stations 1 and 16 are IS, stations $2 - 15$ are PS. Stations are arranged according to the topology shown in Figure 5(c), which represents an architecture composed of four sub-networks which receive jobs from the IS station 1. Service rates are as follows $\mu_1 = 1$, $\mu_2 = 5$, $\mu_3 = 3.33$, $\mu_4 = 10$, $\mu_5 = 2$, $\mu_6 = 1.67$, $\mu_7 = 20$, $\mu_8 = 1$, $\mu_9 = 1.25$, $\mu_{10} = 5$, $\mu_{11} = 1.67$, $\mu_{12} = 3.33$, $\mu_{13} = 5$, $\mu_{14} = 2.5$, $\mu^{-1} = 10$, and $\mu_{16} = 1$. Routing probabilities not specified in Figure 5(c) are $p_{1,2} = p_{1,4} = p_{1,7} = p_{1,11} = 0.25$, $p_{7,8} = 0.2$, $p_{7,9} = 0.8$, $p_{11,12} = 0.1$, $p_{11,13} = 0.2$, $p_{10,1} = 0.5$, $p_{10,7} = 0.5$, $p_{11,14} = 0.3$, $p_{11,15} = 0.4$. Station 7 is a multi-server station with $c_7 = 5$ servers. Numerical results shown in Figure 6 confirm the effectiveness of BE.

BT	*par*(·) structure															
	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}
Cond. Ent.	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
Chow-Liu	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

7. Conclusion

We have presented Bayesian Expansion (BE), a numerical approximation algorithm for estimating passage time distributions in queueing network models, a problem of growing interest for IT service sizing. The technical innovations brought by the BE approximation are several, among which the main ones are: i) the idea of approximating an intractable state space of a queueing model by means of a Bayesian tree; ii) the derivation of a technique to define Bayesian trees driven by the conditional entropy metric, which is shown to be more effective in BE than established methods used in machine learning [9]; iii) the efficient and accurate applications of the above ideas to the passage time distribution analyses in queueing networks; iv) the applicability of the methodology to networks that do not impose cyclic of tree-like topologies considered in several works [10, 12, 16, 5, 8, 19, 14]. Numerical results show that BE typically provides accurate results on random models and case studies.

Open challenges include: i) extension and assessment of the BE approximation on non-product-form models (see Section 5); ii) extension of BE to scheduling disciplines other than IS and PS, possibly the class of symmetric policies, and evaluation of the resulting accuracy; iii) assessing BE applicability to multiclass networks, which involve a large number of random variables in the joint probability density and therefore are inherently more challenging to approximate; iv) a comparison with fluid techniques that have shown to be a valuable approach to transient analysis [6].

Acknowledgement

This work has been supported by the Imperial College Junior Research Fellowship. The author thanks Giuseppe Iazeolla, Rudesindo Núñez-Queija, the anonymous referees, and the Imperial College AESOP group for comments that greatly helped improving this paper.

References

[1] D. Ardagna, M. Trubian, and L. Zhang. SLA based resource allocation policies in autonomic environments. *J. Parallel Distrib. Comput.*, 67(3):259–270, 2007.

[2] P. Bazan, R. German. Approximate transient analysis of large stochastic models with WinPEPSY-QNS. *Computer Networks*, 53(9):1289–1301, 2009.

[3] G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi. *Queueing Networks and Markov Chains*, Wiley, 2006.

[4] R. J. Boucherie and P. G. Taylor. Transient product from distributions in queueing networks. *DEDS*, 3(4):375–396, Sep 1993.

[5] O. J. Boxma, F. P. Kelly, A. G. Konheim. The product form for sojourn time distributions in cyclic exponential queues. *JACM*, 31(1):128–133, Jan 1984.

[6] J. T. Bradley, R. Hayden, W. J. Knottenbelt, and T. Suto Extracting Response Times from Fluid Analysis of Performance Models. *LNCS*, Vol 5119, 29–43, Jun 2008.

[7] G. Casale, N. Mi, E. Smirni. Model-Driven System Capacity Planning under Workload Burstiness. *IEEE Trans. Computers*, 59(1):66-80, 2010.

[8] S. Carbini, L. Donatiello, and G. Iazeolla. An efficient algorithm for the cycle time distribution in two-stages cyclic queues with a non-exponential server. in *Proc. of the Int. Seminar on Teletraffic Analysis and Computer Performance Evaluation*, 99–115, 1986.

[9] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Information Theory*, 14(11):462–467, Nov 1968.

[10] W. Chow. The Cycle Time Distribution of Exponential Cyclic Queues. *JACM*, 27(2):281–286, 1980.

[11] I. Cunha, J. Almeida, V. Almeida, and M. Santos. Self-adaptive capacity management for multi-tier virtualized environments. In *Proc. of the IFIP/IEEE IM*, 129–138, 2007.

[12] H. Daduna. Passage times for overtake-free paths in Gordon-Newell networks. *Advances in Applied Probability*, 14:672–686, 1982.

[13] E. de Souza e Silva and H. R. Gail. Transient solutions for Markov chains. in *Computational probability*, W. K. Grassmann Ed., Kluwer, 2000.

[14] P. G. Harrison. Laplace transform inversion and passage-time distributions in Markov processes. *J. of Appl. Prob.*, 27:74–87, 1990.

[15] P. G. Harrison and W. J. Knottenbelt. Passage time distributions in large Markov chains. In *Proc. of SIGMETRICS*, 129–138, 2002.

[16] P. G. Harrison. An Exact Analysis of the Distribution of Cycle Times in a Class of Queueing Networks. In *Proc. of SIGMETRICS*, 1983.

[17] L. Kleinrock. *Queueing Systems - Vol. 1*. John Wiley and Sons, 1975.

[18] T. Koski and J. M. Noble. *Bayesian Networks*. Springer, 2009.

[19] J. McKenna. Asymptotic Expansions of the Sojourn Time Distribution Functions of Jobs in Closed, Product-Form Queuing Networks. *JACM*, 34(4):985–1003, Oct 1987.

[20] K. C. Sevcik and I. Mitrani. The distribution of queuing network states at input and output instants. *JACM*, 28(2):358–371, 1981.

[21] G. Strang. Linear Algebra and Its Applications. Thomson, Brooks/Cole, 4th edition, 2006.

[22] R. Schassberger, H. Daduna. The Time for a Round Trip in a Cycle of Exponential Queues. *JACM*, 30(1):146–150, 1983.

[23] S. Zhang, I. Cohen, M. Goldszmidt, J. Symons, and A. Fox. Ensembles of models for automated diagnosis of system performance problems. In *Proc. of DSN*, 644–653, Jun 2005.