Performance Evaluation 67 (2010) 61–79

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Trace data characterization and fitting for Markov modeling

Giuliano Casale [a,*], Eddy Z. Zhang [b], Evgenia Smirni [b]

[a] *SAP Research, CEC Belfast, Newtownabbey, UK*

[b] *Department of Computer Science, College of William and Mary, Williamsburg, VA, USA*

ARTICLE INFO

Article history:
Received 16 June 2008
Received in revised form 5 August 2009
Accepted 3 September 2009
Available online 12 September 2009

Keywords:
Markovian arrival process
Fitting algorithm
Time series modeling
Temporal dependence
Correlations

ABSTRACT

We propose a trace fitting algorithm for Markovian Arrival Processes (MAPs) that can capture statistics of any order of interarrival times between measured events. By studying real traffic and workload traces often used in performance evaluation studies, we show that matching higher order statistical properties, in addition to first and second order descriptors, results in increased queueing prediction accuracy with respect to algorithms that only match the mean, the coefficient of variation, and the autocorrelations of the trace. This result supports the approach of modeling traces by the interarrival time process instead of the counting process that is more frequently used in the literature.

We proceed by first characterizing the general properties of MAPs using a spectral approach. Based on this result, we show how different MAPs can be combined together using Kronecker products to define a larger MAP with predefined properties of interarrival times. We then devise an algorithm that is based on this Kronecker composition and can accurately fit data traces. This MAP fitting algorithm uses nonlinear optimization that can be customized to fit an arbitrary number of moments and to meet the desired cost-accuracy tradeoff. Numerical results of the fitting algorithm on real data, such as the Bellcore Aug89 trace and a Seagate disk drive trace, indicate that the proposed fitting technique achieves increased prediction accuracy with respect to other state-of-the-art fitting methods.

1. Introduction

Markovian models provide a convenient way of evaluating the performance of network traffic and system workloads since their queueing analysis enjoys established theoretical results and efficient solution algorithms [1]. Although unable to directly generate sequences with long-range dependent (LRD) behavior, Markovian models can approximate accurately LRD traces in several ways, e.g., by superposition of processes with short-range dependent (SRD) behavior over many time scales [2]. This is known to be sufficient for the evaluation of real systems, e.g., for LRD traffic where the performance effects of statistical correlations becomes nil beyond a finite number of time scales [3].

One of the main obstacles to the Markovian analysis of data traces is model parameterization, which often requires to describe in the fitted Markov model the interaction of several tens or hundreds of states. Even for basic Markov Modulated Poisson Processes (MMPPs) or phase-type (PH) renewal processes, few results exist for their exact parameterization and the focus is usually on models with two or three states only [4–9]. Due to the lack of characterization results, it is also hard to establish detailed properties of these processes in the general case.

In this paper, we tackle the above issues by developing characterization and fitting methods for Markovian Arrival Processes (MAPs), a class of Markovian models developed by Neuts [10] that encompasses MMPP and PH processes as special

[*] Corresponding author. Tel.: +44 28 9078 5733; fax: +44 28 9078 5777.
E-mail addresses: giuliano.casale@sap.com (G. Casale), eddy@cs.wm.edu (E.Z. Zhang), esmirni@cs.wm.edu (E. Smirni).

0166-5316/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2009.09.003

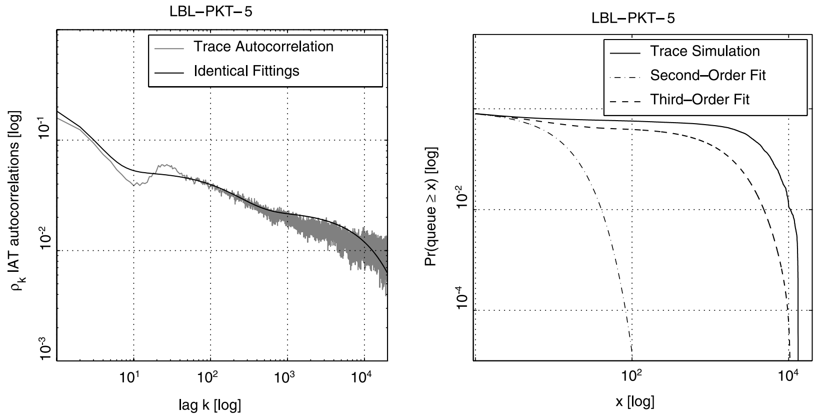

Fig. 1. Autocorrelation and MAP/M/1 queueing behavior (util. 80%) of two MAP(32) fittings of the LBL-PKT-5 trace [15]. The MAPs have identical first and second order properties of the interarrival process, but one has also an accurate fitting of third order properties for which the other is instead loose.

cases. We describe the properties of the interarrival time (IAT) process of a MAP and use these properties to derive accurate fitting algorithms for measured time series.

There are several works in the literature that have focused on fitting Markovian models of measured traces by exactly parameterizing MAPs/MMPPs with two or three states [11,12,5–8,13]. The small state space minimizes the costs of queueing analysis, but it places significant assumptions on the form of the autocorrelations. For instance, a MMPP(2) cannot fit negative autocorrelations, while the MAP(2) autocorrelation function must be geometrically decreasing with constant decay rate [8].

In [2], Andersen and Nielsen develop a fitting algorithm to model LRD traffic traces by superposition of several MMPP(2) sources [14]. The algorithm matches first and second order descriptors of the counting process, i.e., the mean traffic rate, the Hurst parameter, and the lag-1 autocorrelation in counts. This method has low computational costs and captures well the properties of the classic Bellcore LRD traces [15,16].

Following a different approach, Horváth and Telek [17] consider the multifractal traffic model of Riedi et al. [18], and obtain a class of MMPPs which exhibits multifractal behavior [19]. According to this result, one may fit network traffic by first computing an unnormalized Haar wavelet transform of the trace and then determining the MMPPs which best match the variance of the wavelet coefficients at different time scales. Simulation results on the Bellcore Aug89 trace show that this algorithm achieves better accuracy than the superposition method in [2], but at the expense of a larger state space.

Recently, several research efforts [7,6,20,8,21,22] are directed toward the accurate fitting of the IAT process instead of the counting process that is considered in [2] and [17]. IATs can be harder to measure than counts [23], but simple analytical expressions are available for their moments and lag correlations [24]. Instead, only the first three moments of a counting process are known and can be manipulated from closed-form analytical expressions [14,25]. Several authors have shown that fitting the mean, coefficient of variation and autocorrelations is insufficient to predict queueing behavior [26–28], therefore fitting the higher order properties of the IAT process seems a natural way to achieve increased prediction accuracy.

To build intuition on the importance of higher order properties we first present an experiment on the LBL-PKT-5 trace of the Internet Traffic Archive [15]. Fig. 1 shows two different MAP models we have obtained for this trace. The two MAPs have identical first and second order properties of the IAT process, namely same mean, same squared coefficient of variation (*SCV*), and same autocorrelation function. Mean and *SCV* are identical to the sample values, the autocorrelation fit is also quite good as shown in the left graph. However, one model also matches the third order statistics, i.e., the skewness and the bicorrelations [29] of the sample IAT process, while the other has a quite loose fit of these descriptors. The strikingly different queueing predictions of the two models, shown in the right graph of Fig. 1, stress the importance of higher order properties of the measured samples.

In this paper, we propose to fit measured traces using higher order properties of the IAT process in addition to the usual first and second order descriptors. Because of the general difficulty in imposing even basic autocorrelations to the IAT process, we first derive characterization results using a spectral approach, see [30] for a review of previous work on MAP characterization. These characterization results simplify fitting and allow to obtain a MAP fitting algorithm that is based on nonlinear optimization that still matches moments, but can also accurately fit general traces. The latter algorithm is based on a new MAP definition technique, called Kronecker Product Composition (KPC), which is able to generate MAPs with predefined moments, autocorrelations, and higher order statistics in IATs. Compared to the state-of-the-art fitting methods in [2] and [17], the proposed MAP fitting algorithm shows improved queueing prediction accuracy at similar computational costs. In addition, it does not place limitations on the number and order of statistical properties that can be matched for a trace, thus enables the selection of the best cost-accuracy tradeoff.

Furthermore, our approach offers a different computational tradeoff compared to expectation–maximization (EM) algorithms developed in the literature [11,31,32,13,33]. EM algorithms determine a fitting of model parameters to a measured trace by iteratively maximizing the likelihood that the observed data has been sampled from the model defined by the current guess of the parameters. The EM approach has the significant advantage of accounting for *all* the information available from measurement, which is particularly important when the sample size is small. However, EM techniques suffer computational limitations if either the trace size or the number of parameters to be fitted is large. Compared to EM methods, the KPC method proposed in this paper presents the same advantages of a moment matching algorithm. Thus, the trace size does not affect directly computational costs which depend only on the number of moments, autocorrelations, and higher order statistics evaluated in the fitting.

Our detailed contributions can therefore be summarized as follows:

(1) MAP(n) *characterization*: After reviewing the IAT process in MAPs, in Section 3 we propose a general spectral characterization of IAT moments, autocorrelations, and higher order moments. These observations clarify the capabilities of general MAPs, provide necessary conditions for fitting, and simplify the analysis of small processes.

(2) *Compositional definition of* MAP(n): In Section 4 we propose a compositional method based on Kronecker products that can easily generate a MAP(n) with predefined properties of IATs from the composition of smaller processes, e.g., MAP(2)s [9]. While traditional superposition is convenient only for imposing first and second order properties of counts, our method is more flexible and gives complete control of the IAT statistics at all orders.

(3) *MAP fitting*: Exploiting the previous results, we develop in Section 5 the proposed fitting algorithm which first determines the optimal values of IAT moments, autocorrelations, and higher order descriptors using nonlinear optimization, and successively finds the MAP which best matches these target values. The approach is numerically stable and the fitting can be performed in a few minutes. Comparative analyses in Section 6 on the Bellcore Aug89 trace and on the Seagate Web trace [34] with existing fitting methods show that our algorithm achieves the best accuracy. The relative merit of MAP(2)s and of a special class of MAP(3)s as building block for the KPC fitting algorithm is also investigated.

Section 7 draws conclusions. The final appendix reports the MAPs used to fit the traces discussed in Section 6 and a description of the special class of MAP(3)s for KPC fitting. A MATLAB implementation of the proposed MAP fitting algorithm can be found in the KPC-Toolbox [35] which is available for download at http://www.cs.wm.edu/MAPQN/kpctoolbox.html.

2. IAT process in MAPs

A MAP(n) is specified by two $n \times n$ matrices: a stable matrix[1] $\mathbf{D}_0$ with nonnegative off-diagonal entries and a nonnegative matrix $\mathbf{D}_1$ that describe transition rates between n states. Each transition in $\mathbf{D}_1$ produces a job arrival; $\mathbf{D}_0$ describes instead background transitions not associated with arrivals. The matrix $\mathbf{Q} = \mathbf{D}_0 + \mathbf{D}_1$ is the infinitesimal generator of the underlying Markov process. In the special case where $\mathbf{D}_1$ is a diagonal matrix, the process is a MMPP(n).

We focus on the process stationary at arrivals (i.e., interval-stationary) that describes the IATs. For a MAP(n), this is described by the embedded discrete-time chain with stochastic matrix $\mathbf{P} = (-\mathbf{D}_0)^{-1}\mathbf{D}_1$, having embedded probability vector $\boldsymbol{\pi}_e$, $\boldsymbol{\pi}_e\mathbf{P} = \boldsymbol{\pi}_e$, $\boldsymbol{\pi}_e\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of 1's of the appropriate dimension. Let $\mathbf{P}$ be irreducible with a simple unit eigenvalue $\gamma_1 = 1$. Then, IATs are phase-type distributed with k-th moment

$$E[X^k] = k!\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-k}\mathbf{e}, \quad k \geq 0, \tag{1}$$

where X is the random variable representing interarrival times, which implies that $SCV = 2E[X]^{-2}\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-2}\mathbf{e} - 1$. The lag-$k$ autocorrelation coefficient is

$$\rho_k = (E[X]^{-2}\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-1}\mathbf{P}^k(-\mathbf{D}_0)^{-1}\mathbf{e} - 1)/SCV. \tag{2}$$

Higher order moments of the IAT process can be described in terms of *joint moments*. Let X_i be the i-th IAT with respect to an arbitrary starting epoch $i_0 = 0$, and consider a sequence $X_{i_1}, X_{i_2}, \ldots, X_{i_L}$, where $0 \leq i_1 < i_2 < \cdots < i_L$. The joint moments of L IATs are the functions

$$H(\vec{i}, \vec{k}) = E[X_{i_1}^{k_1} X_{i_2}^{k_2} \cdots X_{i_L}^{k_L}],$$

where $\vec{i} = (i_1, i_2, \ldots, i_L)$ and $\vec{k} = (k_1, k_2, \ldots, k_L)$. The moments $H(\vec{i}, \vec{k})$ capture nonlinear temporal relations between samples and are known to completely characterize a MAP [28,36]. They are computed as [36]

$$H(\vec{i}, \vec{k}) = \boldsymbol{\pi}_e \left(\prod_{l=1}^{L} k_l!(-\mathbf{D}_0)^{-k_l}\mathbf{P}^{i_l - i_{l-1}} \right) \mathbf{e}, \tag{3}$$

where, for $l = 1$, i_0 is set to $i_0 = 0$. Noting that it is always $\mathbf{P}^{i_L - i_{L-1}}\mathbf{e} = \mathbf{e}$, (3) reduces in the case $L = 1$ to (1).

In the remainder of this paper and if not otherwise stated, MAP descriptors refer to the IAT process. Further, we use the notation $(\mathbf{D}_0, \mathbf{D}_1)$ or $((-\mathbf{D}_0)^{-1}, \mathbf{P})$ to uniquely specify a MAP. The two representations are equivalent by setting $\mathbf{D}_1 = -\mathbf{D}_0\mathbf{P}$.

[1] A square matrix is said to be stable if its eigenvalues have negative real part.

3. Characterization of MAP(n)s

We now obtain a spectral characterization of IAT moments and autocorrelations, i.e., a scalar representation of (1)–(2) based on spectral properties of $(-\mathbf{D}_0)^{-1}$ and $\mathbf{P}$. This simplifies the analysis of MAP moments and autocorrelations.

3.1. Characterization of moments

We begin by describing the moments (1) in terms of the spectrum of $(-\mathbf{D}_0)^{-1}$. Recall that the characteristic polynomial of an $n \times n$ matrix $\mathbf{A}$ is

$$\phi(\mathbf{A}) = det(s\mathbf{I} - \mathbf{A}) = s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1}s + \alpha_n, \tag{4}$$

which is a polynomial in s with roots s_i equal to the eigenvalues of $\mathbf{A}$. Consider the Cayley–Hamilton Theorem [37], by which the powers of $\mathbf{A}$ satisfy

$$\mathbf{A}^k = -\sum_{j=1\dots n} \alpha_j \mathbf{A}^{k-j}, \quad k \geq n \tag{5}$$

that is, matrix powers are linearly dependent. Because MAP moments are computed in (1) from powers of $(-\mathbf{D}_0)^{-1}$, they are linearly dependent.

Lemma 1. *In a MAP (n), any $n+1$ consecutive moments are linearly dependent according to the relation*

$$E[X^k] = -\sum_{j=1\dots n} \left(\frac{k!m_j}{(k-j)!}\right) E[X^{k-j}], \quad E[X^0] = 1, \quad k \geq n, \tag{6}$$

where m_j is the coefficient of s^{n-j} in $\phi((-\mathbf{D}_0)^{-1})$.

Proof. Using the Cayley–Hamilton theorem,

$$E[X^k] = -k!\boldsymbol{\pi}_e \left(\sum_{j=1\dots n} m_j(-\mathbf{D}_0)^{-(k-j)}\right)\mathbf{e} \tag{7}$$

which immediately proves the lemma by (1). □

Since the coefficients m_j are functions of the eigenvalues of $(-\mathbf{D}_0)^{-1}$ we can derive a closed-form formula for $E[X^k]$.

Theorem 1. *Let $(-\mathbf{D}_0)^{-1}$ have $m \leq n$ distinct eigenvalues $\theta_t \in \mathbb{C}$, $1 \leq t \leq m$. Let q_t be the algebraic multiplicity of θ_t, $\sum_{t=1\dots m} q_t = n$. Then the IAT moments are given by*

$$E[X^k] = \sum_{t=1\dots m} k!\,\theta_t^k \sum_{j=1\dots q_t} M_{t,j}k^{j-1}, \tag{8}$$

$$E[X^0] = \sum_{t=1\dots m} M_{t,1} = 1, \tag{9}$$

where the constants $M_{t,j}$'s are independent of k. In particular,

$$M_{t,1} = \boldsymbol{\pi}_e(-\mathbf{D}_0)_t^{-1}\mathbf{e}, \tag{10}$$

where $(-\mathbf{D}_0)_t^{-1}$ is the t-th spectral projector of $(-\mathbf{D}_0)^{-1}$, i.e., the product of the right and left eigenvectors for θ_t.

Proof. Denoting by $(-\mathbf{D}_0)_t^{-1}$ and $\mathbf{M}_t$ the spectral projector and nilpotent matrix of $(-\mathbf{D}_0)^{-1}$ associated to the Jordan block for θ_t, the generalized spectral decomposition of $(-\mathbf{D}_0)^{-1}$ is [38]

$$(-\mathbf{D}_0)^{-1} = \sum_{t=1\dots m} (\theta_t(-\mathbf{D}_0)_t^{-1} + \mathbf{M}_t), \quad k \geq 0$$

where $\mathbf{M}_t^{q_t} = \mathbf{0}$, $\mathbf{M}_t(-\mathbf{D}_0)_t^{-1} = (-\mathbf{D}_0)_t^{-1}\mathbf{M}_t$, $\mathbf{M}_t(-\mathbf{D}_0)_p^{-1} = (-\mathbf{D}_0)_p^{-1}\mathbf{M}_t = \mathbf{0}$, $t \neq p$, and $(-\mathbf{D}_0)_t^{-1}(-\mathbf{D}_0)_p^{-1} = \mathbf{0}$, $t \neq p$. Therefore, for all $k \geq 0$ we have

$$\begin{aligned}(-\mathbf{D}_0)^{-k} &= \Big(\sum_{t=1\dots m} (\theta_t(-\mathbf{D}_0)_t^{-1} + \mathbf{M}_t)\Big)^k \\ &= \sum_{t=1\dots m} (\theta_t(-\mathbf{D}_0)_t^{-1} + \mathbf{M}_t)^k \\ &= \sum_{t=1\dots m} \theta_t^k \sum_{i=0}^{\min\{q_t-1,k\}} \binom{k}{i} (-\mathbf{D}_0)_t^{-1}(\theta_t^{-1}\mathbf{M}_t)^i,\end{aligned}$$

where we used in the last passage that $\theta_t \neq 0$ which is always true because $(-\mathbf{D}_0)^{-1}$ is an invertible matrix and thus its eigenvalues are all different from zero. Inserting the last formula for $(-\mathbf{D}_0)^{-k}$ into (1) we get after some manipulations

$$E[X^k] = k! \sum_{t=1\dots m} \theta_t^k \sum_{i=1}^{\min\{q_t,k+1\}} \binom{k}{i-1} \widehat{M}_{t,i}, \tag{11}$$

where

$$\widehat{M}_{t,i} = \boldsymbol{\pi}_e (-\mathbf{D}_0)_t^{-1} (\theta_t^{-1}\mathbf{M}_t)^{i-1}\mathbf{e}. \tag{12}$$

The last expression is equivalent to (8) by expanding the binomials and grouping the coefficients of k^j. This yields the following equivalence

$$M_{t,j} = \sum_{i=j}^{q_t} \frac{s(i-1,j-1)}{(i-1)!} \widehat{M}_{t,i} \tag{13}$$

where the $s(m,n)$ is the Stirling number of the first kind giving the coefficient of x^n in $x(x-1)(x-2)\cdots(x-m+1)$. Finally, the condition $\sum_t M_{t,1} = 1$ is obtained by evaluating (1) or (8) for $k = 0$ and noting that it is always $E[X^0] = 1$. □

Corollary 1. *If θ_t has algebraic multiplicity $q_t = 1$, then $M_{t,j} = 0$ for $j \geq 2$.*

Proof. In this case the nilpotent $\mathbf{M}_t$ of the t-th Jordan block is zero and in (13) the only non-zero projector is $M_{t,1}$. □

Note that formula (8) is a Jordan decomposition of $(-\mathbf{D}_0)^{-1}$ since it also holds for defective, i.e., non-diagonalizable, $(-\mathbf{D}_0)^{-1}$. This is extremely important, since well-known processes, e.g., the Erlang process, have $\mathbf{D}_0$ that is not diagonalizable.

Example 1. We show how to apply Theorem 1 for the analytical characterization of a MAP. Consider the MAP(3)

$$\mathbf{D}_0 = \begin{bmatrix} -2\lambda & \lambda & \lambda \\ 0 & -\lambda & \lambda \\ 0 & 0 & -\lambda \end{bmatrix}, \qquad \mathbf{D}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \lambda & 0 & 0 \end{bmatrix}, \quad \lambda \geq 0.$$

The left eigenvector of $\mathbf{P}$ for $\gamma_1 = 1$ is $\boldsymbol{\pi}_e = [1, 0, 0]^{\mathrm{T}}$. Since

$$(-\mathbf{D}_0)^{-k} = \begin{bmatrix} 2^{-k}\lambda^{-k} & (1-2^{-k})\lambda^{-k} & k\lambda^{-k} \\ 0 & \lambda^{-k} & k\lambda^{-k} \\ 0 & 0 & \lambda^{-k} \end{bmatrix},$$

from (1) it is $E[X^k] = k!\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-k}\mathbf{e} = (k+1)!\lambda^{-k}$. However this approach does not generalize easily, because obtaining a closed-form expression for $(-\mathbf{D}_0)^{-k}$ on larger examples can be difficult. We show that the spectral characterization can analyze $E[X^k]$ without the need of a closed-form formula for $(-\mathbf{D}_0)^{-k}$. We first compute $E[X] = 2\lambda^{-1}$ and $E[X^2] = 6\lambda^{-2}$, and observe that the eigenvalues of $(-\mathbf{D}_0)^{-1}$ are $\theta_1 = (2\lambda)^{-1}$ and $\theta_2 = \lambda^{-1}$ with multiplicity $q_1 = 1$ and $q_2 = 2$. Imposing $E[X]$ and $E[X^2]$ in (8), we find $M_{1,1} = 0, M_{2,1} = 1 - M_{1,1} = 1, M_{2,2} = 1$, and substituting back we finally get $E[X^k] = k!\theta_1^{-k}M_{1,1} + k!\theta_2^{-k}(M_{2,1} + kM_{2,2}) = (k+1)!\lambda^{-k}$. □

We also observe that if $(-\mathbf{D}_0)^{-1}$ is diagonalizable, then $m = n$ and the projectors $M_{t,1}$ are in simple relation to the IAT cdf since

$$F(x) = 1 - \boldsymbol{\pi}_e e^{\mathbf{D}_0 x}\mathbf{e} = 1 - \sum_{t=1\dots n} M_{t,1} e^{-x/\theta_t}, \tag{14}$$

which follows by the relation $e^{\mathrm{diag}(-\theta_1^{-1},\dots,-\theta_n^{-1})} = \mathrm{diag}(e^{-\theta_1^{-1}},\dots,e^{-\theta_n^{-1}})$ and the computational formula for $M_{t,1}$. Note that (14) allows an efficient numerical computation of quantities such as percentiles of the IAT distribution.

We remark that the above characterization is sufficient to develop a simple moment matching algorithm for hyper-exponential traces. We point to [39] for a description of this results and examples that illustrate its accuracy using real traffic traces of the Internet Traffic Archive [15].

3.2. Characterization of autocorrelation

The spectral characterization can be extended to autocorrelations using the properties of the powers $\mathbf{P}^k$ in (2).

Lemma 2. *In a MAP (n), any $n+1$ consecutive autocorrelations are linearly dependent according to the relation*

$$\rho_k = -\sum_{j=1\dots n} a_j \rho_{k-j}, \qquad \rho_0 = (1 - 1/SCV)/2, \quad k \geq n, \tag{15}$$

where a_j is the coefficient of s^{n-j} in $\phi(\mathbf{P})$ and $\sum_{j=0}^n a_j = 0$ where $a_0 = 1$.

Proof. We want to prove that $\sum_{j=0\dots n} a_j\rho_{k-j} = 0$, where $a_0 = 1$. By definition of ρ_k, this is equivalent to prove that

$$\sum_j a_j(\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-1}\mathbf{P}^{k-j}(-\mathbf{D}_0)^{-1}\mathbf{e} - E[X]^2) = 0.$$

The last equation is indeed true if we can show that $\sum_{j=0}^n a_j\mathbf{P}^{k-j} = 0$ and $\sum_{j=0}^n a_j = 0$. But the former holds true by the Cayley–Hamilton theorem, while the latter follows by the stochasticity of $\mathbf{P}$, since for the unit eigenvalue $\gamma_1 = 1$ it is $\phi(\mathbf{P}) = 0 = \sum_{j=0}^n a_j$. This proves $\rho_k = -\sum_{j=1\dots n} a_j\rho_{k-j}$. The formula for ρ_0 follows by evaluating (2) for $k = 0$, i.e.,

$$\rho_0 = (E[X]^{-2}\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-2}\mathbf{e} - 1)/SCV = (1 - 1/SCV)\,/2$$

since $\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-2}\mathbf{e} = E[X^2]/2 = (1 + SCV)E[X]^2/2$.

Similarly to Theorem 1, we can obtain a closed-form expression of ρ_k.

Theorem 2. *Let $\gamma_t \in \mathbb{C}$, $1 \le t \le m$, be an eigenvalue of $\mathbf{P}$ with algebraic multiplicity r_t. If $\gamma_t = 0$ assume that its geometric multiplicity equals its algebraic multiplicity, i.e., the r_t associated Jordan blocks have all order one. Then the autocorrelation function of a MAP is*

$$\rho_k = \sum_{t=2\dots m} \gamma_t^k \sum_{j=1\dots r_t} A_{t,j}k^{j-1}, \quad k \ge 1 \tag{16}$$

$$\rho_0 = \sum_{t=2\dots m} A_{t,1} = (1 - 1/SCV)/2, \tag{17}$$

where the $A_{t,j}$'s constants are independent of k. In particular,

$$A_{t,1} = E[X]^{-2}\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-1}\mathbf{P}_t(-\mathbf{D}_0)^{-1}\mathbf{e}/SCV, \tag{18}$$

in which $\mathbf{P}_t$ is the t-th spectral projector of $\mathbf{P}$, that is, the product of the right and left eigenvectors associated to γ_t.

Proof. The proof is similar to the proof of Theorem 1. Let us assume first that $\gamma_t \ne 0$ for all t. If γ_t has multiplicity r_t, the generalized spectral decomposition of $\mathbf{P}$ gives [38]

$$\mathbf{P} = \sum_{t=1\dots m} (\gamma_t\mathbf{P}_t + \mathbf{N}_t), \quad k \ge 0$$

where $\mathbf{P}_1 = \mathbf{e}\boldsymbol{\pi}_e$, $\mathbf{N}_t$ is the nilpotent matrix associated to γ_t, $\mathbf{N}_t^{r_t} = \mathbf{0}$, $\mathbf{N}_t\mathbf{P}_t = \mathbf{P}_t\mathbf{N}_t$, $\mathbf{P}_t\mathbf{P}_p = \mathbf{0}$, $t \ne p$, and $\mathbf{N}_t\mathbf{P}_p = \mathbf{P}_p\mathbf{N}_t = \mathbf{0}$, $t \ne p$. Therefore,

$$\mathbf{P}^k = \Big(\sum_{t=1\dots m} (\gamma_t\mathbf{P}_t + \mathbf{N}_t)\Big)^k = \sum_{t=1\dots m} (\gamma_t\mathbf{P}_t + \mathbf{N}_t)^k$$
$$= \sum_{t=1\dots m} \gamma_t^k \sum_{i=0}^{\min\{r_t-1,k\}} \binom{k}{i} \mathbf{P}_t(\gamma_t^{-1}\mathbf{N}_t)^i, \quad k \ge 0.$$

Inserting the last formula for $\mathbf{P}^k$ into (2) we get after algebraic manipulations

$$\rho_k = \sum_{t=2\dots m} \gamma_t^k \sum_{i=1}^{\min\{r_t,k+1\}} \binom{k}{i-1}\widehat{A}_{t,i},$$

where

$$\widehat{A}_{t,i} = E[X]^{-2}\boldsymbol{\pi}_e(-\mathbf{D}_0)^{-1}\mathbf{P}_t(\gamma_t^{-1}\mathbf{N}_t)^{i-1}(-\mathbf{D}_0)^{-1}\mathbf{e}/SCV$$

and we have noticed that for $\gamma_1 = 1$ it is $\widehat{A}_{t,i} = E[X]^{-2}/SCV$ that simplifies with the similar term appear in (2). Grouping the coefficients of k^j, we have

$$A_{t,j} = \sum_{i=j}^{r_t} \frac{s(i-1, j-1)}{(i-1)!}\widehat{A}_{t,i}, \tag{19}$$

where the $s(m, n)$'s are the Stirling number of the first kind. Note that for $k = 0$ one has immediately from (18)

$$\sum_{t=2\dots m} \widehat{A}_{t,1} = \rho_0 = (1 - 1/SCV)/2,$$

where the value of ρ_0 follows from Lemma 2.

From Lemma 2 we see that the function ρ_k when evaluated in $k = 0$ assumes the value $\rho_0 = (1 - 1/SCV)/2$. Although this coefficient does not admit any statistical interpretation, since the autocorrelation function is by definition $\rho_k = 1$ for $k = 0$, it is useful to consider this limit value since the condition $\rho_0 = \sum_t A_{t,1}$ can simplify the computation of projectors. The value ρ_0 can also help in manipulating the autocorrelation coefficients, since it is often observed that increasing ρ_0 produces a generalized increase of all autocorrelations. For instance, in the special case of a MAP(2), it follows from (16) that $\rho_k = \gamma_2^k \rho_0$ and therefore the autocorrelations increases monotonically as a function of ρ_0.

Corollary 2. *If γ_t has algebraic multiplicity $r_t = 1$, then $A_{t,j} = 0$ for $j \geq 2$.*

Proof. If all nilpotents $\mathbf{N}_t$ are zero, then by definition of $\widehat{A}_{t,i}$ the only non-zero projector in (18) is $A_{t,1}$. □

Without loss of generality, we assume in the rest of the paper that $|\gamma_j| \geq |\gamma_{j+1}|, j = 1, \ldots, n-1$. According to this ordering, the asymptotic decay of the autocorrelation function is geometric with rate γ_2 (unless $\gamma_2 = -1$ and ρ_k does not converge to zero as $k \to +\infty$). We complete the analysis in Theorem 2 by studying the following degenerate case.

Corollary 3. *If $\mathbf{P}$ has zero eigenvalues belonging to r_m Jordan blocks of order $l_0^1, l_0^2, \ldots, l_0^{r_m}$, then*

$$\rho_k = \sum_{j=1 \ldots r_m} \eta_{k,j} + \sum_{t=2 \ldots m-1} \gamma_t^k \sum_{j=1 \ldots r_t} A_{t,j} k^{j-1},$$

where

$$\eta_{k,j} = E[X]^{-2} \boldsymbol{\pi}_e (-\mathbf{D}_0)^{-1} (\mathbf{N}_{0,j})^k (-\mathbf{D}_0)^{-1} \mathbf{e}/SCV, \tag{20}$$

in which $\mathbf{N}_{0,j}$, $\mathbf{N}_{0,j}^{l_0^j} = \mathbf{0}$, is the nilpotent associated to the Jordan block of order l_0^j, and $\eta_{k,j}$ is equal to zero for $k \geq l_0^j$.

Proof. The generalized spectral decomposition of $\mathbf{P}$ is

$$\begin{aligned} \mathbf{P}^k &= \Big(\sum_{j=1 \ldots r_m} \mathbf{N}_{0,j} + \sum_{t=1 \ldots m-1} (\gamma_t \mathbf{P}_t + \mathbf{N}_t) \Big)^k \\ &= \sum_{j=1 \ldots r_m} \mathbf{N}_{0,j}^k + \sum_{t=1 \ldots m-1} \gamma_t^k \sum_{i=0}^{\min\{r_t - 1, k\}} \binom{k}{i} \mathbf{P}_t (\gamma_t^{-1} \mathbf{N}_t)^i, \end{aligned}$$

for $k \geq 0$. The rest of the proof follows by substituting the above expression into (2) and performing passages similar to the ones in the proof of Theorem 2. □

We conclude by remarking that the distinct $A_{t,j}$'s and γ_t's in (16) are no more than $2n - 2$. Thus a non-degenerate MAP(n) can fit up to $2n - 2$ independent autocorrelations ρ_k, $k \geq 0$. If the fitting assigns SCV, then ρ_0 is fixed and the maximum number of independent autocorrelations becomes $2n - 3$. This last result seems apparently in contradiction with the formula in Lemma 2, in which it is said that any $n + 1$ consecutive coefficients are linear dependent. However, this last relation is valid given an existing $(\mathbf{D}_0, \mathbf{D}_1)$ representation which imposes the value of the coefficients $a_1, \ldots, a_n$. However, in fitting one can also assign the $a_1, \ldots, a_n$ values, hence it is possible to fit up to $2n - 3$ autocorrelations as we explain below. To better understand the counter-intuitive derivation of the $2n - 3$ value, consider a MAP(2), then from Lemma 2 we can write

$$\rho_2 = a_1 \rho_1 + a_2 \rho_0, \quad a_0 + a_1 + a_2 = 0, \quad a_0 = 1.$$

This recursive equation appears to have $2n - 1 = 3$ degrees of freedom: the coefficient a_1 (a_2 is linear dependent) and the initial conditions ρ_1 and ρ_0. If SCV is fixed, the degrees of freedom seem to reduce to $2n - 2 = 2$ because $\rho_0 = (1 - 1/SCV)/2$ becomes fixed as well. However, if we apply a spectral expansion, we can clearly see that there are actually only $2n - 3 = 1$ degrees of freedom. In fact, noting that from Theorem 2 it is

$$\rho_1 = \gamma_2 A_{2,1} \Rightarrow \rho_1 = \gamma_2 \rho_0,$$

and similarly $\rho_2 = \gamma_2^2 \rho_0$, then we have

$$\rho_2 = a_1 \rho_1 + a_2 \rho_0 = (a_1 \gamma_2 + a_2) \rho_0,$$

but since $\rho_2 = \gamma_2^2 \rho_0$ we can write $(a_1 \gamma_2 + a_2) = \gamma_2^2$, and from $a_0 + a_1 + a_2 = 0$, $a_0 = 1$ we finally get

$$a_1 = -1 - \gamma_2 = -1 - \rho_1/\rho_0, \quad a_2 = \gamma_2.$$

The last result implies that, unexpectedly, also a_1 cannot be chosen arbitrarily once that we have set ρ_1 and ρ_0 since it is always $a_1 = -1 - \rho_1/\rho_0$. This reduces the degrees of freedom to $2n - 3$ and suggests that it is difficult to evaluate the number of degrees of freedom of a MAP directly from the recurrence equation of Lemma 2. The spectral expansion in (16) does not suffer from this problem since eigenvalues and spectral projectors are independent by definition.

Example 2. The Circulant MMPP is proposed in [26] to introduce complex eigenvalues in the autocorrelation of the counting process. According to our results, this approach can be generalized to the IAT process by simply defining a MAP with circulant $\mathbf{P}$ and/or $(-\mathbf{D}_0)^{-1}$. In particular, if $\mathbf{D}_0$ is diagonal, the resulting circulant MAP admits a quite simple characterization. Define $[p_1, p_2, \dots, p_n]$, $p_n = 1 - \sum_{j\neq n} p_j$, to be the first column of the circulant matrix $\mathbf{P}$. Since in a circulant $\mathbf{P}$ it is $\boldsymbol{\pi}_e = \mathbf{e}/n$, from (1) we have

$$E[X^k] = \left(\frac{k!}{n}\right)\sum_{t=1}^{n}\theta_t^k.$$

Using Theorem 2 we can also study autocorrelations. For instance, in the case $n = 3$ a MAP with circulant $\mathbf{P}$ and diagonal $\mathbf{D}_0$ has structure

$$\mathbf{D}_0 = \begin{bmatrix} -\theta_1^{-1} & 0 & 0 \\ 0 & -\theta_2^{-1} & 0 \\ 0 & 0 & -\theta_3^{-1} \end{bmatrix}, \qquad \mathbf{D}_1 = \begin{bmatrix} p_1\theta_1^{-1} & p_3\theta_1^{-1} & p_2\theta_1^{-1} \\ p_2\theta_2^{-1} & p_1\theta_2^{-1} & p_3\theta_2^{-1} \\ p_3\theta_3^{-1} & p_2\theta_3^{-1} & p_1\theta_3^{-1} \end{bmatrix}.$$

In this case, the circulant matrix has two identical or complex conjugate eigenvalues, which implies from the condition on $A_{2,1} + A_{3,1} = \rho_0$ that $A_{2,1} = A_{3,1} = \rho_0/2$. Now letting $\gamma_2 = |\gamma_2|e^{j\omega_2}$,

$$|\gamma_2| = (1/2)\sqrt{(3p_1 - 1)^2 + 3\Delta_{3,2}^2}, \qquad \omega_2 = \arg\left(\frac{(3p_1 - 1) + j\sqrt{3}\Delta_{3,2}}{2}\right),$$

with $\Delta_{i,j} = p_i - p_j$, the autocorrelation is

$$\rho_k = \rho_0|\gamma_2|^k\frac{(e^{jk\omega_2} + e^{-jk\omega_2})}{2} = \rho_0|\gamma_2|^k\cos(k\omega_2).$$

Higher order cases are similar. For example, for $n = 4$ after few manipulations we get

$$\rho_k = A_{a,1}\gamma_a^k + A_{b,1}|\gamma_b|^k\cos(k\omega_b),$$

with

$$\gamma_a = \Delta_{1,2} + \Delta_{3,4}, \qquad |\gamma_b| = \sqrt{\Delta_{4,2}^2 + \Delta_{1,3}^2}, \qquad \omega_b = \arg(\Delta_{1,3} + j\Delta_{4,2}),$$

$$A_{a,1} = \frac{(\theta_4 - \theta_2 + \theta_3 - \theta_1)^2}{SCV(\theta_3 + \theta_2 + \theta_4 + \theta_1)^2}, \qquad A_{b,1} = \rho_0 - A_{a,1},$$

where the eigenvalues are denoted by the indices a and b since the asymptotic decay rate $|\gamma_2|$ can be either $|\gamma_a|$ or $|\gamma_b|$. □

3.3. *Higher order statistics*

We observe that the characterization given for moments and autocorrelations generalizes in a similar fashion to the joint moments (3), since these functions consist of powers of $(-\mathbf{D}_0)^{-1}$ and $\mathbf{P}$. For example, in the case where both matrices are diagonalizable and $L = 2$, we have

$$H(\vec{i}, \vec{k}) = E[X_{i_1}^{k_1}X_{i_2}^{k_2}] = \sum_{t=1\dots,n}\sum_{l=1\dots,n} H_{t,l}\theta_t^{k_t}\gamma_l^{i_l}, \tag{21}$$

where the joint moment projector $H_{t,l}$ is a constant independent of $\vec{i}$ and $\vec{k}$ and it is computed from the product of the spectral projectors of $(-\mathbf{D}_0)^{-1}$ and $\mathbf{P}$. From (3) it can be seen that for general L the joint moment projector is not in simple relation with the projectors $M_{t,j}$ and $A_{t,j}$, since it is obtained by first multiplying several projectors $((-\mathbf{D}_0)^{-1})_t$ and $\mathbf{P}_t$ and then weighting the result using the $\boldsymbol{\pi}_e$ probabilities. Therefore, moment and autocorrelation fitting algorithms which impose the eigenvalues θ_t and γ_l and the projectors $M_{t,j}$ and $A_{t,j}$, still leave degrees of freedom to assign the projectors of higher order moments. This observation is consistent with the results in [36].

4. Compositional definition of large processes

The accurate fitting of LRD traces requires models composed by many states; e.g., the MAP fittings of the Bellcore Aug89 trace in [2] and [17] employ $n = 16$ and $n = 32$ states, respectively. Since traditional superposition is not meant to impose higher order properties of the IAT process, we define a different process composition method which we call Kronecker Product Composition (KPC). Given J MAPs $\{\mathbf{D}_0^j, \mathbf{D}_1^j\}$, we define the KPC process as the MAP

$$\{\mathbf{D}_0^{\text{kpc}}, \mathbf{D}_1^{\text{kpc}}\} = \{(-1)^{J-1}\mathbf{D}_0^1 \otimes \cdots \otimes \mathbf{D}_0^J, \mathbf{D}_1^1 \otimes \cdots \otimes \mathbf{D}_1^J\}$$

where $\otimes$ is the Kronecker product operator [40]. It can be easily shown by the properties of the Kronecker product that $\mathbf{P}^{\mathrm{kpc}} = -(\mathbf{D}_0^{\mathrm{kpc}})^{-1}\mathbf{D}_1^{\mathrm{kpc}} = \mathbf{P}^1 \otimes \cdots \otimes \mathbf{P}^J$ and $\boldsymbol{\pi}_e^{kpc} = \boldsymbol{\pi}_e^1 \otimes \cdots \otimes \boldsymbol{\pi}_e^J$, thus our composition generates an embedded process $\mathbf{P}^{\mathrm{kpc}}$ with simple compositional structure.

In order to generate a valid MAP, the KPC requires that at least $J-1$ composing processes have diagonal $\mathbf{D}_0^j$ otherwise off-diagonal negative elements appear in $\mathbf{D}_0^{\mathrm{kpc}}$. Nevertheless, because one MAP can be arbitrary, the KPC does not place modeling restrictions.

The basic property of a MAP obtained by KPC is that we can easily impose its eigenvalues and projectors in both moments and autocorrelations as we show later in Theorem 3. Equivalently, one may impose directly moments and autocorrelation values, as described in Theorem 4. This is important because, by the characterization in Section 3, the fitting of real traces can be seen as an inverse eigenvalue problem for the eigenvalues of $\mathbf{P}$ and $(-\mathbf{D}_0)^{-1}$. Inverse eigenvalue problems are notoriously hard, but the KPC method provides an effective solution. A MAP(n) can be defined to match an arbitrary number of autocorrelation and moment values, with the only practical difficulty of limiting the order of the resulting MAP. In the rest of the section, we show how one can a priori determine moments and autocorrelations of the KPC process given the knowledge of the properties of the composing MAPs.

4.1. KPC process characterization

Without loss of generality we study $\{\mathbf{D}_0^{\mathrm{kpc}}, \mathbf{D}_1^{\mathrm{kpc}}\}$ for the case of composing by KPC $J = 2$ MAPs. The results presented here recursively characterize also the case $J > 2$.

Theorem 3. *Let* $\mathrm{MAP}_a = \{\mathbf{D}_0^a, \mathbf{D}_1^a\}$ *and* $\mathrm{MAP}_b = \{\mathbf{D}_0^b, \mathbf{D}_1^b\}$ *be MAPs of order* n_a *and* n_b*, respectively, and assume that* $\mathbf{D}_0^b$ *is a diagonal matrix. Let* $\gamma_p^a, \theta_p^a, A_{p,1}^a$ *and* $M_{q,1}^a$ *be the eigenvalues and projectors of* MAP_a*. Let* $\gamma_q^b, \theta_q^b, A_{p,1}^b$ *and* $M_{q,1}^b$ *be the equivalent descriptors of* MAP_b*. Then the KPC*

$$\mathrm{MAP}_a \otimes \mathrm{MAP}_b = \{-\mathbf{D}_0^a \otimes \mathbf{D}_0^b, \mathbf{D}_1^a \otimes \mathbf{D}_1^b\}$$

is a MAP of order $n_a n_b$ *with eigenvalues* $\gamma_t = \gamma_p^a \gamma_q^b$, $\theta_t = \theta_p^a \theta_q^b$, *and projectors*

$$M_{t,1} = M_{p,1}^a M_{q,1}^b, \quad A_{t,1} = (A_{p,1}^a SCV_a)(A_{q,1}^b SCV_b)/SCV,$$

for all $1 \le p \le n_a$, $1 \le q \le n_b$.

Proof. The relations for the eigenvalues follow from basic properties of the Kronecker product [40]. The projector associated to $\theta_t = \theta_p^a \theta_q^b$ is

$$\begin{aligned} M_{t,1} &= \boldsymbol{\pi}_e((-\mathbf{D}_0)^{-1})_t \mathbf{e} \\ &= (\boldsymbol{\pi}_e^a \otimes \boldsymbol{\pi}_e^b)((-\mathbf{D}_0)^{-1})_p^a \otimes ((-\mathbf{D}_0)^{-1})_q^b (\mathbf{e}^a \otimes \mathbf{e}^b) \\ &= (\boldsymbol{\pi}_e^a((-\mathbf{D}_0)^{-1})_p^a \mathbf{e}^a)(\boldsymbol{\pi}_e^b((-\mathbf{D}_0)^{-1})_q^b \mathbf{e}^b) = M_{p,1}^a M_{q,1}^b. \end{aligned}$$

Similarly, the projector of $\gamma_t = \gamma_p^a \gamma_q^b$ is

$$\begin{aligned} A_{t,1} &= E[X]^{-2} \boldsymbol{\pi}_e(-\mathbf{D}_0)^{-1} \mathbf{P}_t (-\mathbf{D}_0)^{-1} \mathbf{e}/SCV, \\ &= E[X]^{-2}(E[X^a]^2 A_{p,1}^a SCV_a)(E[X^b]^2 A_{q,1}^b SCV_b)/SCV, \\ &= (A_{p,1}^a SCV_a)(A_{q,1}^b SCV_b)/SCV. \quad \square \end{aligned}$$

Theorem 4. *Moments and autocorrelations of the KPC satisfy*

$$E[X^k] = E[X_a^k]E[X_b^k]/k!, \tag{22}$$

$$SCV \rho_k = (SCV_a)\rho_k^a + (SCV_b)\rho_k^b + (SCV_a SCV_b)\rho_k^a \rho_k^b, \tag{23}$$

where the quantities in the right-hand side refer to MAP^a *and* MAP^b*. In particular the relation for* $E[X^k]$ *immediately implies*

$$1 + SCV = (1 + SCV_a)(1 + SCV_b)/2. \tag{24}$$

Proof. We begin by proving (22). Using the properties of the Kronecker product [40] we have

$$\begin{aligned} E[X^k] &= k! \boldsymbol{\pi}_e (-\mathbf{D}_0)^{-k} \mathbf{e} \\ &= k!(\boldsymbol{\pi}_e^a \otimes \boldsymbol{\pi}_e^b)(-((-1)^{2-1}\mathbf{D}_0^a \otimes \mathbf{D}_0^b))^{-k}(\mathbf{e}^a \otimes \mathbf{e}^b) \\ &= k!(\boldsymbol{\pi}_e^a \otimes \boldsymbol{\pi}_e^b)((\mathbf{D}_0^a)^{-k} \otimes (\mathbf{D}_0^b)^{-k})(\mathbf{e}^a \otimes \mathbf{e}^b), \end{aligned}$$

and multiplying by $(-1)^{-2k}$ which equals one for all $k \in \mathbb{N}$

$$\begin{aligned} E[X^k] &= k!(-1)^{-2k}(\boldsymbol{\pi}_e^a(\mathbf{D}_0^a)^{-k}\mathbf{e}^a)(\boldsymbol{\pi}_e^b(\mathbf{D}_0^b)^{-k}\mathbf{e}^b) \\ &= k!(\boldsymbol{\pi}_e^a(-\mathbf{D}_0^a)^{-k}\mathbf{e}^a)(\boldsymbol{\pi}_e^b(-\mathbf{D}_0^b)^{-k}\mathbf{e}^b) \\ &= E[X_1^k]E[X_2^k]/k!. \end{aligned}$$

Eq. (23) follows the same steps as (22) by considering (2). □

The two theorems presented above provide a complete characterization of moments and autocorrelations of the KPC process. The KPC also simplifies the definition of higher order statistics.

Theorem 5. *The joint moments of $MAP_a \otimes MAP_b$ satisfy*

$$H(\vec{i},\vec{k}) = \frac{H^a(\vec{i},\vec{k})H^b(\vec{i},\vec{k})}{k_1!k_2!\cdots k_L!}, \tag{25}$$

with $H^a(\vec{i},\vec{k})$ and $H^b(\vec{i},\vec{k})$ respectively joint moments of MAP_a and MAP_b.

Proof. We have

$$\begin{aligned} H(\vec{i},\vec{k}) &= \boldsymbol{\pi}_e\left(\prod_{l=1}^{L} k_l!(-\mathbf{D}_0)^{-k_l}\mathbf{P}^{i_l-i_{l-1}}\right)\mathbf{e} \\ &= (\boldsymbol{\pi}_e^a \otimes \boldsymbol{\pi}_e^b)\left(\prod_{l=1}^{L} k_l!(-\mathbf{D}_0^a \otimes -\mathbf{D}_0^b)^{-k_l}(\mathbf{P}^a \otimes \mathbf{P}^b)^{i_l-i_{l-1}}\right)\mathbf{e} \end{aligned}$$

and using commutativity of Kronecker products

$$\begin{aligned} H(\vec{i},\vec{k}) &= \left(\pi_e^a\left(\prod_{l=1}^{L} k_l!(-\mathbf{D}_0^a)^{-k_l}(\mathbf{P}^a)^{i_l-i_{l-1}}\right)\mathbf{e}\right)\left(\boldsymbol{\pi}_e^b\left(\prod_{l=1}^{L}(-\mathbf{D}_0^b)^{-k_l}(\mathbf{P}^b)^{i_l-i_{l-1}}\right)\mathbf{e}\right) \\ &= H^a(\vec{i},\vec{k})\frac{H^b(\vec{i},\vec{k})}{k_1!k_2!\cdots k_L!} \end{aligned}$$

which proves the theorem. □

5. MAP fitting algorithm

Using KPC, we define a MAP fitting algorithm for trace data. We illustrate the algorithm in the case where the J composing MAPs used in the KPC are an arbitrary MAP(2) (index $j = 1$) and $J - 1$ MAP(2)s with diagonal $\mathbf{D}_0$, but the method works with minor modifications also with other processes as we discuss in Section 6.5. The fitting algorithm searches for J MAP(2)s that composed by KPC can match accurately the first three moments, the autocorrelations and the bicorrelations of the trace.

5.1. MAP fitting algorithm

The MAP fitting algorithm proceeds in three steps:

Step 1 — *Autocorrelation and SCV fitting.* Let $\widehat{SCV}$ be the sample SCV; similarly, let $\widehat{\rho}_k$ be the sample autocorrelation computed on a set of lags $\mathbf{K}$, and let $\mathbf{J} = \{1, 2, \ldots, J\}$. We fit second order IAT properties by the nonlinear optimization program in Fig. 2. The fitting algorithm is essentially a least square algorithm constrained by the properties of the KPC. The result of the optimization are two sets $SCV(j)$ and $\gamma_2(j)$ for $j \in \mathbf{J}$ which specify the optimal SCV and autocorrelation for each of the J MAPs used in the KPC. For each variable, a set of upper and lower bounds are imposed, e.g., $\mathrm{ub}_{SCV}(j)$ and $\mathrm{lb}_{SCV}(j)$ are respectively upper and lower bounds on the value $SCV(j)$ to be determined by the solver. Since $SCV(j)$ and $\gamma_2(j)$ are constrained by proper bounds, they can be always chosen to be feasible for a MAP(2), see [7,30] for existing bound formulas. In particular, we set the upper bound on the SCV to be $\mathrm{ub}_{SCV}(j) = \infty, j \in \mathbf{J}$. Further, for the arbitrary MAP(2) we have

$$\mathrm{lb}_{SCV}(1) = 0.5, \qquad \mathrm{lb}_{\gamma_2}(1) = -1, \qquad \mathrm{ub}_{\gamma_2}(1) = 1 - \epsilon,$$

where ϵ is an arbitrarily small positive quantity. The $J - 1$ MAP(2)s with diagonal $\mathbf{D}_0$ can be shown to have hyper-exponential marginal probabilities, and thus we set

$$\mathrm{lb}_{SCV}(j) = 1 + \epsilon, \qquad \mathrm{lb}_{\gamma_2}(j) = 0, \qquad \mathrm{ub}_{\gamma_2}(j) = 1 - \epsilon.$$

The value tol_{SCV} is a tolerance on the exact matching of the SCV. On certain traces where the value of the lag-1 autocorrelation ρ_1 differs significantly from $\rho_0 = (1 - 1/SCV)/2$, flexibility on the SCV fitting avoids an excessive constraining to impose the passage through ρ_1 which can result in bad fitting of autocorrelation at high lags.

$$
\begin{array}{ll}
\textbf{minimize } \sum_{k\in K}(\rho_k-\widehat{\rho}_k)^2 & \\
\textbf{subject to} & \\
(SCV-\widehat{SCV})^2 \le \mathrm{tol}_{SCV}, & \\
\mathrm{lb}_{SCV}(j) \le SCV(j) \le \mathrm{ub}_{SCV}(j), & \forall\, j \in \mathbf{J}; \\
\mathrm{lb}_{\gamma_2}(j) \le \gamma_2(j) \le \mathrm{ub}_{\gamma_2}(j), & \forall\, j \in \mathbf{J}; \\
\textbf{where} & \\
\widehat{SCV} \leftarrow \text{sample } SCV, & \\
\widehat{\rho_k} \leftarrow \text{sample autocorrelation}, & \forall\, k \in \mathbf{K} \\
SCV \leftarrow (24) \text{ recursively using } SCV(j), & \forall\, j \in \mathbf{J}, \\
\rho_k(j) \leftarrow \frac{1}{2}\left(1-\frac{1}{SCV(j)}\right)\gamma_2(j)^k, & \forall\, k \in \mathbf{K}, \forall\, j \in \mathbf{J}; \\
\rho_k \leftarrow (23) \text{ recursively using } \rho_k(j), & \forall\, k \in \mathbf{K}, \forall\, j \in \mathbf{J}.
\end{array}
$$

Fig. 2. Autocorrelation and *SCV* fitting [Step 1].

Step 2 — *Moment and higher order fitting*. Once that the optimal values of $SCV(j)$ and $\gamma_2(j)$ are obtained after one or more runs[2] of the previous algorithm, we search for the missing parameters required to define valid MAP(2)s, namely the means $E[X](j)$ and third moments $E[X^3](j)$ for all $j \in \mathbf{J}$. Indeed, the second moments $E[X^2](j)$ are readily obtained from the $SCV(j)$ for given $E[X](j)$. As shown by the motivating example in Fig. 1, given fixed autocorrelation and *SCV* there exist many possible valid processes; we thus solve a new nonlinear optimization program to select the one that results in better fitting of higher order properties of IATs on a set of sample joint moments $\widehat{H}(\vec{i},\vec{k})$ for $(\vec{i},\vec{k}) \in \mathbf{H}$. The nonlinear program is given in Fig. 3. In our MATLAB implementation, at each iteration of the solver a $(\mathbf{D}_0, \mathbf{D}_1)$ representation is obtained for each of the J MAP(2)s similarly to the way described below in Step 3, but yet without assembling the results into a MAP(n). This approach grants MAP feasibility at each point of the iteration. The joint moments of the MAP(2)s generated at each iteration are used to update the value of the objective function since the joint moments of the composed process are in simple relation with those of the composing MAPs [39]. Indeed, this approach imposes an overhead at each iteration, but the experimental results we present in the paper required very modest computational times of the order of a few minutes per fitting and therefore the cost of this interleaved estimation of the MAP(2)s is acceptable.

Because of the approach used, whenever the optimizer steps into the infeasibility region of this arbitrary MAP(2), the value of the objective function is forced back to the value in a point where the MAP(2)s are all feasible because of the feasibility corrections applied to the MAP(2). Indeed, it is possible that the optimizer gets stuck iterating into the same region. This case is detected by evaluating the progress in the objective function over a time window of iterations, then the optimization stops and returns the currently estimated MAP. If unsatisfied with the result, the analyst is left the choice of re-running Step 2 from a different initialization point without the need of restarting from Step 1.

Finally, we remark that in this step we use the following moment bounds for the hyper-exponential MAP(2)s [7]

$$
\begin{aligned}
&\mathrm{lb}_{E[X]}(j) = \sqrt{E[X^2]/2}, \qquad \mathrm{ub}_{E[X]}(j) = +\infty,\\
&\mathrm{lb}_{E[X^3]}(j) = \sqrt{(1.5+\epsilon)E[X^2]^2/E[X]}, \qquad \mathrm{ub}_{E[X^3]}(j) = +\infty.
\end{aligned}
$$

Step 3 — MAP(n) *generation*. Given the target optimal values for the $E[X](j)$, $SCV(j)$, $E[X^3](j)$, and $\gamma_2(j)$ we generate the J MAPs as follows. The $J-1$ diagonal MAPs are always feasible since the constraints on moments and autocorrelations are sufficient for feasibility [7,30]. These are fitted using the analytic hyper-exponential fitting scheme in [39], but more general MAP(2) methods may be used as well. For the arbitrary MAP(2) we use standard fitting algorithms, see e.g., [12,5]. Whenever the fitting gives an infeasible process (e.g., negative rates in $\mathbf{D}_1$ or in the off-diagonal elements of $\mathbf{D}_0$), we perform a simple least square fitting to best match the target $E[X^3](j)$ and $\gamma_2(j)$, while keeping fixed $E[X](j)$ and $E[X^2](j)$. Once that J feasible MAPs are obtained, the final process is immediately computed by Kronecker products according to the KPC definition.

We conclude the section by remarking that with MMPP(2)s/MAP(2)s, the fitting algorithm cannot include complex eigenvalues in the IAT autocorrelations. These may be included by also using one or more circulant MAP(3)s (see [26]), but this may easily yield processes with several tens or hundreds of states. This state space explosion associated to the use of circulant matrices has been pointed out also in the fitting of the counting process [41] and remains an open problem. However, we empirically observe that many traces that exhibit multiple complex eigenvalues in the counting process often have IAT autocorrelation that does not require complex eigenvalues, and this makes MAP(2)-based IAT fitting sufficient more frequently than counting process-based methods. For instance, Fig. 4 compares the Welch power spectrum density (PSD) estimate of the IAT and counting processes on the Bellcore Aug89 trace. The counting process is obtained by computing

[2] The term algorithm "run" is used in the rest of the paper meaning that the optimization algorithm is restarted each time with different random initializations.

$$
\begin{array}{ll}
\textbf{minimize} \sum_{(\vec{i},\vec{k})\in\mathbf{H}} (H(\vec{i},\vec{k}) - \widehat{H}(\vec{i},\vec{k}))^2 & \\
\textbf{subject to} & \\
(E[X] - \widehat{E}[X])^2 \le \mathrm{tol}_{\mathrm{E[X]}}, & \\
(E[X^3] - \widehat{E}[X^3])^2 \le \mathrm{tol}_{\mathrm{E[X^3]}}, & \\
\mathrm{lb}_{\mathrm{E[X]}}(j) \le E[X](j) \le \mathrm{ub}_{\mathrm{E[X]}}(j), & \forall\, j \in \mathbf{J}; \\
\mathrm{lb}_{\mathrm{E[X^3]}}(j) \le E[X^3](j) \le \mathrm{ub}_{\mathrm{E[X^3]}}(j), & \forall\, j \in \mathbf{J}; \\
\textbf{where} & \\
\widehat{E}[X] \leftarrow \text{sample E[X]}, & \\
\widehat{E}[X^3] \leftarrow \text{sample E[X}^3\text{]}, & \\
E[X] \leftarrow \text{(22) recursively using } E[X](j), & \forall\, j \in \mathbf{J}, \\
E[X^3] \leftarrow \text{(22) recursively using } E[X^3](j), & \forall\, j \in \mathbf{J}, \\
E[X^2](j) \leftarrow (1 + SCV(j))(E[X](j))^2, & \forall\, j \in \mathbf{J}.
\end{array}
$$

Fig. 3. Moment and higher order fitting [Step 2].

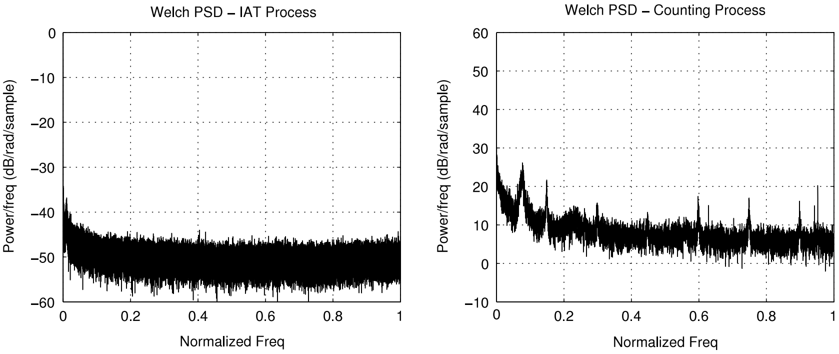

Fig. 4. Comparison of the power spectrum of the IAT process and the counting process for the Bellcore Aug89 trace. The counting process shows power density in the complex spectrum which is instead negligible in IATs.

the arrivals in 10^5 consecutive times slots of identical duration $\Delta T = 10^{-2}$ s. The figure for the counting process indicates power in the low frequency spectrum, whereas the IAT process does not show any significant complex sinusoid and thus can be approximated effectively by real eigenvalues only.

6. Experiments

We present a comparison of our algorithm with two of the best-available algorithms for Markovian analysis of LRD traces, that is, the method of Andersen and Nielsen (A&N) in [2] and the multifractal approach of Horváth and Telek (H&T) in [17]. We have also performed experiments against the methods presented in [32,20], but we have found the techniques in [2,17] the most competitive against our approach on the set of traces considered in this section. We first describe the experimental methodology, later we report fitting results on the Bellcore Aug89 trace [15,16] and a recently measured Web trace of Seagate disk drives and presented in [34]. All the nonlinear optimization problems considered in this section have been solved with the fmincon function of the MATLAB optimization toolkit.

6.1. Experimental methodology

We apply the algorithm described in Section 5 as follows. We first fit the autocorrelation on a set of 10^3–10^4 logarithmically-spaced lags ranging in a large interval, e.g., $[1, 10^5]$. Previous work has often limited to match IAT autocorrelations in $[1, 10^4]$, but we have observed that the choice of a larger lag interval can result in increased modeling accuracy at heavy load where second order properties are fundamental for queueing prediction[42]. The solution of the least squares program in Fig. 2 is usually very efficient (of the order of seconds), and only a few tens of runs are needed for an accurate match. Here we consider four MAPs ($J = 4$); good fitting of the autocorrelation is also possible with only two or three MAPs, but the remaining degrees of freedom are usually insufficient to match accurately higher order properties of IATs.

In the fitting of the joint moments, we have performed several experiments and obtained the best results by matching a set of moments $E[X_{i_1}X_{i_2}X_{i_3}]$, which implicitly define the bicorrelations of the IAT process[29]. In our tests this seemed more important than matching moments $E[X_{i_1}^{k_1}X_{i_2}^{k_2}]$ of the IATs, which did not result in improved queueing prediction accuracy

Table 1
MAP(16) fitting of the Bellcore Aug89 Trace using the algorithm of Section 5.

BC-Aug89	Trace	MAP(16)
$E[X]$	3.1428×10^{-3}	3.1428×10^{-3}
SCV	3.2236×10^{0}	3.2235×10^{0}
$E[X^3]$	2.0104×10^{-6}	1.1763×10^{-5}
γ_2	n/a	9.9995×10^{-1}

with respect to a standard second order fitting. Without loss of generality we set $i_1 = 1$ and fit $E[X_1X_{i_2}X_{i_3}]$ on a square grid of 10^2 or 25^2 points (i_2, i_3) generated by the Cartesian product of two identical sets of logarithmically-spaced points in $[1, 10^4]$. The point $E[X_1X_1X_1] = E[X^3]$ is always included in this grid, thus in Step 2, see Fig. 3, we set $\text{tol}_{E[X^3]} = +\infty$ to give more flexibility to the least squares; in all experiments we instead impose exact matching of $E[X]$ and SCV, thus $\text{tol}_{E[X]} = 0$ and $\text{tol}_{SCV} = 0$. Compared to the autocorrelation, the least square fitting of joint moments seems more difficult and the nonlinear optimizer can occasionally return infeasible solutions. Thus, several runs may be needed to find a good local optimum, which is nevertheless obtained in a few minutes.

The computational costs of the final MAP(n) generation is negligible. We also remark that small manual corrections of erroneous behaviors are possible without the need of re-running the entire fitting algorithm. For instance, to obtain a slower asymptotic decay rate for the autocorrelations it is possible to increase the value of the largest $\gamma_2(j)$ and regenerate the MAP(n).

Finally, the evaluation of the queueing behavior of the fitted MAP is done with an implementation of the analytical method for the solution of a MAP/D/1 process in [43] and using a numerical tolerance for convergence of $\epsilon = 10^{-10}$. Details on the experimental results are given in the rest of the section.

6.2. Bellcore Aug89, $-/D/1$ queue

We first compare with the queueing predictions of the models in [2,17] using the Bellcore Aug89 on a first-come-first-served queue with deterministic service and different utilization levels. This is the standard case for evaluating the quality of LRD trace fitting, e.g., [2,17,20]. The traffic trace consists of 10^6 IAT samples collected in 1989 at the Bellcore Morristown Research and Engineering facility and shows a clear LRD behavior, see [16] for details. We run the algorithm described in Section 5 to determine a MAP(16) which accurately fits the trace.

The size of this MAP is similar to those employed in previous work, which are composed by 16 states (A&N) or 32 states (H&T). We have experimented with other MAP orders, but we have never obtained results qualitatively better than the MAP(16) ones using less states (i.e., MAP(2), MAP(4), and MAP(8)), while the MAP(32) does not improve significantly prediction accuracy. As we show later in Section 6.5, the last observation applies only to KPC fitting based on MAP(2)s as building blocks. Furthermore, we point the interested reader to [35] for a MAP(2) fitting of the Bellcore Aug89 trace that shows the severe errors of small MAPs in predicting accurately queue-lengths under correlated workloads. Due to the limited length of the trace, we fit all autocorrelations in the interval $[1, 2 \times 10^4]$, since at higher lags the sample values are significantly affected by noise. The result of this fit is rather accurate, as shown in Fig. 5, and is obtained in less than one minute[3]. In the second phase of the algorithm, the joint moments $E[X_1X_{i_2}X_{i_3}]$ are matched on a square grid of 25^2 points. On this instance, the computational cost of the program in Fig. 3 is low, approximately thirty seconds. The values of the first three moments of the MAP(16) are given in Table 1; the entries of each composing MMPP(2)/MAP(2) are given in the Appendix.

In order to assess the accuracy of the fit, we compare the queueing prediction of our model with the MMPPs obtained in [2,17] for utilization levels of 20%, 50%, and 80%. All traces have a quite good match of the individual queue probabilities. In Fig. 6 we plot the complementary cumulative distribution function (ccdf) of queue-length probabilities, i.e., the function Pr(queue $\geq x$), which accounts also for the residual queueing probability mass and thus shows the impact of the tail probability. At 20% utilization the effects of the long-range dependence seems minimal, and the probability mass is spread over few lags. Our method gives almost the same results of the multifractal technique, while the method of A&N seems to underestimate the queueing probability for the smallest values of x, which also affects the rest of the ccdf.

The intermediate case for 50% utilization is generally difficult to capture, since the network is approaching heavy traffic, but the dependence effects are still not as strong as in slightly higher utilization values, i.e., for 60%–70% utilization (see, e.g., [2]). All methods initially overestimate the real probability, but for higher values of x our method is closer to the trace values than A&N and H&T which predict a large probability mass also after $x = 10^3$.

Finally, in the case of 80% utilization all three methods perform well, with our algorithm and the H&T being more precise than A&N. The final decay of the curve is again similar, but the KPC method resembles better the simulated trace.

Overall, the result of this trace indicates that the KPC approach is more effective than both H&T and the A&N methods, while preserving the smallest representation (16 states) of the A&N method. It also interesting to point out that the fitting leaves room for further improvements, especially in the 50% case which is difficult to approximate.

[3] In both Figs. 5 and 8 we do not report the acf fitting of A&N and H&T since these methods do not match IAT autocorrelations, but autocorrelations in counts.

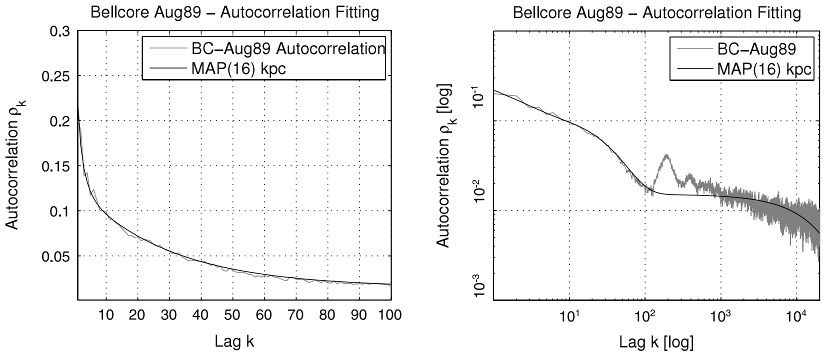

Fig. 5. Fitted autocorrelation for the Bellcore Aug89 trace using the program in Fig. 2.

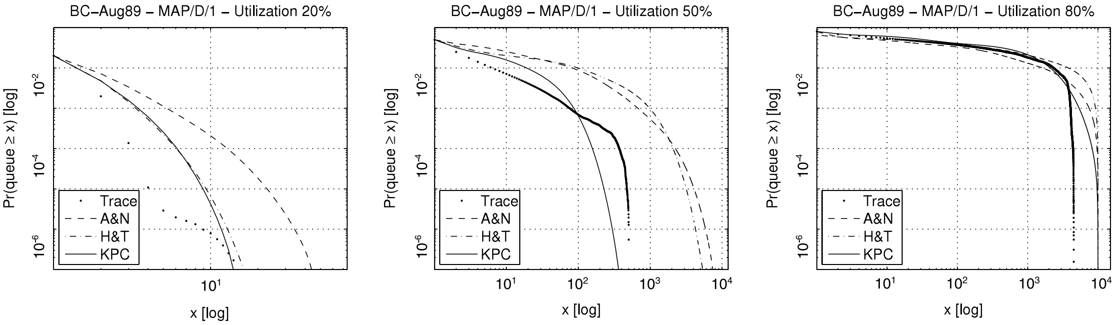

Fig. 6. Queueing predictions for the Bellcore Aug89 trace on a queue with deterministic service.

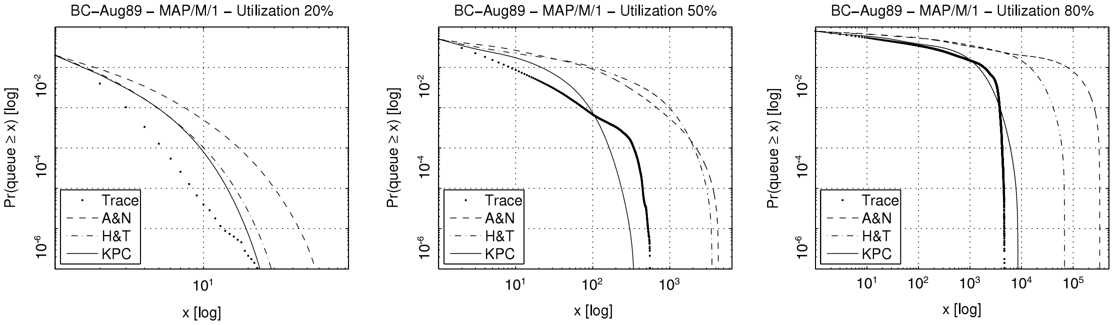

Fig. 7. Queueing predictions for the Bellcore Aug89 trace on a queue with exponential service.

6.3. Bellcore Aug89, $-/M/1$ queue

In the second experiment we evaluate the robustness of the fitting under different variability in the service process. This is important to assess that the fitting captures the essential properties of the traffic process, and thus can provide accurate results regardless of the context in which the fitted MAP is used. In Fig. 7 we plot comparative results for a $-/M/1$ queue using as input the same MAPs considered before. As we can see, KPC performs better than in the $-/D/1$ case, and it is now able to capture well the tail decay also for the 80% utilization. A possible explanation of this behavior is that the autocorrelation in the flow becomes more important if the queueing process is more variable, therefore more accurate autocorrelation fitting becomes necessary under such conditions. In comparison, the other methods seem instead to suffer by the increase in variability of the process, as shown by the overestimates which are significantly greater than in the $-/D/1$ case. This indicates that KPC is more robust than counting-process-based fittings.

6.4. Seagate web trace, $-/D/1$ queue

In order to provide a comparison on traces that are representative of other workloads, we have implemented the A&N method and compared its counting-process fitting with our method on the HTTP web trace presented in [34]. The trace is composed by 3.6×10^6 interarrival times of requests at the storage system of a Web server, and has a long-range

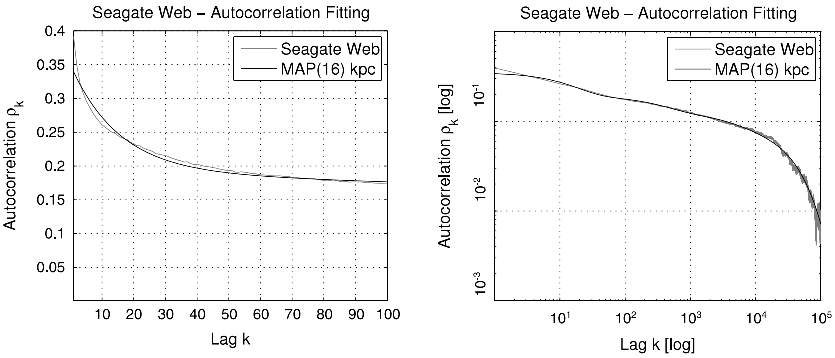

Fig. 8. Fitted autocorrelation for the Seagate Web trace using the program in Fig. 2.

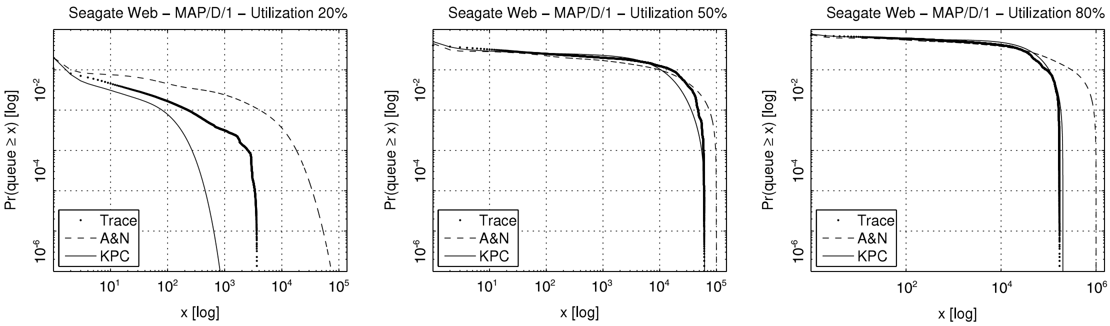

Fig. 9. Queueing predictions for the Seagate Web trace on a queue with deterministic service.

Table 2
MAP(16) fitting of the Seagate Web Trace using the algorithm of Section 5.

Seagate	Trace	MAP(16)
$E[X]$	3.0134×10^0	3.0134×10^0
SCV	3.3285×10^0	3.3285×10^0
$E[X^3]$	1.5986×10^3	1.1414×10^3
γ_2	n/a	9.9997×10^{-1}

dependence that is stronger than the BC-Aug89, see [34] for the Hurst coefficient estimates. Thanks to the larger size of the sample, we now fit the autocorrelation in the larger set of lags $[1, 10^5]$ using only 10^3 logarithmically-spaced points since the autocorrelation function is less noisy than for the Bellcore trace, see Fig. 8. The joint moments are then fitted on a grid of 10^2 points. The order of the target MAP is 16 states similarly to the Bellcore Aug89 case. The values of the first three moments of the fitted MAP(16) are reported in Table 2; each of the composing MMPP(2)/MAP(2) are given in the Appendix.

Queueing results for this trace are shown in Fig. 9. Here we compare with an implementation of the A&N algorithm [2]. The A&N MAP(16) fitting is obtained by the algorithm parameters $H = 0.85682$, $\rho = 0.74503$, $\lambda^{\star} = 3.3185$, $n = 5$, $d^{\star} = 4$.

Although the performance effects of Web traffic on a server is more often modeled by a queue with exponential service, we perform the comparison here by assuming a deterministic service time, since the results on the Bellcore trace indicate that this case is more difficult to approximate. Predictions on a $-/D/1$ queue at utilization levels of 20%, 50%, 80% are shown in Fig. 9. The KPC method is more accurate than the A&N fitting in the cases 50% and 80% while the case 20% is hard to approximate for both methods. This reinforces the validity of the observations on the Bellcore trace: IAT fitting is more effective as soon as the effect of the temporal dependence becomes evident. The 50% and 80% utilization levels for the KPC method are cases of almost perfect fits. In particular, for the 50% case the analytical results indicate that the tail probability is zero with respect to machine accuracy for $x = 61891$, while the simulated queue drops to zero for $x = 61002$.

6.5. Generalization of KPC fitting

We conclude the experimental part by providing some discussion on the generalization of the KPC fitting algorithm to building blocks different from the MAP(2). After choosing a new building block, the KPC fitting algorithm presented in Section 5 should be modified by changing the control variables considered in the optimization to those needed to specify the new building block. Upper and lower bounds that constraint the feasibility range of these control variables should also be added to the optimization programs. Furthermore, it is required that Step 1 of SCV and autocorrelation fitting can be

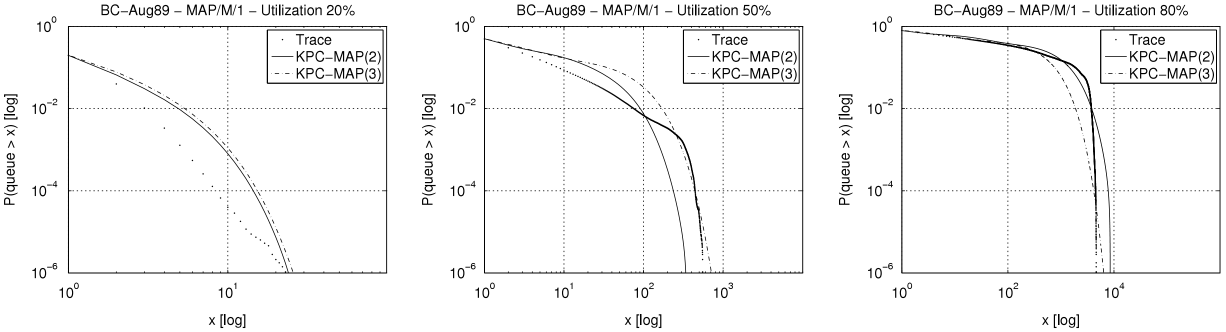

Fig. 10. Comparison of queueing predictions of two independent KPC fittings for the Bellcore Aug89 trace on a queue with exponential service: the MAP(16) built from MAP(2)s presented in Section 6.3 and a MAP(9) obtained from KPC of two MAP(3)s with special structure.

performed separately from Step 2; this condition is always granted whenever the projectors $A_{t,j}$ of the autocorrelation function and the eigenvalues γ_k can be determined without need of information on the control variables determined in Step 2.

As an example, we have considered as building block a special class of MAP(3) described in the Appendix. These MAP(3)s enjoy simple formulas for the characterization of moments and autocorrelations. Also the range of feasibility of the model parameters can be obtained easily. These MAP(3)s have autocorrelation coefficients expressed as

$$\rho_k = \frac{\rho_0}{2}(\gamma_2^k + \gamma_3^k), \tag{26}$$

where γ_2 and γ_3 are two independent eigenvalues of **P**. Thus, the autocorrelation function of the considered MAP(3) is more versatile than the one of a MAP(2) process which has a single eigenvalue. Fig. 10 illustrates fitting results on the BC-Aug89 trace obtained with KPC using two MAP(3)s as building blocks, instead of the four MAP(2)s considered in previous fittings; the process fitted by KPC is thus a MAP(9) instead of a MAP(16). The figure illustrates queueing results of a MAP/M/1 system for different utilization levels; the same results shown in Fig. 7 for the MAP(2)-based KPC are here reported for the sake of comparison with the MAP(3)-based approach. The results show that an increase of flexibility in the description of the temporal dependence as in (26) immediately improves the queueing prediction accuracy. Strikingly, a MAP process with only 9 states is now able to achieve accuracy levels that are comparable to the MAP(16) process fitted by MAP(2)-based KPC; in particularly, it is remarkable that for utilization 50% the fitted MAP(9) has much increased accuracy compared to MAP(2)-based KPC which underestimates the decay rate of the queueing probabilities.

Summarizing, this subsection has shown that generalization of the proposed approach to building blocks other than MAP(2) is viable and can be profitable to increase MAP fitting accuracy. However, the preliminary characterization of moments, autocorrelation, and ranges of parameters required to perform this generalization inevitably limits the attention to building blocks of MAPs of order no more than 2 or 3 states which are often the only ones that can be characterized and fitted analytically. We remark that we have only experimented the KPC fitting algorithm using as building blocks MAPs with 2 or 3 states only. This is because analytical characterization is important to make the KPC-based fitting algorithm efficient computationally, thus we believe that only small with no more than 2 or 3 states should be used as building blocks of the proposed fitting algorithm.

7. Conclusion

We have presented several contributions to the Markovian analysis of measured traces described in terms of packet or request interarrival times. We have obtained a spectral characterization of moments and autocorrelation which simplifies the analysis of MAPs. Then, we have studied the definition of large MAPs by Kronecker Product Composition (KPC), and shown that this provides a simple way to create processes with predefined moments and correlations at all orders. A least square fitting procedure based on the properties of these processes has been described. Detailed comparisons with other state-of-the-art fitting methods based on the counting process show that KPC provides improved fitting of LRD traces that require models that capture their higher order properties, including the challenging BC-Aug89 trace of the Internet Traffic Archive. An open-source MATLAB implementation of the MAP(2)-based fitting algorithm can be found in the KPC-Toolbox [35] at http://www.cs.wm.edu/MAPQN/kpctoolbox.html.

Acknowledgments

This work was supported by the National Science Foundation under grants CNS-0720699 and CCF-0811417 and by the InvestNI/SAP MORE project.

The authors thank M. Telek for his comments on an earlier version of this manuscript; B. Van Houdt for his help for the analytical solution of the MAP/D/1 queue; A. Riska for providing the Seagate Web traffic trace; A. Horváth for comments and material related to the multifractal method.

Appendix

KPC fitting — Bellcore Aug89 trace. The Bellcore Aug89 trace is fitted by the KPC process $(\mathbf{D}_0^{kpc}, \mathbf{D}_1^{kpc})$ defined by

$$\begin{aligned}\mathbf{D}_0^{kpc} &= -\mathbf{D}_0^a \otimes \mathbf{D}_0^b \otimes \mathbf{D}_0^c \otimes \mathbf{D}_0^d\\ \mathbf{D}_1^{kpc} &= \mathbf{D}_1^a \otimes \mathbf{D}_1^b \otimes \mathbf{D}_1^c \otimes \mathbf{D}_1^d.\end{aligned}$$

The four composing processes have

$$\mathbf{D}_1^a = \begin{bmatrix} 2.5582 \times 10^0 & 4.3951 \times 10^{-2} \\ 1.1369 \times 10^{-2} & 6.6173 \times 10^{-1} \end{bmatrix}, \qquad \mathbf{D}_1^b = \begin{bmatrix} 2.6769 \times 10^0 & 6.6924 \times 10^{-5} \\ 4.2706 \times 10^{-5} & 1.7082 \times 10^0 \end{bmatrix},$$

$$\mathbf{D}_1^c = \begin{bmatrix} 4.3309 \times 10^0 & 2.7061 \times 10^{-4} \\ 6.7564 \times 10^{-2} & 2.2578 \times 10^{-2} \end{bmatrix}, \qquad \mathbf{D}_1^d = \begin{bmatrix} 3.5552 \times 10^1 & 2.9355 \times 10^{-1} \\ 2.6962 \times 10^0 & 4.8230 \times 10^0 \end{bmatrix}$$

and the corresponding $\mathbf{D}_0$ are diagonal with i-th element equal in modulus to the sum of the i-th row of the associated $\mathbf{D}_1$ matrix, e.g., $\mathbf{D}_0^a = \mathrm{diag}(-(2.6769 \times 10^0 + 6.6924 \times 10^{-5}), -(4.2706 \times 10^{-5} + 1.7082 \times 10^0))$.

KPC fitting — Seagate web trace. The MAP(16) fitting the Seagate Web trace is the process $(\mathbf{D}_0^{kpc}, \mathbf{D}_1^{kpc})$ where

$$\begin{aligned}\mathbf{D}_0^{kpc} &= -\mathbf{D}_0^a \otimes \mathbf{D}_0^b \otimes \mathbf{D}_0^c \otimes \mathbf{D}_0^d\\ \mathbf{D}_1^{kpc} &= \mathbf{D}_1^a \otimes \mathbf{D}_1^b \otimes \mathbf{D}_1^c \otimes \mathbf{D}_1^d\end{aligned}$$

in which the composing processes have

$$\mathbf{D}_1^a = \begin{bmatrix} 6.0174 \times 10^{-4} & 1.9726 \times 10^{-5} \\ 5.4983 \times 10^{-6} & 1.6772 \times 10^{-4} \end{bmatrix}, \qquad \mathbf{D}_1^b = \begin{bmatrix} 4.7919 \times 10^1 & 6.4534 \times 10^{-2} \\ 2.8556 \times 10^{-2} & 2.1204 \times 10^1 \end{bmatrix},$$

$$\mathbf{D}_1^c = \begin{bmatrix} 4.4827 \times 10^0 & 5.7367 \times 10^{-5} \\ 1.6440 \times 10^{-5} & 1.2846 \times 10^0 \end{bmatrix}, \qquad \mathbf{D}_1^d = \begin{bmatrix} 2.9941 \times 10^1 & 3.6688 \times 10^{-3} \\ 1.9573 \times 10^{-3} & 1.5974 \times 10^1 \end{bmatrix}$$

and the corresponding $\mathbf{D}_0$ matrices are again diagonal with i-th element equal in modulus to the sum of the i-th row of the associated $\mathbf{D}_1$ matrix.

A special class of MAP(3)s. We define a class of specialized MAP(3)s that is useful for KPC fitting. This class of MAP(3) is defined by

$$\mathbf{D}_0 = \begin{bmatrix} -\theta_1^{-1} & 0 & 0 \\ 0 & -\theta_2^{-1} & 0 \\ 0 & 0 & -\theta_3^{-1} \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} 1-q-p & p & q \\ p & 1-2p & p \\ q & p & 1-q-p \end{bmatrix}$$

where $0 \le p \le 1$, $0 \le q \le 1$, and $\theta_k \ge 0$ for $k = 1, 2, 3$; a $(\mathbf{D}_0, \mathbf{D}_1)$ representation is immediately obtained by setting $\mathbf{D}_1 = -\mathbf{D}_0\mathbf{P}$. Because of the diagonal $\mathbf{D}_0$, the MAP(3) always describes a process with hyper-exponential marginal probabilities ($SCV \ge 1$). From the characterization results presented in Section 3, it follows with simple algebra that the process admits the following characterization of moments and autocorrelations

$$E[X^k] = \frac{k!}{3}\sum_{t=1}^{3}\theta_t^k, \quad \rho_k = A_{2,1}\gamma_2^k + A_{3,1}\gamma_3^k, \quad A_{2,1} + A_{3,1} = \rho_0,$$

where $A_{2,1}$ and $A_{3,1}$ depend on the θ_k values. In order to obtain a class of process that can be useful for the KPC fitting algorithm in Section 5, it is useful to consider a case where we can decouple the analysis of SCV and autocorrelations from the analysis of the other moments and of the bicorrelations. This decoupling can be obtained by removing the dependence on the θ_k's in the autocorrelation function expression and specifically in the $A_{2,1}$ and $A_{3,1}$ terms. We have found that by setting

$$\theta_3^{-1} = (2+\sqrt{3})\theta_2^{-1}\theta_1^{-1}/(\theta_2^{-1} + (1+\sqrt{3})\theta_1^{-1})$$

one immediately defines a specialized MAP(3) in which $A_{2,1} = A_{3,1} = \rho_0/2$ and the autocorrelation function thus becomes

$$\rho_k = \frac{\rho_0}{2}(\gamma_2^k + \gamma_3^k),$$

which is more flexible than the MAP(2) autocorrelation since it depends on two eigenvalues, instead of only γ_2, and yet keeps the explicit dependence on the term $\rho_0 = (1 - 1/SCV)/2$, which immediately relates autocorrelations with SCV.

However, this increased flexibility in the autocorrelation function is paid by a reduction of the degrees of freedom to assign moments, since the specific value given to θ_3 makes it possible to fit only two given moments instead of the three moments of a MAP(2). However, fitting experiments reported in the paper show that this does not significantly affect the queueing prediction accuracy. Fitting expressions as a function of the eigenvalues of the autocorrelation function and the first two moments are as follows:

$$p = (1 - \gamma_2)/3, \qquad q = (2 + \gamma_2 - 3\gamma_3)/6,$$

$$\theta_1 = E[X] + \sqrt{(E[X^2] - 2E[X]^2)\left(\frac{2+\sqrt{3}}{4}\right)}, \qquad \theta_2 = \sqrt{3}E[X] + (1 - \sqrt{3})\theta_1.$$

Positivity conditions on the variables immediately translate into simple feasibility conditions for moments and autocorrelations. In particular, it follows that the proposed class of MAP(3)s has $1 \le SCV < 3$ and that, while γ_2 can assume arbitrary value in its natural range $-1 \le \gamma_2 < 1$, the other eigenvalue γ_3 should be always chosen so that $0 \le q \le 1$. Note that for particular choices of the parameters it can be $\gamma_2 < \gamma_3$. We remark that the condition $1 \le SCV < 3$ is quite limiting for general fitting, but this is not the case if the process is used as a building block in KPC since the composition of several MAP(3)s can result in MAPs with $SCV \gg 3$; furthermore, we observe that the MAPs used for KPC of the BC-Aug89 and Seagate Web traces have low $SCV < 3$ thus suggesting that the MAP(3) model could be equally useful in fitting these traces. Experiments confirming this observation are reported in Section 6.5 proving that the proposed special class of MAP(3) can be a more powerful building block than general MAP(2)s.

References

[1] R.D. Nelson, Probability, Stochastic Processes and Queueing Theory, Springer-Verlag, 1995.

[2] A.T. Andersen, B.F. Nielsen, A Markovian approach for modeling packet traffic with long-range dependence, IEEE J. Sel. Areas Commun. 16 (5) (1998) 719–732.

[3] M. Grossglauser, J.C. Bolot, On the relevance of long-range dependence in network traffic, in: Proc. of SIGCOMM Conf., 1996, pp. 15–24.

[4] A.W. Berger, On the index of dispersion for counts for user demand modeling, in: ITU, Madrid, Spain, AT&T Study Group 2, Question 17/2, June 27–29, 1994.

[5] H.W. Ferng, J.F. Chang, Connection-wise end-to-end performance analysis of queueing networks with MMPP inputs, Perf. Eval. 43 (1) (2001) 39–62.

[6] S.H. Kang, H.K. Yong, D.K. Sung, B.D. Choi, An application of Markovian arrival process (MAP) to modeling superposed ATM cell streams, IEEE Trans. Commun. 50 (4) (2002) 633–642.

[7] A. Heindl, Analytic moment and correlation matching for MAP(2)s, in: Proc. of PMCCS Workshop, 2003, pp. 39–42.

[8] A. Heindl, K. Mitchell, A. van de Liefvoort, Correlation bounds for second-order MAPs with application to queueing network decomposition, Perf. Eval. 63 (6) (2006) 553–577.

[9] L. Bodrog, A. Heindl, G. Horváth, M. Telek, A Markovian canonical form of second-order matrix-exponential processes, European J. Oper. Res. 190 (2008) 457–477.

[10] M.F. Neuts, Structured Stochastic Matrices of M/G/1 Type and Their Applications, Marcel Dekker, New York, 1989.

[11] S. Asmussen, O. Nerman, M. Olsson, Fitting phase-type distributions via the EM algorithm, Scand. J. Statist. 23 (1996) 419–441.

[12] J.E. Diamond, A.S. Alfa, On approximating higher-order MAPs with MAPs of order two, Queueing Syst. 34 (2000) 269–288.

[13] T. Ryden, An EM algorithm for estimation in Markov-modulated Poisson processes, Comput. Statist. Data Anal. 21 (4) (1996) 431–447.

[14] H. Heffes, D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE J. Sel. Areas Commun. 4 (1986) 856–868.

[15] The internet traffic archive, Nov. 2005 release http://ita.ee.lbl.gov.

[16] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Trans. Netw. 2 (1) (1994) 1–15.

[17] A. Horváth, M. Telek, Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples, in: Performance Evaluation of Complex Systems: Techniques and Tools, IFIP Performance 2002, in: LNCS Tutorial Series, vol. 2459, 2002, pp. 405–434.

[18] R.H. Riedi, M.S. Crouse, V.J. Ribeiro, R.G. Baraniuk, A multifractal wavelet model with application to network traffic, IEEE Trans. Inform. Theory 45 (3) (1999) 992–1018.

[19] M.S. Taqqu, V. Teverovsky, W. Willinger, Is network traffic self-similar or multifractal?, Fractals 63 (5) (1997) 833–846.

[20] G. Horváth, P. Buchholz, M. Telek, A MAP fitting approach with independent approximation of the inter-arrival time distribution and the lag correlation, in: Proc. of 2nd Conf. on Quantitative Evaluation of Systems, QEST, 2005, pp. 124–133.

[21] G. Horváth, M. Telek, A canonical representation of order 3 phase type distributions, in: LNCS 4748: Formal Methods and Stochastic Models for Performance Evaluation, 2007, pp. 48–62.

[22] R. Sadre, Decomposition-Based Analysis of Queueing Networks, Ph.D. Thesis, University of Twente, 2007.

[23] D.P. Heyman, D. Lucantoni, Modeling multiple IP traffic streams with rate limits, IEEE/ACM Trans. Netw. 11 (6) (2003) 948–958.

[24] A. Heindl, Traffic-based decomposition of general queueing networks with correlated input processes, Ph.D. Thesis, Shaker Verlag, Aachen, 2001.

[25] Ken'ichi Kawanishi, On the counting process for a class of Markovian arrival processes with an application to a queueing system, Queueing Syst. 49 (2) (2005) 93–122.

[26] S. Li, C. Hwang, Queue response to input correlation functions: Continuous spectral analysis, IEEE/ACM Trans. Netw. 1 (6) (1993) 678–692.

[27] B. Hajek, L. He, On variations of queue response for inputs with the same mean and autocorrelation function, IEEE/ACM Trans. Netw. 6 (5) (1998) 588–598.

[28] A.T. Andersen, B.F. Nielsen, On the use of second-order descriptors to predict queueing behavior of MAPs, Naval Res. Logistics 49 (4) (2002) 391–409.

[29] J. Fan, Q. Yao, Nonlinear Time Series: Nonparametric and Parametric Methods, Springer-Verlag, New York, USA, 2003.

[30] L. Bodrog, A. Heindl, G. Horváth, M. Telek, A. Horváth, Current results and open questions on PH and MAP characterization, in: Dagstuhl Seminar Proceedings 07461, Schloss Dagstuhl, Germany 2008.

[31] L. Breuer, An EM algorithm for batch markovian arrival processes and its comparison to a simpler estimation procedure, Ann. Oper. Res. 112 (1–4) (2002) 123–138.

[32] P. Buchholz, A. Panchenko, A two-step EM algorithm for MAP fitting, in: Proc. of ISCIS, in: LNCS Tutorial Series, vol. 3280, Springer, 2004, pp. 217–227.

[33] A. Riska, V. Diev, E. Smirni, An EM-based technique for approximating long-tailed data sets with PH distributions, Perf. Eval. 55 (1–2) (2004) 147–1674.

[34] A. Riska, E. Riedel, Long-range dependence at the disk drive level, in: QEST 2006 Conference, IEEE Press, 2006, pp. 41–50.

[35] G. Casale, E.Z. Zhang, E. Smirni, KPC-toolbox: Simple yet effective trace fitting using Markovian arrival processes, in: Proc. of the 5th Conf. on Quantitative Evaluation of Systems (QEST), 2008, pp. 83–92.

[36] M. Telek, G. Horváth, A minimal representation of Markov arrival processes and a moments matching method, Perf. Eval. 64 (9–12) (2007) 1153–1168.
[37] F.E. Hohn, Elementary Linear Algebra, Dover, 1973.
[38] A. Saar, Numerical Methods for Large Eigenvalue Problems, Manchester University Press, 1992.
[39] G. Casale, E.Z. Zhang, E. Smirni, Interarrival times characterization and fitting for Markovian traffic analysis, Technical Report WM-CS-2008-02, available at http://www.wm.edu/as/computerscience/documents/cstechreports/WM-CS-2008-02.pdf, College of William and Mary, 2008.
[40] J.W. Brewer, Kronecker products and matrix calculus in system theory, IEEE Trans. Circuits Sys. 25 (9) (1978).
[41] H. Che, S. Li, Fast algorithms for measurement-based traffic modeling, in: Proc. of INFOCOM Conf., 1997, pp. 177–186.
[42] K. Sriram, W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, IEEE J. Sel. Areas Commun. 4 (6) (1986) 833–846.
[43] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stoch. Models 7 (1991) 1–46.

Giuliano Casale received the M.E.E. and Ph.D. degrees in computer engineering from the Politecnico di Milano, Milan, Italy, in 2002 and 2006, respectively. He is currently a full-time researcher at SAP Research, CEC Belfast. From January 2007 he was postdoctoral research associate at the College of William and Mary, Williamsburg, Virginia, where he studied the performance impact of burstiness in systems. In Fall 2004 he was a visiting scholar at UCLA studying bounds for queueing networks. His research interests include performance evaluation, modeling, capacity planning, and simulation. He is a member of the ACM, IEEE, and IEEE Computer Society.

Eddy Z. Zhang is currently a Ph.D. student in the Department of Computer Science at the College of William and Mary. She received her B.S. in Electrical Engineering from Shanghai Jia Tong University and her M.S. in Computer Science from the College of William and Mary. Her main research interests include program locality and memory management, adaptive compilation, runtime scheduling policies, workload characterization, program parallelization and parallel computing.

Evgenia Smirni is a Professor at the College of William and Mary, Department of Computer Science, Williamsburg, Virginia 23187-8795 (esmirni@cs.wm.edu). She received her Diploma in Computer Engineering and Informatics from the University of Patras, Greece, in 1987, and her M.S. and Ph.D. in Computer Science from Vanderbilt University in 1993 and 1995, respectively. Her research interests include analytic modeling, stochastic models, Markov chains, queueing networks, resource allocation policies, Internet systems, storage systems, workload characterization, and modeling of distributed systems and applications. She has served as program co-chair of QEST'05, ACM SIGMETRICS/Performance'06, of ACM HOTMETRICS 2010 and as general co-chair of QEST'10. She is a member of ACM, IEEE, and the Technical Chamber of Greece.