Compact Markov-Modulated Models for Multiclass Trace Fitting

Giuliano Casale[a], Andrea Sansottera[b], Paolo Cremonesi[b]

[a] *Imperial College London, Department of Computing, 180 Queen's Gate, SW7 2AZ, UK*
[b] *Politecnico di Milano, DEIB, Via Ponzio 34/5, 20133, Milan, Italy*

Abstract

Markov-modulated Poisson processes (MMPPs) are a popular class of stochastic models for fitting empirical traces. In this paper, we develop the first study of the counting process of the marked MMPP (M3PP), the multiclass generalization of the MMPP. We initially explain how to fit two-state M3PPs to empirical datasets. Stemming from this result, we propose a novel form of composition, called *interleaving* which enables the superposition of several two-state M3PPs in a way that preserves their statistical characteristics without suffering state space explosion, the major drawback and limiting factor of MMPP superposition. Experimental results indicate that the proposed M3PP fitting methods provide accurate results using a much more compact representation than superposed MMPPs.

Keywords: Counting process, marked Markov-modulated Poisson process, trace, fitting

1. Introduction

The Markov-modulated Poisson process (MMPP) is a general fitting tool for traces with correlated arrivals [11] which finds application in modelling network traffic [26], disk I/O patterns [31], grid and cloud workloads [22], and self-adaptive software systems [28]. Its main feature is the composability with other Markov models, such as queueing systems [17, 18, 30] or stochastic Petri nets [28], which enables integrating non-renewal processes in performance models, without violating the Markov property.

The MMPP is a special case of the Markovian Arrival Process (MAP) [23] where the departure process is a modulation of Poisson processes and the currently active process is selected according to the active state of a continuous-time Markov chain. In this paper, we study a generalization of the MMPP, the *marked MMPP* (M3PP), and develop a scalable methodology for its fitting to empirical datasets. The M3PP can be considered as a specialization of the MMAP, the marked extension of the MAP [13]. A marked point process associates to each arrival a class label [13, 12], thus it is useful to model traces with events

Email addresses: `g.casale@imperial.ac.uk` (Giuliano Casale), `sansottera@elet.polimi.it` (Andrea Sansottera), `paolo.cremonesi@polimi.it` (Paolo Cremonesi)

of multiple classes (e.g., read and write requests in disk drives). Marked processes are also important in the analysis of priority queues and multiclass models in general [17, 8, 18, 30]. However, few techniques exist for their fitting and they all focus on matching moments of the inter-arrival time process for marked MAPs (MMAPs) [17, 8]. Therefore, today no techniques exist for fitting *marked* Markov-modulated processes to count data. Still, to limit monitoring overheads, only count data is extracted from many computer systems (e.g., network links, web servers), which calls for methodologies to fit marked datasets.

In this paper, we fill this gap by developing approximate fitting algorithms for the counting process of the M3PP. In particular, we first explain how to *approximately* fit a two-state MMPP counting process and then extend the concept to two-state M3PPs. The proposed approach is applicable to traces with aggregated or coarse-sampled measurements, which cannot be analysed using approaches based on inter-arrival times, which require moments from a trace recording *all* the arrivals. The main drawback is that counting processes are mathematically less tractable than inter-arrival processes, therefore one normally focuses on low-order moments of counts.

Two-state models are often insufficient to fit complex traces, therefore we then study the approximate fitting of large M3PPs. In the single class setting, a known limitation of MMPPs is the inability to fit simultaneously many statistical descriptors due to the non-linearity of their underlying equations [20, 3, 16]. This has led to the definition of several approaches to fit complex traces by composing, using Kronecker operators, multiple small-sized MMPPs or MAPs based on their counting process, as done for example in [1, 15], or their inter-arrival time process, as for example in [10]. These methods employ composition operators for moment fitting, offering a different trade-off between computational cost and fitting accuracy compared to fitting methods based on the EM algorithm [5, 19, 21]. In particular, the *superposition* operator allows one to describe a trace by the statistical multiplexing of several MMPPs, at the expense of an exponential growth of the number of states in the process representation [29]. This state space explosion is an obstacle for the application of MMPPs and MAPs to real systems, for example it considerably slows down, or even renders infeasible, the numerical evaluation of queueing models by matrix geometric methods [27, 2].

In this paper, we tackle the state space explosion problem of superposition by showing that M3PPs admit a more sophisticated form of composition, which we call *interleaving*, that enables several MMPPs to share the same state space without mutually affecting the marginal distributions of their counting processes. However, this introduces spurious covariances between class arrivals that may be seen as the cost of the state-space reduction. The interleaving method defines an original approach to build large Markov-modulated processes, in which a set of J two-state M3PPs is merged into a single M3PP process with just

$J + 1$ states, without affecting the marginal counting processes of the initial M3PPs. Thus, interleaving is a composition methodology that can generate compact Markov-modulated processes for fitting complex traces. We identify general conditions for interleaving to result in a valid MMAP and observe that these condition can be satisfied by M3PPs, but not by general MMAPs. The possibility to interleave processes makes a case for using M3PPs, instead of general MMAPs, for fitting count data. We then define a method to automatically identify the M3PPs to be interleaved and a fast mixed-integer linear (MILP) program that can help in automatically identifying the order of interleaving of the M3PPs.

We conclude the paper reporting fitting experiments for a set of artificial and real-world traces. We show that the proposed M3PP fitting algorithms are widely applicable and run efficiently even in the case of approximate fitting. We also find that interleaving M3PPs outperforms superposition in terms of both accuracy and compactness of the representation.

Summarizing, our main contributions are as follows:

- Section 3 defines fitting algorithms for the counting process of two-state M3PPs with arbitrary number of classes. Our formulas are in closed-form for exact fitting and use quadratic programming for approximate fitting. As a by-product, our approach also introduces the first infeasibility adjustment methodology for approximate fitting of unmarked MMPPs.
- Section 4 introduces the new notion of interleaved process to compose multiple M3PPs, while preserving their statistical properties, as we rigorously establish in Theorem 1.
- Section 5 develops a methodology for fitting empirical traces using the interleaving operator.

In addition to the above, a description of necessary background is given in Section 2 and Section 7 draws conclusions.

2. Background

2.1. Model and Notation

An m-state MMPP is defined by a continuous-time Markov chain (CTMC) with infinitesimal generator $\boldsymbol{Q}$, having m states, and rate vector $(\lambda(1), \ldots, \lambda(m))$. When the CTMC is in state j, the MMPP generates arrivals according to a Poisson process with rate $\lambda(j)$. The net effect of the modulating action of the underlying CTMC is to enable the modelling of non-Poisson, possibly autocorrelated, arrival processes. We assume the underlying CTMC to be initialized according to its steady-state distribution so that the process is time-stationary. For ease of comparison with the literature, we use throughout the paper the MAP $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ notation, where for a MMPP $\boldsymbol{D}_1 = \text{diag}(\lambda(1), \ldots, \lambda(m))$ and $\boldsymbol{D}_0 = \boldsymbol{Q} - \boldsymbol{D}_1$.

We define a M3PP[K] as a MMPP in which arrivals are marked with one out of K available classes. This may be seen as a marking of the Poisson processes of the MMPP where one decomposes each rate $\lambda(j)$, $1 \leq j \leq m$, into arrival rates $q_{j,k}\lambda(j)$, $1 \leq k \leq K$, subject to $\sum_k q_{j,k} = 1$, $q_{j,k} \geq 0$. When an arrival is generated by a Poisson process with rate $q_{j,k}\lambda(j)$ it is said to be of class k. Equivalently, in matrix notation, augmenting a MMPP $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ with marks defines a set of matrices $\boldsymbol{D}_{1,k} = \mathrm{diag}(q_k\lambda(1), \ldots, q_{m,k}\lambda(m))$, such that $\boldsymbol{D}_1 = \sum_{k=1}^{K} \boldsymbol{D}_{1,k}$. The tuple $(\boldsymbol{D}_0, \boldsymbol{D}_{1,1}, \ldots, \boldsymbol{D}_{1,K})$ is called the *representation* of the M3PP[K]. Also, $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ is the embedded MMPP, which we refer to as the *unmarked* process.

In the rest of the paper we will often deal with processes having $m = 2$ states. For readability, we will use λ and λ' in place of $\lambda(1)$ and $\lambda(2)$.

2.2. Problem formulation

Let $n_k(t) \geq 0$ be the number of arrivals of class k in a M3PP[K] at time t after initialization, subject to $n_k(0) = 0$ for all classes. The *counting process* of the M3PP[K] is the CTMC with state $\boldsymbol{n}(t) = (n_1(t), n_2(t), \ldots, n_K(t))$. We shall denote by $n(t) = \sum_{k=1}^{K} n_k(t)$ the aggregate arrival count. Given an initial state probability vector, the evolution over time of this process is characterized by a matrix $\boldsymbol{P}(\boldsymbol{n}, t)$, with element $p_{i,j}(\boldsymbol{n}, t)$ in row i and column j representing the probability that a M3PP initialized in state i is in state j at time t with an arrival count $\boldsymbol{n}(t)$. Thus, the counting process evolves according to the Kolmogorov forward equations

$$\frac{\partial \boldsymbol{P}(\boldsymbol{n}, t)}{\partial t} = \boldsymbol{P}(\boldsymbol{n}, t)\boldsymbol{D}_0 + \sum_{k=1}^{K} \boldsymbol{P}(\boldsymbol{n} - \boldsymbol{e}_k, t)\boldsymbol{D}_{1,k}, \tag{1}$$

where $\boldsymbol{e}_k$ is a column vector of zeros except for a one in position k. From this equation, it is simple to derive factorial moments functions, from which moments of the counting process can be obtained in closed form [13]. For example, this yields the mean arrival count of class k at time t as

$$\mu_k(t) = \mathrm{E}[n_k(t)] = \mu_k t, \qquad \mu_k = \boldsymbol{\pi}\boldsymbol{D}_{1,k}\boldsymbol{1}, \tag{2}$$

where $\boldsymbol{\pi}$ is the stationary distribution of the embedded CTMC, $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$, $\boldsymbol{\pi}\boldsymbol{1} = 1$, μ_k is the *mean arrival rate* of class k in the time-stationary process, $\boldsymbol{0}$ and $\boldsymbol{1}$ are respectively column vectors of zeros and ones. The *variance* of class k in counts (also called variance-time curve) is [13]

$$\mathrm{Var}[n_k(t)] = \left(\mu_k - 2\mu_k^2 + 2\boldsymbol{c}_k\boldsymbol{D}_{1,k}\boldsymbol{1}\right)t - 2\boldsymbol{c}_k\left(\boldsymbol{I} - e^{\boldsymbol{Q}t}\right)\boldsymbol{d}_k \tag{3}$$

where $\boldsymbol{c}_k = \boldsymbol{\pi}\boldsymbol{D}_{1,k}\left(\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{Q}\right)^{-1}$ and $\boldsymbol{d}_k = \left(\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{Q}\right)^{-1}\boldsymbol{D}_{1,k}\boldsymbol{1}$. The above formulas and notations are also valid for the unmarked process ($K = 1$), but in that case we omit the dependence on class k. The *covariance* of counts for classes k and h is given by

$$\mathrm{Cov}[n_k(t), n_h(t)] = \frac{1}{2}\left(\mathrm{Var}[n_k(t) + n_h(t)] - \mathrm{Var}[n_k(t)] - \mathrm{Var}[n_h(t)]\right), \tag{4}$$

where $\mathrm{Var}[n_k(t) + n_h(t)]$ can be obtained from (3) by replacing μ_k with $\mu_k + \mu_h$ and $\boldsymbol{D}_{1,k}$ with $\boldsymbol{D}_{1,k} + \boldsymbol{D}_{1,h}$.

The M3PP[K] fitting problem is to find a representation $(\boldsymbol{D}_0, \boldsymbol{D}_{1,1}, \ldots, \boldsymbol{D}_{1,K})$ such that (2)-(4) match, to the best possible extent, the corresponding empirical values at given timescales t. Such a representation needs to be valid, meaning that all rates are non-negative real numbers and the generator $\boldsymbol{Q}$ of the embedded CTMC is both valid and irreducible.

2.3. Two-State MMPP Fitting

In order to fit a two-state M3PP[K], we propose to separately fit the MMPP for the unmarked process and the markings needed to specify the M3PP. Since MMPP fitting is well-known, we here just summarize the MMPP fitting procedure proposed in [14]. The fitting algorithm receives in input the following descriptors of the counting process:

$$\mu = \frac{\mathrm{E}[n(t)]}{t}, \quad I(t) = \frac{\mathrm{Var}[n(t)]}{\mathrm{E}[n(t)]}, \quad I = \lim_{t\to\infty} I(t), \quad \mu^{(3)}(t) = \mathrm{E}\left[(n(t) - \mu t)^3\right],$$

where μ is the rate, $I(t)$ is the *index of dispersion for counts* [29] at timescale t, I is the asymptotic index of dispersion value for $t \to \infty$, and $\mu^{(3)}(t)$ is the third central moment of counts at timescale t. The objective is to fit the rates used in the MMPP representation

$$\boldsymbol{D}_0 = \begin{bmatrix} -(\lambda + r) & r \\ r' & -(\lambda' + r') \end{bmatrix}, \qquad \boldsymbol{D}_1 = \mathrm{diag}(\lambda, \lambda').$$

subject to all the parameters being non-negative and $\lambda + \lambda' > 0$. Notice in particular that we exclude the trivial cases $\lambda = \lambda'$ and $r = r' = 0$, where the MMPP degenerates into a Poisson process, which does not require a two-state model.

Fitting the above descriptors to a two-state MMPP requires to solve for the unknowns r, r', λ, and λ' using a nonlinear system composed by the following four equations [14]

$$\mu = \frac{\lambda r' + \lambda' r}{x}, \quad I = 1 + \frac{2(\lambda - \lambda')^2 r r'}{x^2(\lambda r' + \lambda' r)}, \quad \frac{I - I(t)}{I - 1} = \frac{1 - e^{-xt}}{xt}, \quad h = (\lambda - \lambda')x', \tag{5}$$

where $x = r + r'$, $x' = r - r'$, t is an arbitrary finite timescale at which we want to fit the index of dispersion, and $h = (g^{(3)}(1, t) - k_1 - k_2 + k_3\mu - k_4\mu x)(k_4 + (k_3/x) - k_5)^{-1}$. The parameters in the last equation are $k_1 = \mu^3 t^3$, $k_2 = 3\mu^2(I - 1)t^2$, $k_3 = 3\mu(I - 1)/xt$, $k_4 =$

$3\mu(I-1)te^{-xt}/x^2$, $k_5 = 6\mu(I-1)(1-e^{-xt})/x^3$, and $g^{(3)}(1,t)$ is the third factorial moment of counts at timescale t. We point to [14, Eq. (14d)] for explicit expressions of $g^{(3)}(1,t)$.

Heffes and Lucantoni [14] propose the following fitting procedure. First, compute $x = r + r'$, solving at an arbitrary timescale $t = t_1$ the fixed point equation

$$x = \frac{1}{t}\left(\frac{I-1}{I-I(t)}\right)\left(1-e^{-xt}\right) \tag{6}$$

obtained by rewriting the third equation appearing in (5). Then, we compute h at a second arbitrary timescale $t = t_2$. If $h \neq 0$, the fitting formulas are

$$r = \frac{x}{2}\left(1 + \frac{1}{\sqrt{4y+1}}\right), \quad r' = x - r, \quad \lambda' = \mu - \frac{hr'}{x'x}, \quad \lambda = \mu + \frac{hr}{x'x}.$$

where $y = (I-1)\mu x^3(2h^2)^{-1}$. If $h = 0$, the explicit fitting formulas become

$$r = r' = \frac{x}{2}, \quad \lambda' = \mu - \frac{1}{2}\sqrt{2(I-1)\mu x}, \quad \lambda = \mu + \frac{1}{2}\sqrt{2(I-1)\mu x}.$$

The main drawback of this procedure is that the formulas can return negative rates for some combinations of the input parameters. In such cases, approximate fitting techniques are required, which to the best of our knowledge are however not available in the literature on the counting process of a MMPP.

3. Fitting the Counting Process of a Two-State M3PP[$\boldsymbol{K}$]

In this section, we develop a procedure to fit two-state M3PP[K]s with representation

$$\boldsymbol{D}_0 = \begin{bmatrix} -(\lambda + r) & r \\ r' & -(\lambda' + r') \end{bmatrix}, \quad \boldsymbol{D}_{1,k} = \mathrm{diag}(q_k\lambda, q'_k\lambda'),$$

where $\lambda + \lambda' > 0$, $\sum_{k=1}^{K} q_k = \sum_{k=1}^{K} q'_k = 1$, $q_k \geq 0$, $q'_k \geq 0$, for all $1 \leq k \leq K$. As mentioned before, the availability of exact methods to fit unmarked MMPPs justifies an approach where we fit separately the embedded MMPP first, followed by the marking process that defines the M3PP. In practice, this means that we first fit $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ using a MMPP fitting algorithm applied to the unmarked trace. Then we determine the individual $\boldsymbol{D}_{1,k}$ matrices using the marked trace and subject to the condition $\boldsymbol{D}_1 = \sum_{k=1}^{K} \boldsymbol{D}_{1,k}$, which ensures that the embedded MMPP does not change. In Section 3.1, we discuss how to perform approximate fitting of the unmarked process using an extension of the algorithm by Heffes and Lucantoni [14]. In Section 3.2, we discuss the fitting of the marked process.

3.1. Step 1: Approximate MMPP Fitting

The fitting algorithm of Heffes and Lucantoni described in Section 2.3 cannot be applied to cases where the input set of descriptors μ, $I(t)$, I and $\mu^{(3)}(t)$ is infeasible for the considered MMPP. We have found this to happen frequently in real traces, and despite one may repeatedly attempt to fit different timescales t_1 and t_2 until finding a feasible set of descriptors, we believe that better solutions are needed and possible. We have not found previous works addressing this problem, at least for fitting counting processes. Our approximation consists in sacrificing the degree of freedom of the third moment of counts $\mu^{(3)}(t)$ to restore feasibility of all second-order descriptors. We here focus on cases where the M3PP process is not Poisson, thus we assume $\lambda \neq \lambda'$, $r > 0$, $r' > 0$.

We being by characterizing the feasibility region of the index of dispersion for a two-state MMPP.

Proposition 1. *In a two-state MMPP with $\lambda \neq \lambda'$, the index of dispersion satisfies $I > I(t) > 1$ for any timescale t.*

Proof. From the nonlinear system of equations (5), it readily follows that $I > 1$ since $\lambda \neq \lambda'$. Moreover, since x is also strictly positive, it follows from (6) that $I > I(t)$. Notice that (6) admits a solution $x > 0$ if and only if the derivative of the right-hand side evaluated at 0 is greater than 1. Hence, $(I-1)/(I-I(t)) > 1$ and therefore $I(t) > 1$. □

The previous proposition states the well-known fact that MMPPs can only fit traces that are more variable than a Poisson process. We now show that asking for $r > 0$ and $r' > 0$, to avoid the MMPP being Poisson, is implied by the previous statement and thus always verified.

Proposition 2. *For $I > 1$, the conditions $r > 0$ and $r' > 0$ are always satisfied.*

Proof. For the case $h \neq 0$, observe that $y > 0$ if $I > 1$. Therefore $1/\sqrt{4y+1} < 1$. Since x is positive, it also follows that $0 < r < x$ and $0 < r' < x$.

For the case $h = 0$, the result follows immediately given that $x > 0$ since $r = r' = 0$ would imply a Poisson process which would have $I = 1$ against the assumption. □

Therefore, provided that $I > I(t) > 1$, we simply need to ensure that arrival rates are non-negative and $\lambda + \lambda' > 0$. Using (5), we can reformulate this requirement as:

$$-\frac{\mu x}{r} \leq \frac{h}{x'} \leq \frac{\mu x}{r'} \tag{7}$$

which implicitly characterizes the feasible region for the rate μ and, via the h term, for the third moment $g^{(3)}(1,t)$. If one of these conditions is not met, we propose to relax the fitting procedure by ignoring the matching of third moment $g^{(3)}(1,t)$ as follows. Consider first the following lower bound on r.

Proposition 3. *Without loss of generality, assume $\lambda > \lambda'$. Then $r \geq u$, where*

$$u = \frac{(I-1)x^2}{2\mu + (I-1)x}. \tag{8}$$

where $x = r + r'$ is the solution of (6).

Proof. We substitute the first equation in (5) into the second one and find after simple algebra

$$\lambda = \lambda' + \sqrt{\frac{(I-1)\mu x^3}{2rr'}} \tag{9}$$

Obtaining λ from the first equation in system (5) and equating to (9), we get

$$\lambda' = \mu - \frac{r'}{x}\sqrt{\frac{(I-1)\mu x^3}{2rr'}} \geq 0 \tag{10}$$

The result follows using $r' = x - r$ and solving for r, which yields the lower bound u. $\square$

Any value $u \leq r < x$ leads to a feasible solution, thus we can for example choose the middle point of this interval $r = u + (u - x)/2$. Our suggestion of the middle point is convenient to keep the formulas in closed-form, however other choices are possible. After choosing r, the other parameters are easily obtained by setting $r' = x - r$ and using the equations developed in the proof of Proposition 3 to obtain λ' and λ.

Summarizing, provided that $I > I(t_1) > 1$, the approximate procedure we have defined fits μ, $I(t_1)$ and I exactly at the expense of sacrificing an infeasible third moment $\mu^{(3)}(t_2)$, where t_1 and t_2 are arbitrary timescales.

3.2. Step 2: Fitting the Marking Process

In the second step of the fitting algorithm we determine the $\boldsymbol{D}_{1,k}$ matrices. After Step 1, all the statistical properties of the unmarked process are constrained by the $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ matrices of the MMPP, thus the focus of this step is to fit the class-specific properties and class covariances. We consider both exact and approximate fitting procedures.

There are several ways of choosing what descriptors to fit with a M3PP and each choice leads to equations that may differ in tractability compared to other choices. We have experimented with several possibilities, and found that fitting a set of central moments, such as mean class rates μ_k, class variances, or covariances, typically leads to a difficult non-convex formulation. However, we have found that it is instead efficient to fit a two-state M3PP[K]

using the mean class rates and per-class contributions G_k to the asymptotic index of dispersion I. Other efficient approaches are possible, such as fitting mean class and relative covariances or fitting mean class rates and differences between the class variances. However, we here focus on the first method, which we believe provides the best combination of efficiency (quadratic programming) and ease of interpretation.

3.2.1. Fitting mean class rates and per-class contributions to the index of dispersion

We begin by considering the more general problem of fitting the mean class rates μ_k and $g_k(t) = \mathrm{Var}\left[n_k(t)\right] + \mathrm{Cov}[n_k(t), \sum_{i \neq k} n_i(t)]$, i.e., the sum of class variance and per-class contributions to the unmarked variance. The $g_k(t)$ terms have the simple interpretation of modelling the contribution of class k to the variance-time curve of the unmarked process, since $\sigma(t) = \mathrm{Var}\left[\sum_{k=1}^{K} n_k(t)\right] = \sum_{k=1}^{K} g_k(t)$. We will then specialize the method to the index of dispersion, using the fact that $I(t) = \sigma(t)/E[n(t)] = \sum_{k=1}^{K} g_k(t)/E[n(t)]$.

Our goal is to compute the $\boldsymbol{D}_{1,k}$ matrices from the $g_k(t)$ values, given $\boldsymbol{D}_0$ and $\boldsymbol{D}_1 = \sum_{k=1}^{K} \boldsymbol{D}_{1,k}$. Recall that $\boldsymbol{D}_{1,k} = \mathrm{diag}(q_k\lambda, q'_k\lambda')$, the problem under consideration is to determine the values of the probabilities q_k and q'_k that uniquely define the $\boldsymbol{D}_{1,k}$ matrices, given that λ and λ' are known from the $\boldsymbol{D}_1$ matrix fitted in Step 1.

Exact matching. We now give formulas to fit the probabilities q_k and q'_k from μ_k and $g_k(t)$. Note that the arbitrary timescale t used in the statement does not necessarily need to be equal to any of the arbitrary timescales used for fitting the embedded MMPP.

Proposition 4. *Given ($\boldsymbol{D}_0$, $\boldsymbol{D}_1$) and, for each class k, μ_k and $g_k(t)$ at an arbitrary timescale t, the parameters of the M3PP can be computed as follows:*

$$q_k = f_{1,1}\mu_k + f_{1,2}g_k(t) \tag{11}$$

$$q'_k = f_{2,1}\mu_k + f_{2,2}g_k(t) \tag{12}$$

where $f_{1,1} = -F_2x/F$, $f_{1,2} = \lambda' r/F$, $f_{2,1} = F_1 x/F$, $f_{2,2} = -\lambda r'/F$, in which $F = (F_1\lambda' r - F_2\lambda r')$,

$$F_1 = \lambda r' x^{-4}\Big(2(\lambda - \lambda')re^{-xt} + tx\Big(x^2 - 2(\lambda' - \lambda)r\Big) + 2(\lambda' - \lambda)r\Big),$$

$$F_2 = \lambda' r x^{-4}\Big(2(\lambda' - \lambda)r'e^{-xt} + tx\Big(x^2 - 2(\lambda - \lambda')r'\Big) + 2(\lambda - \lambda')r'\Big),$$

and $x = r + r'$ is the solution of (6).

Proof. The expressions of μ_k and $g_k(t)$ for a two-state M3PP[K] process are found from the definitions to be

$$\mu_k = \frac{\lambda q_k r' + \lambda' q_k' r}{x} \tag{13}$$

$$g_k(t) = F_1 q_k + F_2 q_k' \tag{14}$$

where F_1 and F_2 are constant coefficients given $\boldsymbol{D}_0$, $\boldsymbol{D}_1$ and the reference timescale t. Solving (13) for q_k we obtain

$$q_k = \frac{\mu_k x - \lambda' q_k' r}{\lambda r'} \tag{15}$$

Substituting (15) into (14), we obtain the fitting formulas (16) and (17). □

We are now ready to determine the contributions to the index of dispersion. Note that Proposition 4 can also be used to fit the contribution of class k to $I(t)$ of the embedded MMPP, i.e., $G_k(t) = g_k(t)/E[n(t)]$. This is because we can rewrite the fitting formulas as

$$q_k = c_{1,1}\mu_k + c_{1,2}G_k(t) \tag{16}$$

$$q_k' = c_{2,1}\mu_k + c_{2,2}G_k(t) \tag{17}$$

where $c_{1,1} = f_{1,1}$, $c_{1,2} = f_{1,2}\mu t$, $c_{2,1} = f_{2,1}$, $c_{1,1} = f_{2,2}\mu t$. The asymptotic expressions of the coefficients as $t \to \infty$ are readily obtained as

$$c_{1,1} = -\frac{2(\lambda' - \lambda)r' + x^2}{2\lambda r'(\lambda - \lambda')},\ c_{1,2} = \frac{\mu x^2}{2\lambda r'(\lambda - \lambda')},\ c_{2,1} = \frac{2(\lambda - \lambda')r + x^2}{2\lambda' r(\lambda - \lambda')},\ c_{2,2} = -\frac{\mu x^2}{2\lambda' r(\lambda - \lambda')}$$

Combining these expressions with the requirements $q_k \geq 0$ and $q_k' \geq 0$, we find after simple passage this tight lower bound for the feasibility of the per-class contributions to the asymptotic index of dispersion:

$$G_k \geq \frac{\mu_k}{\mu}\left(1 + 2\frac{\max((\lambda' - \lambda)r', (\lambda - \lambda')r)}{x^2}\right), \tag{18}$$

For two-state M3PPs, the minimum value of the right-hand side is achieved for Poisson processes, where $G_k = \mu_k/\mu$ and $\sum_k G_k = I = 1$.

Approximate matching. Some values of the descriptors might lead to unfeasible values of the parameters (e.g., negative probabilities). In this case, we choose to fit the per-class arrival rates μ_k exactly and find the feasible values $\tilde{G}_k$ that minimize the following L^2-norm

$$\sum_{k=1}^{K}\left(\frac{\tilde{G}_k - G_k}{G_k}\right)^2 = \frac{1}{2}\boldsymbol{x}^T H\boldsymbol{x} + f^T\boldsymbol{x} + K \tag{19}$$

with $\boldsymbol{x} = (\tilde{G}_1, \ldots, \tilde{G}_K)$, $H = \mathrm{diag}\,(2/G_1^2, \ldots, 2/G_K^2)$, $f^T = (-2/G_1, \ldots, -2/G_K)$, K being the number of classes of the M3PP, and subject to the constraints

$$-c_{i,2}\tilde{G}_k \leq c_{i,1}\mu_k \qquad \forall i = 1,2;\ k = 1,\ldots,K \tag{20}$$

$$c_{i,2}\sum_{k=1}^{K}\tilde{G}_k = 1 - c_{i,1}\mu \qquad \forall i = 1,2. \tag{21}$$

These constraints are derived using equations (16) and (17) for q_k and q'_k. In particular, (20) stem directly from imposing $q_k \geq 0$ and $q'_k \geq 0$, whereas (21) follow from the conditions $\sum_{k=1}^{K} q_k = 1$ and $\sum_{k=1}^{K} q'_k = 1$.

The above optimization program can be used also for fitting mean class rates μ_k and the contributions $g_k(t)$ to the index of dispersion $I(t)$. In order to do so, it is sufficient to replace the coefficients $c_{i,j}$ with the coefficients $f_{i,j}$, G_k with $g_k(t)$, and $\tilde{G}_k$ with $\tilde{g}_k(t)$.

In either case, the program has a convex quadratic objective function, thus its minimizer can be obtained efficiently using standard quadratic programming solvers. Once the problem is solved and the feasible values $\widetilde{G}_k$ or $\tilde{g}_k(t)$ are obtained, the parameters q_k and q'_k are readily fitted using (16) and (17).

4. Compositional Fitting of M3PP[$\boldsymbol{K}$]s

In the previous section, we have defined a general purpose fitting method for two-state M3PPs. We now consider the problem of exploiting additional degrees of freedom of a M3PP[K] to increase the flexibility of the fitting. For this case, exact fitting procedures capable of exploiting all available degrees of freedom do not exist even for unmarked processes. Therefore we focus on compositional fitting, where one builds a large process by composition of smaller processes which are simpler to fit to count data than a m-state process. The drawback of compositional approaches of this kind is to use only some of the degrees of freedom of a general process in return for ease of fitting.

We first review and generalize superposition methods for unmarked MMPPs. Afterwards, we introduce a novel form of composition, called *interleaving*, which offers a different trade-off between accuracy and compactness of the representation. Note that we do not include in the analysis methods that are specific to inter-arrival processes, such as the product composition [10].

4.1. Superposition

Unmarked case. Consider a set of J MMPP $(\boldsymbol{D}_0^j, \boldsymbol{D}_1^j)$, $1 \leq j \leq J$, their superposition is the MMPP process $(\boldsymbol{D}_0^+, \boldsymbol{D}_1^+)$ where $\boldsymbol{D}_0^+ = \bigoplus_{j=1}^{J} \boldsymbol{D}_0^j$, $\boldsymbol{D}_1^+ = \bigoplus_{j=1}^{J} \boldsymbol{D}_1^j$, in which $\oplus$ denotes the Kronecker sum operator [6]. Superposition naturally arises in networking to describe the traffic process obtained by merging separate traffic flows, each described by a MMPP. That is, the superposed process defines the multiplexing of J channels, each with inter-arrival times described by the j-th MMPP. The superposed process has mean arrival rate and variance-time curve equal to the summation of the individual MMPPs' mean arrival rates and variance-time curves. The index of dispersion is a weighted sum of the IDCs of the individual MMPPs, i.e., $I(t) = \sum_j (\mu_j/\mu) I_j(t)$, where $\mu = \sum_j \mu_j$ is the mean arrival rate of the superposition. Finally, since the Kronecker sum allows us to expand the state space in such a way to keep track of the individual state of each MMPP, if MMPP j has m_j states, the superposed process has $\prod_{j=1}^{J} m_j$ states, thus its size grows exponentially with J.

Marked case. The introduction of class markings does not change the mathematical nature of the superposition of M3PPs compared to MMPP superposition. Let $\mathcal{K} = \{1, \ldots, K\}$ be a set of classes. We consider J M3PPs and assume that M3PP j generates arrival of classes $\mathcal{K}_j \subset \mathcal{K}$, with $\mathcal{K}_1 \cup \mathcal{K}_2 \cup \cdots \cup \mathcal{K}_J = \mathcal{K}$. In this case, only a few M3PPs contribute to arrivals of class k, thus $\boldsymbol{D}_0^{+} = \bigoplus_{j=1}^{J} \boldsymbol{D}_0^{j}$, $\boldsymbol{D}_{1,k}^{+} = \bigoplus_{j=1}^{J} \boldsymbol{D}_{1,k}^{j}$, $\forall k \in \mathcal{K}$, where we define $\boldsymbol{D}_{1,k}^{j} = \mathbf{0}$ if $k \notin \mathcal{K}_j$, being $\mathbf{0}$ a matrix of all zeros of order m_j.

The statistical properties of the embedded unmarked process are obtained similarly to the case of an unmarked superposition. Lastly, note that if each M3PP has two states, the resulting process is a M3PP with 2^J states and K classes. Thus, the main drawback of the superposition method remains the state-space explosion, which limits the superposition to a small number of processes.

4.2. Interleaving

We now propose a new form of composition for Markov-modulated processes that tackles the state-space explosion problem of superposition. Informally, our idea is to define an operator by which several M3PPs can by defined upon the same state space, without interfering with their respective marginal counting processes. This allows us to preserve in the interleaving the counting process properties fitted in isolation for each composed M3PP. We characterize the main feature of this new composition operator later in Theorem 1. Note that the results in this section are also applicable to unmarked MMPPs, and thus represent advances also for unmarked processes.

Consider a set of J two-state M3PPs where the ith process has K_i classes. Assume classes to be labelled such that each class appears in one and only one M3PP and note that this does not lead generality, since classes belonging to different M3PPs can still be stochastically equivalent. The M3PPs have representation

$$\boldsymbol{D}_0^i = \begin{bmatrix} -r_i - \lambda_i & r_i \\ r'_i & -r'_i - \lambda'_i \end{bmatrix}, \quad \boldsymbol{D}_{1,k}^i = \mathrm{diag}(q_{i,k}\lambda_i, q'_{i,k}\lambda'_i), \quad k = 1\ldots, K_i,$$

where we define the probabilities $\sum_{k=1}^{K_i} q_{i,k} = 1$, $q_{i,k} \geq 0$, and $\sum_{k=1}^{K_i} q'_{i,k} = 1$, $q'_{i,k} \geq 0$, and the rates $r_i = \sum_{j=i}^{J} \alpha_j$ and $r'_i = \sum_{j=1}^{i} \beta_j$, for given constants $\alpha_j \geq 0$ and $\beta_j \geq 0$. Equivalently, given the values of the r_i and r'_i constants, we can define the rate differences $\alpha_i = r_i - r_{i+1}$ and $\beta_i = r'_i - r'_{i-1}$, with boundary values $\beta_1 = r'_1$ and $\alpha_J = r_J$.

We now define the interleaving operator for M3PPs. Given the set of K two-state M3PPs $(\boldsymbol{D}_0, \boldsymbol{D}_{1,1}^i, \ldots, \boldsymbol{D}_{1,K_i}^i)$, the interleaved process $(\boldsymbol{D}_0^*, \boldsymbol{D}_{1,1}^*, \ldots, \boldsymbol{D}_{1,K}^*)$ is the M3PP with $J+1$ states, $K = \sum_{i=1}^{J} K_i$ classes, and rates

$$\boldsymbol{D}_0^* = \begin{bmatrix} -\Sigma & \alpha_1 & \alpha_2 & \ldots & \alpha_J \\ \beta_1 & -\Sigma & \alpha_2 & \ldots & \alpha_J \\ \vdots & \ddots & -\Sigma & \ddots & \vdots \\ \beta_1 & \ldots & \beta_{J-1} & -\Sigma & \alpha_J \\ \beta_1 & \ldots & \beta_{J-1} & \beta_J & -\Sigma \end{bmatrix}, \quad \boldsymbol{D}_{1,k}^* = \begin{bmatrix} q_{i,c}\lambda_i \boldsymbol{I}_i & \mathbf{0} \\ \mathbf{0} & q'_{i,c}\lambda'_i \boldsymbol{I}_{J+1-i} \end{bmatrix},$$

where $\boldsymbol{I}_n$ is the identity matrix of order n, class $k = \sum_{j=1}^{i-1} K_j + c$ is the class in the interleaved process associated to class c of the i-th composed M3PP, and the diagonal elements of $\boldsymbol{D}_0^*$ are such that $(\boldsymbol{D}_0^* + \sum_k \boldsymbol{D}_{1,k}^*)\mathbf{1} = \mathbf{0}$. Throughout this section, we denote by $\mathcal{K}_i = \{\sum_{j=1}^{i-1} K_j, \ldots, \sum_{j=1}^{i} K_j\}$ the set of class indexes in the interleaved process associated to the i-th composed M3PP, and by $\bar{\mathcal{K}}_i = \{1, \ldots, K\} \setminus \mathcal{K}_i$ its complement.

The interleaved process may be seen as a M3PP modulated by a CTMC with an infinitesimal generator $\boldsymbol{Q}^* = \boldsymbol{D}_0^* + \sum_{k=1}^{K} \boldsymbol{D}_{1,k}^*$ that allows *exact* CTMC aggregation [4, Chap 4.]. Specifically, for each initial M3PPs, we can define a partition of the states of the interleaved process such that, by exact aggregation of the CTMC $\boldsymbol{Q}^*$ one recovers the corresponding CTMC of the composed M3PP. For example, we may consider the aggregation

$$\boldsymbol{Q}^* = \left[\begin{array}{c:cccc} -r_1 & \alpha_1 & \alpha_2 & \ldots & \alpha_J \\ \hdashline \beta_1 & -r_2 - r'_1 & \alpha_2 & \ldots & \alpha_J \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \beta_1 & \ldots & \beta_{J-1} & -r_J - r'_{J-1} & \alpha_J \\ \beta_1 & \ldots & \beta_{J-1} & \beta_J & -r'_J \end{array}\right] \Rightarrow \begin{bmatrix} -r_1 & \sum_{i=1}^{J} \alpha_i \\ \beta_1 & -\beta_1 \end{bmatrix} = \begin{bmatrix} -r_1 & r_1 \\ r'_1 & -r'_1 \end{bmatrix} = \boldsymbol{Q}_1,$$

where $\boldsymbol{Q}_1 = \boldsymbol{D}_0^1 + \sum_k \boldsymbol{D}_{1,k}^1$ is the infinitesimal generator of the ith composed M3PP. Similarly, for each partition, the definition of $\boldsymbol{D}^*_{1,k}$ ensures that the departure rates are identical to the ones in the composed M3PP.

In order to prove that the marginal counting process of each i-th composed M3PP, for all classes $k \in \mathcal{K}_i$, is preserved by the interleaving operator, we require additional notation. Let $\boldsymbol{P}(\boldsymbol{n}_i, t)$, $\boldsymbol{n}_i = (n_1, \dots, n_{K_i})$, be the counting process matrix for the i-th M3PP at time t and let $\boldsymbol{P}^*(\boldsymbol{n}, t)$ be the corresponding matrix for the interleaved M3PP. Let $\mathbf{0}_n$ and $\mathbf{1}_n$ be row vectors of size n of all zeros and all ones, respectively, and define the aggregation matrix

$$\boldsymbol{V}_i = \begin{bmatrix} \mathbf{1}_i & \mathbf{0}_{J+1-i} \\ \mathbf{0}_i & \mathbf{1}_{J+1-i} \end{bmatrix}$$

so that $diag(\boldsymbol{e}^T \boldsymbol{V}_i^T)^{-1} \boldsymbol{V}_i \boldsymbol{X} \boldsymbol{V}_i^T$ aggregates the state space of any matrix $\boldsymbol{X}$ of the interleaved process returning the corresponding matrix $\boldsymbol{X}_i$ of the i-th composed M3PP [7]. We can now prove the main characterization result for the interleaved process.

Theorem 1. *Assume all processes to be time-stationary. The marginal counting process of the i-th composed M3PP is unchanged in the interleaved process, i.e.,*

$$\boldsymbol{P}(\boldsymbol{n}_i, t) = \boldsymbol{P}^*_i(\boldsymbol{n}, t), \qquad \forall t \geq 0, \tag{22}$$

where $\boldsymbol{P}^*(\boldsymbol{n}_i, t) = \sum_{n_h : h \in \bar{\mathcal{K}}_i} \boldsymbol{P}^*(\boldsymbol{n}, t)$ *is the marginal counting process matrix for the* $i-th$ *composed M3PP and* $\boldsymbol{P}^*_i(\boldsymbol{n}, t) = \boldsymbol{W}_i \boldsymbol{P}^*(\boldsymbol{n}, t) \boldsymbol{V}_i^T$, *where* $\boldsymbol{W}_i = diag(\boldsymbol{e}^T \boldsymbol{V}_i^T)^{-1} \boldsymbol{V}_i$. *Furthermore,* $\boldsymbol{P}(\boldsymbol{n}_i, t)$ *and* $\boldsymbol{P}^*_i(\boldsymbol{n}, t)$ *are initialized in the same state.*

Proof. Note that $\boldsymbol{P}^*_i(\boldsymbol{n}, t)$ differs from $\boldsymbol{P}^*(\boldsymbol{n}, t)$ for being defined on the aggregate state space where we lump the first i states into a new state labelled by 1 and the last $J+1-i$ states into a new state labelled by 2. Also it is simple to verify that, for the interleaved M3PP, $\boldsymbol{V}_i$ has the following properties

$$\boldsymbol{Q}^* \boldsymbol{V}_i^T = \boldsymbol{V}_i^T \boldsymbol{Q}_i, \qquad \boldsymbol{D}^*_{1,k} \boldsymbol{V}_i^T = \boldsymbol{V}_i^T \boldsymbol{D}^i_{1,k}, \qquad \forall k : K_{i-1} + 1 \leq k \leq K_i, \tag{23}$$

Note that conditions of this kind arise naturally also in the study of minimal representations of Markovian and rational arrival processes [9].

Let us consider the Kolmogorov forward equation for $\boldsymbol{P}^*(\boldsymbol{n}, t)$. Pre-multiplying (1) by $\boldsymbol{W}_i = diag(\boldsymbol{e}^T \boldsymbol{V}_i^T)^{-1} \boldsymbol{V}_i$ and post-multiplying by $\boldsymbol{V}_i^T$, we obtain

$$\frac{\partial \boldsymbol{P}^*_i(\boldsymbol{n}, t)}{\partial t} = \boldsymbol{W}_i \boldsymbol{P}^*(\boldsymbol{n}, t) \boldsymbol{D}^*_0 \boldsymbol{V}_i^T + \textstyle\sum_{k \in \mathcal{K}_i} \boldsymbol{P}^*_i(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}^i_{1,k} + \textstyle\sum_{k \in \bar{\mathcal{K}}_i} \boldsymbol{W}_i \boldsymbol{P}^*(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}^*_{1,k} \boldsymbol{V}_i^T$$

where we have used (23) to simplify the expression. Now using the fact that $\boldsymbol{D}^*_0 = \boldsymbol{Q}^* -$

$\sum_{k=1}^{K_i} \boldsymbol{D}_{1,k}^*$, we can write

$$\frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n},t)}{\partial t} = \boldsymbol{P}_i^*(\boldsymbol{n},t)\boldsymbol{D}_0^i + \textstyle\sum_{k\in\mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n}-\boldsymbol{e}_k,t)\boldsymbol{D}_{1,k}^i + \sum_{k\in\bar{\mathcal{K}}_i} \boldsymbol{W}_i(\boldsymbol{P}^*(\boldsymbol{n}-\boldsymbol{e}_k,t) - \boldsymbol{P}^*(\boldsymbol{n},t))\boldsymbol{D}_{1,k}^*\boldsymbol{V}_i^T.$$

where the $\boldsymbol{D}_0^i$ matrix appears in the first term of the right-hand side follows after noting that $\boldsymbol{Q}_i - \sum_{k\in\mathcal{K}_i} \boldsymbol{D}_{1,k}^i = \boldsymbol{D}_0^i$. We now compute the marginal probabilities on the $\boldsymbol{n}_i$ vector as

$$\begin{aligned}\frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n}_i,t)}{\partial t} = {} & \textstyle\sum_{n_h:h\in\bar{\mathcal{K}}_i} \boldsymbol{P}_i^*(\boldsymbol{n},t)\boldsymbol{D}_0^i + \sum_{n_h:h\in\bar{\mathcal{K}}_i}\sum_{k\in\mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n}-\boldsymbol{e}_k,t)\boldsymbol{D}_{1,k}^i \\ & + \textstyle\sum_{n_h:h\in\bar{\mathcal{K}}_i}\sum_{k\in\bar{\mathcal{K}}_i} \boldsymbol{W}_i(\boldsymbol{P}^*(\boldsymbol{n}-\boldsymbol{e}_k,t) - \boldsymbol{P}^*(\boldsymbol{n},t))\boldsymbol{D}_{1,k}^*\boldsymbol{V}_i^T \end{aligned} \quad (24)$$

Now note that, since the infinite summations in the last term of (24) are on the same class indexes (set $\bar{\mathcal{K}}_i$), for each $\boldsymbol{P}^*(\boldsymbol{n}-\boldsymbol{e}_k,t)$ term with $n_k > 0$, the outer summation on the n_h vectors introduces in the summation a term equivalent to $-\boldsymbol{P}^*(\boldsymbol{n}-\boldsymbol{e}_k,t)$. Since there are infinite states, the last term of (24) vanishes in the summation. For the boundary states where $n_k = 0$, $\boldsymbol{P}^*(\boldsymbol{n}-\boldsymbol{e}_k,t) = 0$ and thus it does not appear in the expression as well. Thus (24) reduces to

$$\frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n}_i,t)}{\partial t} = \boldsymbol{P}_i^*(\boldsymbol{n}_i,t)\boldsymbol{D}_0^i + \textstyle\sum_{k\in\mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n}_i-\boldsymbol{e}_k,t)\boldsymbol{D}_{1,k}^i$$

which is identical to the Kolmogorov forward equation for the counting process of the i-th composed M3PP. This shows that $\boldsymbol{P}^*(\boldsymbol{n}_i,t)$ and $\boldsymbol{P}_i^*(\boldsymbol{n}_i,t)$ are stochastically equivalent, provided that they start from the same initial state. Since we are considering time-stationary processes, the initial state is determined by the equilibrium distribution of $\boldsymbol{Q}^*$. It is straightforward to see that the exact aggregations preserve the structure of the global balance equations of the CTMC of the i-th initial M3PPs, since they apply an exact aggregation. Thus the distribution of the probability mass across the two states at equilibrium in the i-th composed M3PP must be identical to the one in the marginal counting process for M3PP i in the interleaved process. □

Importantly, equation (23) in the proof provides a condition for a *general* MMAP to admit an interleaving process. It is simple to see that the diagonal structure of the $\boldsymbol{D}_{1,k}$ matrices in a M3PP satisfies these assumptions. While our analysis does not exclude that other kinds of MMAPs may be used for interleaving, the conditions required to satisfy (23) do not seem to readily suggest alternatives to the M3PP structure. This motivates the use of M3PPs as a building block for the interleaving composition.

Feasibility of an Interleaving. We now give examples of valid and invalid interleaved processes. Consider the following two-state M3PP[2]s:

$$\boldsymbol{A}_0 = \begin{bmatrix} -6 & 3 \\ 1 & -5 \end{bmatrix}, \ \boldsymbol{B}_0 = \begin{bmatrix} -3 & 1 \\ 4 & -7 \end{bmatrix}, \ \boldsymbol{G}_0 = \begin{bmatrix} -9 & 7 \\ 2 & -5 \end{bmatrix}, \ \boldsymbol{E}_0 = \begin{bmatrix} -7 & 4 \\ 1 & -3 \end{bmatrix},$$

where $\boldsymbol{A}_{1,1} = \text{diag}(1,3)$, $\boldsymbol{A}_{1,2} = \boldsymbol{E}_{1,1} = \text{diag}(2,1)$, $\boldsymbol{B}_{1,1} = \boldsymbol{G}_{1,1} = \text{diag}(1,2)$, and $\boldsymbol{B}_{1,2} = \boldsymbol{G}_{1,2} = \boldsymbol{E}_{1,2} = \text{diag}(1,1)$. It is easy to see that the interleaving of processes $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfies the assumptions on the non-negativity of the rates α_j and β_j and yields the interleaved M3PP[4]

$$\boldsymbol{D}_0^* = \begin{bmatrix} -8 & 2 & 1 \\ 1 & -8 & 1 \\ 1 & 3 & -11 \end{bmatrix},$$

with $\boldsymbol{D}_{1,1}^* = \text{diag}(1,3,3)$, $\boldsymbol{D}_{1,2}^* = \text{diag}(2,1,1)$, $\boldsymbol{D}_{1,3}^* = \text{diag}(1,1,2)$, $\boldsymbol{D}_{1,4}^* = \text{diag}(1,1,1)$. The per-class arrival rates and variances of counts in the interleaved process match the corresponding statistics of the two classes of $\boldsymbol{A}$ and the two classes of $\boldsymbol{B}$.

Conversely, the interleaving of processes $\boldsymbol{A}$ and $\boldsymbol{C}$ is instead invalid because the non-diagonal rates of $\boldsymbol{G}_0$ are both greater than the corresponding rates in $\boldsymbol{A}_0$. Similarly, the interleaving of $\boldsymbol{A}$ and $\boldsymbol{E}$ is invalid, even though $\boldsymbol{B}$ and $\boldsymbol{E}$ are the same process, because their states are ordered in a different manner, which affects the computation if the α_j and β_j coefficients. This last example provides intuition on the fact that, in order to foster feasibility of the interleaved process, one may need to re-order the states of the M3PPs. In the next section, we describe an algorithm to find automatically a feasible interleaving of a given set of M3PPs, if one exists.

Class Covariances. While the interleaved process preserves the marginal counting properties of the composed M3PPs, when compared to superposition this comes at the expense of introducing spurious covariances between arrivals of different M3PPs. This is because, by definition of the inter-leaved process, a transition in the state space of a composed M3PP also changes the current state of the other composed M3PPs. In order to quantify the magnitude of these covariances, we look at the asymptotic covariances, which contribute to the index of dispersion. Observe that, by plugging (3) into (4), we find after some simplifications

$$\sigma_{k,h} = \lim_{t\to\infty} \text{Cov}[n_k(t), n_h(t)] = -2\mu_k\mu_h + \boldsymbol{\pi}\boldsymbol{D}_{1,h}(\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{Q})^{-1}\boldsymbol{D}_{1,k}\boldsymbol{1} + \boldsymbol{\pi}\boldsymbol{D}_{1,k}(\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{Q})^{-1}\boldsymbol{D}_{1,h}\boldsymbol{1}$$

Let $\boldsymbol{P} = (\boldsymbol{1}\boldsymbol{\pi} - \boldsymbol{Q})^{-1}$ and observe that this by construction is a stochastic matrix with equilibrium vector $\boldsymbol{\pi}$. Using the fact that $\boldsymbol{\pi}\boldsymbol{D}_{1,k} = \mu_k$, for any class k, and after determining the structure of the $\boldsymbol{P}$ matrices for the interleaved process, it is possible to compute their Jordan canonical form, which after algebraic simplifications yields the formula

$$\sigma_{k,h} = \frac{r'_i r_j (q_{j,h}\lambda_j - q'_{j,h}\lambda'_j)(q_{i,k}\lambda_i - q'_{i,k}\lambda'_i)(x_j + x_i)}{x_j^2 x_i^2}, \tag{25}$$

where it is assumed that $k \in \mathcal{K}_i$ and $h \in \mathcal{K}_j$, and $i < j$.

Some remarks on the formulas are as follows. First, when any of the two M3PPs tends to a Poisson process, a pair of departure rates at the numerator of (25) annihilates and $\sigma_{k,h}$ goes to zero. Next, the order of the denominator suggests that a way to reduce the covariance introduced by interleaving consists in spending degrees of freedom to maximize x_i and x_j in the embedded MMPPs, for fixed arrival rates. When applying such a scheme, one should however take into account the bounds in Proposition 3, since an increase of the value of x_i also reduces fitting flexibility in the embedded MMPP. Lastly, since $x_i > 0$ for any i, we note that the sign of the covariance is determined only by the differences between the per-class arrival rates within each of the composed M3PPs.

5. Fitting the Interleaved Process

In this section we consider two issues that arise in the compositional fitting with interleaving. First, we consider the decision problem involving the mapping of a marked trace into a set of two-states M3PPs. Then we show that the problem of finding a feasible interleaving of a set of J second-order M3PPs, where the i-th M3PP models any subset $\mathcal{K}_i$ of the K classes, can be formulated as a Mixed-Integer Linear Program and hence solved efficiently within the class of integer programming problems.

5.1. Fitting a marked trace into a set of M3PPs

Given a trace with arrivals of K classes, if the arrivals are uncorrelated it would be appropriate to fit each of the classes using a superposition of K independent MMPPs, one for each class, and then reduce the size of the resulting process using interleaving. However, if the classes have a significantly large covariance, this procedure may not result into a good fitting. Therefore, we propose two heuristic fitting methods that reflect these two cases. We assess the effectiveness of these heuristics in Section 6.

Independent Method. In this method, we ignore class covariances and fit each class into a separate two-state MMPP. The method is similar to superposition, but returns a much smaller M3PP[K] with $K+1$ states, instead of 2^K states. Interleaving uses the order of composition obtained from the MILP method that we present in Section 5.2.

Covariance-based Method. We initially build the asymptotic co-variance matrix $\mathbf{\Sigma} = [\sigma_{k,h}] = \lim_{t\to\infty} \mathrm{Cov}[n_k(t), n_h(t)]$ between each pair of classes. We approximate the asymptotic timescale with the largest finite t for which we can average counts over 100 samples. The user is then requested to specify a covariance threshold δ. We then iterate on the classes, in decreasing order of variance $\sigma_{k,k}$. For each class k, we first determine all the classes $h \neq k$ with $\sigma_{h,k} \geq \delta$ and record the decision to fit these classes into the same M3PP[K]. The algorithm then continues with analyzing in a similar way the other classes not already planned for fitting in a M3PP. After this stage, each class k is mapping to a unique M3PP i. In order to obtain a representation for each M3PP i, we run the algorithm given in Section 5.2, which returns the parameters of the embedded MMPPs and the order of composition of the M3PPs in the interleaving. With these, we can conclude the fitted by fitting the M3PPs with the method in Section 3.2.1 and the applying the interleaving operator.

5.2. Finding a feasible inter-leaving of M3PPs

The definition of the interleaved process implies that the $\boldsymbol{D}_{1,k}$ matrices are always feasible by construction. Thus, infeasible M3PPs may arise only if the $\boldsymbol{D}_0$ matrix has some negative element. Since the entries of the $\boldsymbol{D}_0$ matrix depend only on the MMPPs embedded in the M3PPs to be composed, a procedure to determine a feasible interleaving should be run prior to fitting the MMPPs.

To find a feasible interleaving, we consider permutations of the assignment of the rates to the states of the underlying MMPP and the order in which the J M3PPs are composed. Furthermore, when deciding the structure of the MMPPs, we exploit the degree of freedom given by ignoring the fitting of the third moment of counts. Let x_i and u_i, be the values of x and u for the i-th embedded MMPP. Sacrificing the fitting of the third moment allows us to decide the value for r_i, given x_i, which then implies $r'_i = x_i - r_i$. Recall from Section 3 that when ignoring the third moment of counts, infinite feasible MMPPs exist as long as $u_i \leq r_i < x_i$ and $r'_i = x_i - r_i$, where x_i is found by (6). These bounds have been obtained in Section 3 taking the assumption that $\lambda_i > \lambda'_i$. While this assumption does not have any impact on the feasibility of individual M3PPs, it has an impact on the feasibility of the interleaving. Making the opposite assumption $\lambda_i \leq \lambda'_i$, we obtain the specular constraints $u_i \geq r'_i > x_i$ and $r_i = x_i - r'_i$. Thus, the above bounds will need to be both considered and only one will be active depending on the relative value of λ_i to λ'_i, which we decide using a binary variable b_i that is responsible for state order. Summarizing, the above bounds allow us to obtain a feasible interleaving by choosing from a continuous set of feasible MMPPs.

Before solving the MILP formulation that decides the order of states and the order of composition, we assume that the fixed point equation (6) has been solved for each M3PP i

for the values of x_i and u_i. For the MILP formulation, we then set the following decision variables:

- the rates $r_i > 0$ and $r'_i > 0$ of the i-th M3PP, $\forall i \in \{1, ..., J\}$;
- the integer variables $p_i \in \mathbb{N}$, $\forall i \in \{1, ..., J\}$, for the position of the i-th M3PP in the interleaved process, i.e., the order of composition.
- the auxiliary binary variables $z_{i,j} \in \{0, 1\}$, $\forall (i, j) \in \{1, ..., J\}^2$, $i \neq j$, such that $z_{i,j} = 1$ if and only if $p_i \leq p_j$;
- the binary variables $b_i \in \{0, 1\}$, $\forall i \in \{1, ..., J\}$, encoding the choice of the state order for the i-M3PP, with $b_i = 0$ if $\lambda_i > \lambda'_i$ and $b_i = 1$ if $\lambda_i < \lambda'_i$.

Overall, the number of variables is $\mathcal{O}(J^2)$, i.e., quadratic in the number of composed M3PPs.

Let M be a large constant. We consider the following MILP formulation:

$$
\begin{aligned}
\text{minimize} \quad & 0 \quad \text{(feasibility problem)} && && (26)\\
\text{s.t.:} \quad & u_i(1 - b_i) \leq r_i < x_i - u_i b_i && \forall i \in \{1, ..., J\} && (27)\\
& u_i b_i \leq r'_i < x_i - u_i(1 - b_i) && \forall i \in \{1, ..., J\} && (28)\\
& z_{i,j} + z_{j,i} = 1 && \forall (i, j) \in \{1, ..., J\}^2 : i \neq j && (29)\\
& z_{i,j} \geq p_j - p_i && \forall (i, j) \in \{1, ..., J\}^2 && (30)\\
& r_i \geq r_j - (1 - z_{i,j})M && \forall (i, j) \in \{1, ..., J\}^2, i \neq j && (31)\\
& r'_i \leq r'_j + (1 - z_{i,j})M && \forall (i, j) \in \{1, ..., J\}^2, i \neq j && (32)\\
& p_i \neq p_j && \forall (i, j) \in \{1, ..., J\}^2 : i \neq j && (33)
\end{aligned}
$$

Observe that the objective function is a constant, since we just need a feasible solution. Constraint (27) imposes the bounds on r_i, with different bounds depending on whether the state order of the i-M3PP is inverted or not. Strict bounds are obtained by adding small tolerance to the corresponding formulation with $\leq$ inequalities. Constraint (28) imposes specular bounds on r'_i. Antisymmetry constraints on the auxiliary variables $z_{i,j}$ are in (29). Constraint (30) ensures that $z_{i,j} = 1$ if and only if $p_i \leq p_j$. This constraint together with the uniqueness constraint on p_i ensures the transitivity property for the binary variables $z_{i,j}$. Ordering constraints on r_i and r'_i are expressed by (31) and (32), respectively. Two M3PP processes cannot be in the same position due to (33); note that inequality constraints can be handled for integer variables in MILP.

After solving the MILP, we have the parameters r_i and r'_i of each M3PP. Applying the fitting formulas of Section 3 we obtain the remaining parameters of the embedded MMPP, λ_i and λ'_i, taking care to swap the two values when states are in inverted order, i.e., $b_i = 1$.

Then we compute the class parameters $q_{i,k}$ and $q'_{i,k}$ for each class of the i-th M3PP and build the interleaving of the J M3PPs as in Section 4.

6. Fitting Results

6.1. Fitting Random MMAPs

We being by examining the ability of two-state M3PPs and the interleaving process to fit the characteristics of randomly generated processes. We consider random M3PP[K]s, with $m \in \{4, 8, 16\}$ states and $K \in \{2, 3, 4\}$ classes. For each choice of m and K, we generate 100 random M3PP[K] as follows. First, we generate a random infinitesimal generator $\boldsymbol{Q} = \boldsymbol{D}_0 + \boldsymbol{D}_1$ using uniform random numbers. Then, for each class k, we first compute a vector of m uniform random numbers $\boldsymbol{u}$ and then we set $\boldsymbol{D}_{1,k} = \exp(\text{diag}(5\boldsymbol{u}))$. Lastly, we set $\boldsymbol{D}_1 = \sum_k \boldsymbol{D}_{1,k}$ and $\boldsymbol{D}_0 = \boldsymbol{Q} - \boldsymbol{D}_1$. The expression used to compute $\boldsymbol{D}_{1,k}$ provides a set of processes with index of dispersion I that is 3-6 times larger than the squared coefficient of variation (SCV), as shown in Table 6.1, which gives statistics for the randomly generated M3PPs. The statistics in each row are averaged across the 100 random instances and we report mean ± standard deviation. The third columns gives the average ratio

$$\text{Cov }\% = \lim_{t\to\infty} \frac{\sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} |\text{Cov}[n_k(t), n_{k'}(t)]|}{\sum_{c=1}^{K}\sum_{c'=1}^{K} |\text{Cov}[n_c(t), n_{c'}(t)]|} \tag{34}$$

M3PPs are fitted as follows. We first compute the statistical descriptors of the random process using theoretical expressions. Two-state M3PPs are then obtained using the method described in Section 3.2.1. The interleaving process is fitted using the two methods described in Section 5.1 and the feasibility method in Section 5.2. The latter is implemented in MATLAB using the YALMIP [24] language and the CBC branch-and-cut solver [25]. Superposition is implemented as by the definition after fitting an independent MMPP for each class. Here and in the following sections, class counts used to fit a class are obtained from inter-arrival times between successive arrivals of that same class.

Let $n_k(t)$ be the number of arrivals of class k in t time units for the randomly-generated M3PP and let $\tilde{n}_k(t)$ be the analogous quantity for the fitted model. The metrics used to assess the quality of the fitting are the ability of the model to capture the asymptotic class variances and covariances, as quantified by the absolute relative errors

$$\epsilon_{var} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K} |\text{Var}[n_k(t)] - \text{Var}[\tilde{n}_k(t)]|}{\sum_{c=1}^{K} \text{Var}[n_c(t)]} \tag{35}$$

$$\epsilon_{cov} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} |\text{Cov}[n_k(t), n_{k'}(t)] - \text{Cov}[\tilde{n}_k(t), \tilde{n}_{k'}(t)]|}{\sum_{c=1}^{K}\sum_{c'=1,c'\neq c}^{K} |\text{Cov}[n_c(t), n_{c'}(t)]|} \tag{36}$$

We focus on second order class descriptors of the counting process since mean arrival rates can always be fitted exactly. We also quantify the absolute relative error on the variance of the underlying unmarked process

$$\epsilon_{unmark} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K} |\mathrm{Cov}[n_k(t), n_{k'}(t)] - \mathrm{Cov}[\tilde{n}_k(t), \tilde{n}_{k'}(t)]|}{\sum_{c=1}^{K}\sum_{c'=1}^{K} |\mathrm{Cov}[n_c(t), n_{c'}(t)]|} \tag{37}$$

Lastly, we assess the total number of states m^{fit} of the fitted model and the time T required for the fitting algorithm to return a valid M3PP.

Results. Results of the validation against random M3PPs are given in Table 6.1. Remarks are as follows:

- The results indicate that interleaving has the same efficiency of superposition in capturing class variances, whereas two-state M3PPs are better at capturing class covariances. This aligns with expectations, given that interleaving matches marginal counting processes, whereas the method presented in Section 3.2.1 for two-state M3PPs matches asymptotic covariances.
- It is fairly surprising to note the good performance of two-state M3PPs in fitting even large processes, with several states and classes. However, we conjecture this to be because the arrivals have the same inter-arrival time distribution of the embedded MMPP. In fact, we will show that when this assumption is removed in histogram-based fitting of real traces, two-state M3PPs perform worse.
- Interleaving is on most instances superior to superposition. The number of states is significantly lower, without appreciable loss of accuracy in matching variances. It should be noted that, while in some cases interleaving performs worse than superposition in matching covariances, superposition always return zero covariance, thus it is uninformative. Conversely, the error on covariance of interleaving appears to decrease with growing number of states and classes.
- The covariance-based fitting method used for interleaving typically fares better than the interleaving of independent MMPPs. The approach has similar accuracy for variance matching, it is generally more accurate in describing covariances, and requires less states.
- Computational times are small for all methods. However, there is an appreciable difference in the time to compute an interleaving for large processes, due to the cost of solving the integer program introduced in Section 5.2. Still, it should be noted that when the number of classes grows beyond $K = 4$, superposition requires tens or even hundreds of states, thus it becomes far less tractable than interleaving.

m	K	Cov %	SCV	$I(\infty)$
4	2	0.29	5.25 ± 4.17	32.06 ± 35.47
8	2	0.31	5.47 ± 2.54	21.57 ± 11.99
16	2	0.30	4.55 ± 1.15	11.43 ± 36.50
4	4	0.57	5.25 ± 3.90	27.27 ± 22.12
8	4	0.57	5.28 ± 2.22	19.93 ± 9.13
16	4	0.56	4.40 ± 1.07	11.94 ± 3.42

Table 1: Statistics for randomly generated M3PPs. Each row refers to 100 random instances.

6.2. Fitting Real Traces

The analysis for random MMAPs is now repeated for fitting the BC-pAug89 and LBL-TCP-3 traces from the Internet Traffic Archive[1]. While these traces are commonly used in the literature of Markov-modulated processes, they are unmarked and thus cannot be readily used for validation of marked processes. In order to do so, we introduce a labelling that associates each arrival to a different bin of the histogram of packet sizes. We first consider the case where arrivals belong only to one of $K = 2$ classes, putting the separator at the p-th percentile of the inter-arrival time distribution, where $p \in \{25, 50, 75, 90\}$. In addition, we consider a separation of the histogram into five bins, defined by the same set of percentiles, which leads to a model with $K = 5$ classes.

Results are given in tables 3 and 4. The trends in the two tables are qualitatively similar and indicate that interleaving can fit better the traces than a 2-state model and superposition on most instances. There are also several instances on which the interleaved M3PPs have better ϵ_{cov} than the 2-state M3PP. This suggests that the good results on random instances for 2-state models may be biased by the fact that inter-arrival times of different classes are identically distributed in the random instances. This does not happen with the real traces, since the histogram bins do not overlap with each other.

It can also be noted that, as the number of classes grows to $K = 5$, the performance of the different methods get closer to each other, presumably due to the increased difficulty in matching class covariances.

6.2.1. Queueing Analysis

Lastly, we consider a queueing analysis problem. We use again the LBL-TCP-3 trace, but with the class marking defined in [8]. This marking defines 4 classes, each associated to a different bin of the histogram of packet sizes, and having service rates $\mu_1 = 300$, $\mu_2 = 250$, $\mu_3 = 200$, $\mu_4 = 100$. We simulate a MMAP/M/1/10 queue with non-preemptive priority scheduling (small packet sizes are favored) and buffer size of 10 jobs. In the simulations, we

[1] http://ita.ee.lbl.gov/

m	K	*Method*	ϵ_{unmark}	ϵ_{var}	ϵ_{cov}	m^{fit}	T [s]
4	2	2-state	0.184	0.118	0.972	2.00	0.181
4	2	superposition	0.296	0.002	1.000	4.00	0.171
4	2	interleaving-independent	0.218	0.002	2.103	3.00	1.722
4	2	interleaving-covariance	0.197	0.048	1.870	2.55	1.533
8	2	2-state	0.197	0.129	0.805	2.00	0.183
8	2	superposition	0.310	0.001	1.000	4.00	0.177
8	2	interleaving-independent	0.212	0.001	1.464	3.00	1.825
8	2	interleaving-covariance	0.188	0.022	1.321	2.72	1.681
16	2	2-state	0.227	0.153	0.600	2.00	0.175
16	2	superposition	0.300	0.000	1.000	4.00	0.184
16	2	interleaving-independent	0.202	0.000	1.024	3.00	1.814
16	2	interleaving-covariance	0.197	0.018	0.858	2.88	1.705
4	4	2-state	0.160	0.117	0.216	2.00	0.187
4	4	superposition	0.573	0.002	1.000	16.00	0.333
4	4	interleaving-independent	0.409	0.002	0.755	5.00	4.439
4	4	interleaving-covariance	0.280	0.074	0.491	3.64	2.788
8	4	2-state	0.180	0.147	0.223	2.00	0.200
8	4	superposition	0.574	0.001	1.000	16.00	0.361
8	4	interleaving-independent	0.306	0.001	0.569	5.00	5.157
8	4	interleaving-covariance	0.275	0.042	0.477	4.02	3.700
16	4	2-state	0.222	0.201	0.246	2.00	0.175
16	4	superposition	0.560	0.000	1.000	16.00	0.331
16	4	interleaving-independent	0.320	0.000	0.594	5.00	4.978
16	4	interleaving-covariance	0.307	0.019	0.557	4.72	4.570

Table 2: Fitting results for randomly generated M3PPs.

use 10^8 samples, record mean throughput and mean queue-length, and obtain mean response times by Little's law. The following fitting methods are compared: (i) superposition of 4 MMPPs, one per class (16 states); (ii) interleaving of the independent MMPPs used for the superposition, one per class (5 states); (iii) the interleaved process obtained by the covariance-based method, which interleaves 3 M3PP[2], the first for classes 1 and 3, which have the largest covariances, the second for class 2, the third for class 4 (4 states); (iv) the fitting of a M3PP[4] (2 states).

In Table 5 we give the average errors on the mean throughput, mean queue-length, and mean response time as the services rates of the 4 classes are multiplied by a factor $s = 0.5, 0.75, \ldots, 2.00$, for a total of 7 experiments. The result indicate that interleaving is the best method for mean queue-length and mean response time prediction, whereas the 2-state M3PP[4] is the best in throughput prediction. Combining the average errors, the two interleaving methods fare quite similarly and better than superposition and the 2-state M3PP[4]. Summarizing, the experiment suggests that 2-state M3PPs and interleaving com-

Trace	K	*Cut-points* p	*Method*	ϵ_{unmark}	ϵ_{var}	ϵ_{cov}	m	T
BC-pAug89	2	25	2-state	0.718	0.635	0.984	2	1.159
BC-pAug89	2	25	superposition	0.277	0.052	1.000	4	0.922
BC-pAug89	2	25	interleaving-independent	0.278	0.052	1.005	3	1.436
BC-pAug89	2	25	interleaving-covariance	0.163	0.052	0.519	3	1.489
BC-pAug89	2	50	2-state	0.358	0.357	0.360	2	1.263
BC-pAug89	2	50	superposition	0.377	0.048	1.000	4	1.156
BC-pAug89	2	50	interleaving-independent	0.070	0.048	0.113	3	2.564
BC-pAug89	2	50	interleaving-covariance	0.070	0.047	0.112	3	2.064
BC-pAug89	2	75	2-state	0.257	0.186	0.646	2	0.880
BC-pAug89	2	75	superposition	0.195	0.047	1.000	4	0.699
BC-pAug89	2	75	interleaving-independent	0.083	0.047	0.278	3	1.435
BC-pAug89	2	75	interleaving-covariance	0.082	0.046	0.277	3	1.642
BC-pAug89	2	90	2-state	0.257	0.186	0.646	2	0.963
BC-pAug89	2	90	superposition	0.195	0.047	1.000	4	0.777
BC-pAug89	2	90	interleaving-independent	0.195	0.047	1.001	3	1.550
BC-pAug89	2	90	interleaving-covariance	0.195	0.046	1.001	3	1.425
BC-pAug89	5	$25, 50, 75, 90$	2-state	0.382	0.399	0.363	2	1.342
BC-pAug89	5	$25, 50, 75, 90$	superposition	0.487	0.045	1.000	16	1.223
BC-pAug89	5	$25, 50, 75, 90$	interleaving-independent	0.313	0.045	0.624	5	2.363
BC-pAug89	5	$25, 50, 75, 90$	interleaving-covariance	0.680	0.045	1.416	5	2.380

Table 3: Fitting results for the BC-pAug89 trace for different cut-points of the packet size histogram.

pare favourably to a basic superposition of class arrivals, despite that their state space is considerably smaller. Overall, it seems possible to conclude that the methods introduced in this paper can be helpful for real-world queueing analysis.

7. Conclusion

In this paper, we have presented novel methods to fit multiclass arrival processes using marked Markov-modulated Poisson processes (M3PPs). We have defined exact and approximate algorithms to fit two-state M3PPs with arbitrary number of classes and introduced a new composition operator, called interleaving, which enables composing several M3PPs while preserving their marginal counting processes. Remarkably, the state space of the interleaved process grows linearly with the total number of composed M3PPs, instead than exponentially as in ordinary superposition. Experiments reveal that both the two-state M3PP and the interleaved process can help in fitting real-world traces, in matching the descriptors of more complex marked processes, and in queueing analysis.

We also remark that even though we did not develop this case in the paper, inter-leaving can also be applied to MMPPs, thus it may also contribute to the fitting of unmarked processes. However, the main challenge in this case is to define a fitting method for unmarked

Trace	*K*	*Cut-points p*	*Method*	ϵ_{unmark}	ϵ_{var}	ϵ_{cov}	*m*	*T*
LBL-TCP-3	2	25	2-state	0.468	0.661	0.185	2	1.768
LBL-TCP-3	2	25	superposition	0.426	0.034	1.000	4	1.600
LBL-TCP-3	2	25	interleaving-independent	0.709	0.034	1.699	3	1.826
LBL-TCP-3	2	25	interleaving-covariance	0.064	0.085	0.034	2	2.502
LBL-TCP-3	2	50	2-state	0.454	0.357	0.700	2	1.647
LBL-TCP-3	2	50	superposition	0.313	0.043	1.000	4	1.516
LBL-TCP-3	2	50	interleaving-independent	0.032	0.042	0.005	3	1.765
LBL-TCP-3	2	50	interleaving-covariance	0.324	0.060	0.995	3	1.845
LBL-TCP-3	2	75	2-state	0.563	0.383	1.207	2	1.661
LBL-TCP-3	2	75	superposition	0.252	0.043	1.000	4	1.428
LBL-TCP-3	2	75	interleaving-independent	0.061	0.043	0.125	3	1.925
LBL-TCP-3	2	75	interleaving-covariance	0.308	0.118	0.989	3	2.053
LBL-TCP-3	2	90	2-state	0.175	0.107	8.916	2	2.090
LBL-TCP-3	2	90	superposition	0.045	0.037	1.000	4	1.528
LBL-TCP-3	2	90	interleaving-independent	0.045	0.037	1.009	3	2.200
LBL-TCP-3	2	90	interleaving-covariance	0.045	0.037	1.009	3	2.650
LBL-TCP-3	5	$25, 50, 75, 90$	2-state	0.660	0.493	0.858	2	2.646
LBL-TCP-3	5	$25, 50, 75, 90$	superposition	0.482	0.047	1.000	32	2.484
LBL-TCP-3	5	$25, 50, 75, 90$	interleaving-independent	0.663	0.047	1.397	6	3.690
LBL-TCP-3	5	$25, 50, 75, 90$	interleaving-covariance	0.611	0.047	1.284	6	4.571

Table 4: Fitting results for the LBL-TCP-3 trace for different cut-points of the packet size histogram.

Fitting Method	*Queue-length Mean Error*	*Response time Mean Error*	*Throughput Mean Error*	*Average Combined*
2-state	0.160	0.181	0.055	0.132
superposition	0.131	0.148	0.062	0.137
interleave-independent	0.119	0.138	0.083	0.113
interleave-covariance	0.120	0.128	0.084	0.111

Table 5: Queueing prediction errors for MMAP/M/1/10 priority queue fed by the marked LBL-TCP-3 trace.

traces capable of identify a set of MMPPs that simultaneously satisfy the requirements for interleaving. Future work should also investigate how interleaving and exact aggregations may be fruitfully applied to the general class of MMAPs. Also, it would be interesting to apply M3PPs to decomposition analysis of multiclass queueing networks, for example by iterative matching the departure flows of the requests from a queue processing marked arrivals.

Acknowledgement

The work of Giuliano Casale is supported by EPSRC, under grant no. EP/M009211/1.

References

[1] A. T. Andersen and B. F. Nielsen. A Markovian approach for modeling packet traffic with long-range dependence. *IEEE Journal on Selected Areas in Communications*, 16(5):719–732, 1998.

[2] D. Bini, B. Meini, S. Steffé, Juan F. Pérez, and B. Van Houdt. SMCSolver and Q-MAM: tools for matrix-analytic methods. *ACM Performance Evaluation Review*, 39(4):46, 2012.

[3] L. Bodrog, A. Heindl, G. Horváth, and M. Telek. A Markovian canonical form of second-order matrix-exponential processes. *European Journal of Operations Research*, 190(2):459–477, 2008.

[4] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains.* John Wiley and Sons, 1998.

[5] L. Breuer. An EM algorithm for batch markovian arrival processes and its comparison to a simpler estimation procedure. *Annals of Operations Research*, 112:123–138, 2002.

[6] J.W. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, 25(9):772–781, 1978.

[7] P. Buchholz. Exact and Ordinary Lumpability in Finite Markov Chains. *Journal of Applied Probability*, 31(1):59–75, 1994.

[8] P. Buchholz, P. Kemper, and J. Kriege. Multi-class Markovian arrival processes and their parameter fitting. *Performance Evaluation*, 67(11):1092 – 1106, Elsevier, 2010.

[9] P. Buchholz, M. Telek. On Minimal Representations of Rational Arrival Processes. *Annals of Operations Research*, 202(1):35 – 58, Elsevier, 2013.

[10] G. Casale, E. Z. Zhang, and E. Smirni. Trace data characterization and fitting for markov modeling. *Performance Evaluation*, 67(2):61–79, Elsevier, 2010.

[11] W. Fischer and K. S. Meier-Hellstern. The markov-modulated poisson process (MMPP) cookbook. *Performance Evaluation*, 18(2):149–171, Elsevier, 1993.

[12] Q.-M. He. The versatility of MMAP[K] and the MMAP[K]/G[K]/1 queue. *Queueing Systems: Theory and Applications*, 38(4):397–418, Springer, 2001.

[13] Q.-M. He and M.F. Neuts. Markov chains with marked transitions. *Stochastic Processes and their Applications*, 74(1):37–52, 1998.

[14] H. Heffes and D. Lucantoni. A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, 1986.

[15] A. Horváth and 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*, *Lecture Notes in Computer Science*, 2459:405–434, Springer, 2002.

[16] A. Heindl, G. Horváth, and K. Gross. Explicit inverse characterizations of acyclic MAPs of second order. In *EPEW*, *Lecture Notes in Computer Science*, 4054:108–122, Springer, 2006.

[17] A. Horváth, G. Horváth, and M. Telek. A traffic based decomposition of two-class queueing networks with priority service. *Computer Networs*, 53(8):1235 – 1248, 2009.

[18] G. Horváth. Efficient analysis of the queue length moments of the MMAP/MAP/1 preemptive priority queue. *Performance Evaluation*, 69(12):684–700, Elsevier, 2012.

[19] G. Horváth and Hiroyuki Okamura. A fast EM algorithm for fitting marked Markovian arrival processes with a new special structure. In *EPEW*, *Lecture Notes in Computer Science*, 8168:119–133. Springer, 2013.

[20] G. Horváth and M. Telek. On the canonical representation of phase type distributions. *Performance Evaluation*, 66(8):396–409, Elsevier, 2009.

[21] A. Klemm, C. Lindemann, and M. Lohmann. Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation*, 54(2):149 – 173, Elsevier, 2003.

[22] H. Li, M. Muskulus, and L. Wolters. Modeling job arrivals in a data-intensive grid. In *Proceedings of JSSPP*, pages 210–231, 2006.

[23] M.F. Neuts. A Versatile Markovian Point Process. *Journal of Applied Probability*, 16(4):764–779, 1979.

[24] J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[25] Coin-or branch and cut project: `https://projects.coin-or.org/Cbc`.

[26] H. Okamura, T. Dohi, and K. S. Trivedi. Markovian arrival process parameter estimation with group data. *IEEE/ACM Transactions on Networking*, 17(4):1326–1339, August 2009.

[27] J. F. Pérez, J. Van Velthoven, and B. Van Houdt. Q-MAM: a tool for solving infinite queues using matrix-analytic methods. In *Proc. of ValueTools*, pages 16:1–16:9, ICST, 2008.

[28] D. Perez-Palacin, J. Merseguer, and R. Mirandola. Analysis of bursty workload-aware self-adaptive systems. In *Proc. of ICPE*, pages 75–84, ACM, 2012.

[29] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, 4(6):833–846, 1986.

[30] B. Van Houdt. Analysis of the adaptive MMAP[K]/PH[K]/1 queue: A multi-type queue with adaptive arrivals and general impatience. *European Journal of Operations Research*, 220(3):695–704, 2012.

[31] A. Verma and A. Anand. General store placement for response time minimization in parallel disks. *Journal of Parallel and Distributed Computing*, 67(12):1286–1300, Elsevier, December 2007.