On capturing dependence in point processes: Matching moments and other techniques

Barry L. Nelson
Northwestern University
nelsonb@northwestern.edu

Ira Gerhardt
Manhattan College
ira.gerhardt@manhattan.edu

January 13, 2010

Abstract

Providing probabilistic analysis of queueing models can be difficult when the input distributions are non-Markovian. In response, a plethora of methods have been developed to approximate a general renewal process by a process with the time between renewals being distributed as a phase type random variable, which allows the resulting queueing models to become analytically or numerically tractable. However, from previous studies on the manufacturing sector, and more recently in analysis of telecommunications systems, assumptions of independence do not always hold and efforts have been made to approximate nonrenewal processes with Markovian Arrival Processes. In this paper we survey techniques for deriving the appropriate parameters of a Markovian process to accurately capture relevant characteristics of the original point process.

Keywords: Markovian arrival process, phase type distribution, Markov-modulated Poisson process, dependence, moment-matching, maximum-likelihood estimation, time-series analysis, parameter estimation.

1 Introduction

Providing analytical results for specific real-world queueing models is made more difficult if characteristics of the input processes—such as interarrival and service times—do not correspond to the i.i.d. exponential random variables that are the building blocks of queueing theory. For example, studies of internet protocol (IP) traffic have shown that the times between connection attempts typically are not mutually independent, while the resulting counting processes are frequently more variable than Poisson, with connection attempts occurring in bursts (e.g., see [35, 42, 92]). This may lead to models that are computationally and analytically intractable. The task then for the engineer intending to calculate relevant performance measures or predict future queueing behavior begins with fitting models to these processes that allow for tractability.

In response, much queueing literature over the last 40 years has been devoted to developing and describing techniques for fitting processes with the Markov property to arbitrary point processes. Notice the term "fitting" is somewhat misleading, as it is often impossible to perfectly match the cumulative distribution function (cdf) or probability density function (pdf) along its entire support as well as a complete set of dependence measures. Rather, these fitting techniques frequently target a subset of properties of the original process (such as marginal moments, shape characteristics, or measures of autocovariance) or estimate parameters for the fitted process from empirical samples of the original process. Although not always guaranteed to obtain accurate predictions of queueing behavior (e.g., see [3]), these matching techniques typically yield analytically tractable queueing models.

The majority of this literature has focused on approximating point processes with the *versatile Markovian point process*, first described by Neuts [85], which is a generalized class of processes with interevent times characterized as the time to absorption of a finite-state continuous-time Markov chain (CTMC). Two subclasses of this process are particularly prevalent in the fitting literature: Markovian Arrival Processes (MAPs) and phase-type (Ph) renewal processes. Reasons for selecting Ph processes or MAPs as fitting tools are detailed below.

In this paper we survey some of the extensive literature devoted to fitting Markovian point processes, with a focus on those techniques that aim to capture some measure of dependence. The remainder of the paper is organized as follows: First we introduce relevant notation and describe classes of Markovian processes that are the tools of the fitting techniques we survey (Section 2). In Section 3 we briefly review work on approximating a general renewal process with a Ph renewal process, both in terms of techniques and developed technology. In Section 4 we provide a discussion of efforts to capture properties of general nonrenewal processes with MAPs. We also briefly review efforts to fit MAPs to data and cite examples of the use of maximum-likelihood methods to estimate MAP parameters. We conclude with Section 5 where we discuss future directions for this research area.

2 Relevant Terminology

2.1 General Notation for Point Processes

We begin with a set of nonnegative identically distributed interevent times $\{X_n,\, n \geq 1\}$, such that X_1 is from cumulative distribution function G (i.e., $G(t) = \Pr\{X_1 \leq t\}$, for $t \geq 0$). We let S_n denote the time of the n^{th} event; that is, $S_0 = 0$ and $S_n = \sum_{i=1}^{n} X_i$, for $n = 1, 2, \ldots$. We assume that $\{X_n,\, n \geq 1\}$ is stationary; that is, the joint distribution of $(X_{n_1+m}, X_{n_2+m}, \ldots, X_{n_k+m})$ is independent of m for all $k \geq 1$, $\{n_1, n_2, \ldots, n_k\} \in (\mathbb{Z}^+)^k$ [67]. We further assume that $\lim_{\delta \downarrow 0} G(\delta) = 0$.

For $i = 1, 2, \ldots$, we define $m_i \equiv \mathbb{E}\{X_1^i\}$ and $m_i' \equiv \mathbb{E}\{(X_1 - m_1)^i\}$; we say m_i is the i^{th} ordinary moment of X_1, while m_i' is its i^{th} centralized moment. We further define μ_2, such that $(\mu_2)^2 \equiv m_2'/m_1^2$, and $\mu_i \equiv m_i'/(m_2')^{i/2}$ for $i = 3, 4, \ldots$; we say μ_i is the i^{th} standardized moment of X_1, for $i = 2, 3, \ldots$. The second standardized moment μ_2 is worth further discussion; it is commonly known as the coefficient of variation, or *cv*. The squared coefficient of variation, or *scv* $(= \mu_2^2)$, may also be useful. Notice that throughout this paper we refer to *cv* and *scv* rather than μ_2 and μ_2^2, respectively.

Many papers cited here describe a *moment-matching technique*. For shorthand we let the vector $\mathbf{m}_n$ denote the first n noncentral moments of X_1, and let vector $\boldsymbol{\mu}_n$ denote its first n standardized moments (by convention, $\mu_1 = m_1$). Notice that we can compute $\boldsymbol{\mu}_n$ from $\mathbf{m}_n$ and vice versa.

We let ρ_k denote the lag-k interevent time autocorrelation; that is, $\rho_k \equiv \mathrm{Corr}\{X_1, X_{1+k}\}$ $= \mathrm{Cov}\{X_1, X_{1+k}\}/m_2'$, for $k = 1, 2, \ldots$. A useful tool is the Index of Dispersion for Intervals (IDI), defined as $c_n^2 = \mathrm{Var}\{S_n\}/(nm_1^2)$ [111]; c_n^2 is also referred to as the n-interval *scv* sequence. Several papers cited here utilize $c_\infty^2 \equiv \lim_{n\to\infty} c_n^2$; it can be shown that

$$c_\infty^2 = scv\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right). \tag{1}$$

When $\{X_n,\, n \geq 1\}$ are independent as well as identically distributed (i.i.d.), then $\rho_k = 0$ for all $k \geq 1$, and $c_n^2 = scv$ for all $n \geq 1$ (including $n = \infty$).

We have now described the interval process, consisting of interevent times $\{X_n,\, n \geq 1\}$ (with first n marginal moments $\mathbf{m}_n$) and autocorrelation structure $\{\rho_k,\, k \geq 1\}$. For the purpose of this paper, we define an event as an arrival of entities in a batch of (random) size ℓ, for $\ell \in \mathbb{Z}^+$. Thus, we define the counting process $N(t)$ which describes the number of entities that have arrived at or before time $t \geq 0$.

Analogous to the IDI is the Index of Dispersion for Counts (IDC) at time t, defined as $I(t) = \text{Var}\{N(t)\}/\mathbb{E}\{N(t)\}$ [39]. The IDC curve, $\{I(t), t \geq 0\}$, may also be referred to as the *variance-time* curve. The limiting value of the IDC curve, $I_\infty = \lim_{t\to\infty} I(t)$, appears in several of the papers we cite here.

2.2 BMAPs, MAPs, and Ph Renewal Processes

The most general Markovian process cited in this survey is the Batch Markovian Arrival Process (BMAP) [75], which is equivalent to the versatile Markovian process first investigated by Neuts [85], referred to elsewhere (in tribute to Neuts) as the N-Process [94]. The interevent times in a BMAP describe the time it takes an underlying CTMC to reach $m_C \geq 1$ absorbing phases from a finite number $m_T < \infty$ of transient phases; the chain reaching an absorbing phase triggers an arrival of random size $\ell \in \{1, 2, \ldots, M\}$, where M may be infinity. Let $J(t)$ denote the current phase of the CTMC at time t. We utilize the shorthand BMAP(m_T) to describe a BMAP of order m_T, meaning that the underlying CTMC for the BMAP has m_T transient phases.

We utilize a representation here for the BMAP(m_T) that characterizes the interevent distribution by transitions within the embedded discrete-time Markov chain (DTMC) along with a vector of transition rates (one for each transient phase) and a matrix of the initial transient phase probabilities. This representation is used by Nelson and Taaffe [84] and recounted here.

We let $\mathbf{A}$ denote the one-step transition probability matrix of the embedded DTMC:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \boldsymbol{\alpha} & \mathbf{0} \end{pmatrix}.$$

The $m_T \times m_T$ matrix $\mathbf{A}_1$ represents the one-step transition probabilities between the m_T transient phases, while the $m_T \times m_C$ matrix $\mathbf{A}_2$ represents the one-step transition probabilities from the m_T transient phases to the m_C absorbing phases. "Absorbing phase" is really a misnomer in this representation, because rather than being absorbed the process is reinitialized for the next interevent time by $m_C \times m_T$ initial probability matrix $\boldsymbol{\alpha}$. By convention we assume self-transitions in the embedded DTMC are not permitted (i.e., $(\mathbf{A}_1)_{jj} = 0$, for all $j = 1, 2, \ldots, m_T$).

We define the $m_T \times 1$ vector $\boldsymbol{\upsilon}$, whose j^{th} argument is υ_j, the non-negative rate corresponding to phase j, for $j = 1, 2, \ldots, m_T$. We use the convention $\upsilon_{m_T+k} = \infty$, for $k = 1, 2, \ldots, m_C$, corresponding to an instantaneous sojourn time in any absorbing phase. Thus, the Nelson and Taaffe BMAP representation is the pair $(\mathbf{A},\boldsymbol{\upsilon})$.

The key to the Nelson and Taaffe BMAP representation is that we construct matrices $\mathbf{A}_2$ and $\boldsymbol{\alpha}$ such that there is a unique absorbing phase for each pair (j, ℓ) of transient phase $j = 1, 2, \ldots, m_T$ and batch size $\ell = 1, 2, \ldots, M$; thus, $m_C = Mm_T$. To do this, we construct $\mathbf{A}_2$ as the concatenation of M diagonal matrices, each $m_T \times m_T$; that is, we specify that the DTMC cannot transition in one-step from transient phase j to an absorbing state with label (h, ℓ), for $h \neq j \in \{1, 2, \ldots, m_T\}$.

It is worth mentioning that with matrices $\mathbf{A}_2$ and $\boldsymbol{\alpha}$ constructed as such, we can connect the Nelson and Taaffe BMAP representation to a related representation from Lucantoni [75]. The Lucantoni BMAP representation is the set of $m_T \times m_T$ matrices $\{\mathbf{D}_\ell,\ \ell = 0, 1, \ldots, M\}$, such that $(\mathbf{D}_\ell)_{jh}$ is the transition rate from transient phase j to transient phase h upon an arrival of size ℓ, for $\ell \geq 0$. We can construct the Lucantoni representation from the Nelson and Taaffe representation ($\mathbf{A}$,$\boldsymbol{\upsilon}$):

$$\mathbf{D}_0 = \mathbf{U}(\mathbf{A}_1 - \mathbf{I}), \tag{2}$$

where $\mathbf{U}$ is a diagonal matrix with nonzero elements υ_j, for $j = 1, 2, \ldots, m_T$, and $\mathbf{I}$ is the identity matrix, while

$$(\mathbf{D}_\ell)_{jh} = \upsilon_j \cdot (\mathbf{A}_2)_{j,(\ell-1)m_T+j} \cdot (\boldsymbol{\alpha})_{(\ell-1)m_T+j,h}, \tag{3}$$

for $j, h = 1, 2, \ldots, m_T$ and $\ell = 1, 2, \ldots, M$.

Notice the Lucantoni representation explicitly describes the stochastic process $\{(N(t), J(t)),\ t \geq 0\}$, which has infinite state space, while the Nelson and Taaffe representation describes interevent times, characterized by transitions on the embedded DTMC, whose (typically finite) space consists of m_T transient phases and Mm_T absorbing phases. The papers cited in this survey typically approximate properties of the interval process, not the counting process, which is why we employ the Nelson and Taaffe representation.

For simplicity, we refer to this representation as *the BMAP representation* for the remainder of this paper without further attribution. We provide the BMAP representation ($\mathbf{A}$,$\boldsymbol{\upsilon}$) for several example BMAPs; readers interested in translating from the BMAP representation to the Lucantoni representation can do so using (2) and (3).

A MAP(m_T) is a special case of BMAP(m_T) where $M = 1$. For a stationary MAP(m_T) (as we examine here), we utilize $\boldsymbol{\beta}$, the steady-state $m_T \times 1$ vector for the embedded DTMC at arrival instants; it is the solution to

$$\boldsymbol{\beta}^\top[(\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{A}_2\boldsymbol{\alpha}] = \boldsymbol{\beta}^\top, \;\; \boldsymbol{\beta}^\top\mathbf{e} = 1,$$

where $\mathbf{e}$ is a $m_T \times 1$ vector with all coordinates equal to 1. Then

$$G(t) = 1 - \boldsymbol{\beta}^\top \exp\{\mathbf{U}(\mathbf{A}_1 - \mathbf{I})t\}\mathbf{e},$$

and

$$m_i = i!\boldsymbol{\beta}^\top \left[\mathbf{U}(\mathbf{I} - \mathbf{A}_1)\right]^{-i} \mathbf{e}, \tag{4}$$

for $i = 1, 2, \ldots$ [67]. Further, it can be shown that

$$\rho_k = \frac{\boldsymbol{\beta}^\top \left[\mathbf{U}(\mathbf{A}_1 - \mathbf{I})\right]^{-1} (\mathbf{I} - \mathbf{e}\boldsymbol{\beta}^\top) \left[(\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{A}_2\boldsymbol{\alpha}\right]^k \left[\mathbf{U}(\mathbf{A}_1 - \mathbf{I})\right]^{-1} \mathbf{e}}{\boldsymbol{\beta}^\top \left[\mathbf{U}(\mathbf{A}_1 - \mathbf{I})\right]^{-1} (2\mathbf{I} - \mathbf{e}\boldsymbol{\beta}^\top) \left[\mathbf{U}(\mathbf{A}_1 - \mathbf{I})\right]^{-1} \mathbf{e}}, \tag{5}$$

for $k = 1, 2, \ldots$ [30]. Notice for a MAP(m_T), the matrix $\mathbf{A}_2$ is diagonal; in fact,

$$(\mathbf{A}_2)_{jh} = \begin{cases} 1 - \sum_{r=1}^{m_T} (\mathbf{A}_1)_{jr}, & \text{if } h = j, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

for j, $h = 1, 2, \ldots, m_T$. Therefore, to characterize a MAP, we need only specify the probability matrices $\mathbf{A}_1$ and $\boldsymbol{\alpha}$ and rate vector $\boldsymbol{\upsilon}$; the matrix $\mathbf{A}_2$ is defined completely by the matrix $\mathbf{A}_1$, as in (6). The BMAP representation of the MAP(m_T) has $m_T(2m_T - 1)$ free parameters; we discuss the possible over-parameterization of MAPs later in this paper.

A Ph renewal process is a special case of MAP where the $\{X_n,\, n \geq 1\}$ are i.i.d; therefore, $\rho_k = 0$ in (5), for all $k = 1, 2, \ldots$. For this to hold, all m_T rows in the initial probability matrix $\boldsymbol{\alpha}$ must equal $\boldsymbol{\beta}^\top$. Thus, for a Ph renewal process, the initial transient phase visited by the CTMC immediately after an absorbing phase is independent of the absorbing phase index.

A renewal process is completely defined by its interrenewal distribution; therefore, we describe a Ph renewal process in terms of its Ph interrenewal distribution. Various Ph distributions are utilized in the papers we cite here; we specify the matrix $\mathbf{A}_1$, rate vector $\boldsymbol{\upsilon}$, and steady-state initial probability vector $\boldsymbol{\beta}$ for their corresponding Ph renewal processes here:

- Coxian (C_{m_T}): Define the set $\{p_1, p_2, \ldots, p_{m_T - 1}\} \in [0, 1]^{m_T - 1}$. If λ_j^{-1} is the mean sojourn time the underlying CTMC spends in phase j (with $\lambda_j > 0$), for $j = 1, 2, \ldots, m_T$, then the BMAP representation of the Coxian renewal process (generated by a Coxian interrenewal distribution) is

  $$\upsilon_j = \lambda_j, \;(\mathbf{A}_1)_{jh} = \begin{cases} p_j, & \text{if } h = j + 1, \\ 0, & \text{otherwise}, \end{cases} \quad \beta_j = \begin{cases} 1, & \text{if } j = 1, \\ 0, & \text{otherwise}, \end{cases}$$

  for j, $h = 1, 2, \ldots, m_T$, where β_j is the j^{th} component of vector $\boldsymbol{\beta}$. Several cases of Coxian distributions are worth calling out:

  - The Generalized Erlang distribution ($GE_{m_T}(\lambda)$) is a special case of a Coxian distribution where $p_j = 1$ for $j = 2, 3, \ldots, m_T - 1$ (but $p_1 \in [0, 1]$), while $\lambda_j = \lambda$ (with constant $\lambda > 0$) for all $j = 1, 2, \ldots, m_T$.
  - The Erlang distribution ($E_{m_T}(\lambda)$) is a special case of a Generalized Erlang distribution where $p_1 = 1$.
  - The exponential distribution ($E_1(\lambda)$) is a special case of an Erlang distribution where $m_T = 1$. A renewal process generated by an exponential interrenewal distribution is Poisson.

- Hyperexponential (H_{m_T}): Define the set $\{p_1, p_2, \ldots, p_{m_T}\} \in [0, 1]^{m_T}$, such that $\sum_{j=1}^{m_T} p_j = 1$. If λ_j^{-1} is the mean sojourn time the underlying CTMC spends in phase j (with $\lambda_j > 0$), then the BMAP representation of the hyperexponential renewal process (generated by a hyperexponential interrenewal distribution) has $\mathbf{A}_1 = \mathbf{0}$, while $v_j = \lambda_j$ and $\beta_j = p_j$, for $j = 1, 2, \ldots, m_T$.

We frequently use the Ph renewal process' shorthand to describe a random variable from the Ph interrenewal distribution.

3 Renewal Processes: Fitting Ph Interrenewal Distributions

Phase-type, or Ph, distributions are attributed to Neuts [86] and are frequently used in fitting renewal processes, for two reasons. First, the Markovian properties of Ph distributions make the resulting queueing models more analytically tractable [77]. Second, Ph distributions are dense on the set of all distributions with support on $[0, \infty)$ [5].

The question then arises: how do we *approximate* a general renewal process by one with times between renewals governed by a Ph distribution? What properties of the original process can we capture? Which properties are important to replicate to properly represent the original process? An expansive literature has been created to answer these questions;

most papers specify a small but flexible family of Ph distributions, setting values for its BMAP parameters to satisfy (4) for $i = 1$ and $i = 2$ (and possibly, $i = 3$). Although the emphasis of our paper is nonrenewal MAPs, in this section we provide a brief overview of Ph-fitting literature as well as a description of some of the software that has been developed to fit Ph distributions.

3.1 Modeling Techniques

Early work on fitting Ph renewal processes targets the first two moments of the original interval process (i.e., $\mathbf{m}_2$). Using the notion that the mean of a Ph distribution acts as a scaling factor, these papers focus on developing methods to match the *scv* of the time between renewals.

In the earliest of these papers, Sauer and Chandy [105] fit non-exponential service processes with $scv > 1$ to H_2's and processes with $scv < 1$ to $GE_{m_T}(\lambda)$'s. Similarly, Marie [78] fits service processes with $scv > 0.5$ to C_2's and $scv = 0.5$ to $E_2(\lambda)$'s. While noting that an $E_{m_T}(\lambda)$ has $scv = 1/m_T$, he conjectures that $E_k(\lambda)$ distributions might be viable to fit intervals with $scv = 1/k + \epsilon$, for ϵ small and $k = 3, 4, \ldots$. Bux and Herzog [23] develop a nonlinear technique that targets a sample $\mathbf{m}_2$ while minimizing a measure of difference from the empirical cdf. Whitt [119] also develops a two-moment technique, establishing parameters in H_2, $GE_2(\lambda)$, and a shifted exponential distribution (i.e., an $E_1(\lambda)$ shifted by a constant value) to approximate an arrival process in an effort to assess the effect (on congestion in the system) of changing the service parameters. Tijms [114] cites a two-moment technique mixing a pair of Erlang distributions of consecutive orders for $scv < 1$; Weerstra [117] describes a similar technique utilizing an adjusted Erlang, with different means for the last two phases than the common mean for the earlier phases in the chain.

Altiok [2] moves beyond the two-moment approach, citing Whitt [122] on the importance of shape considerations in approximating arrival processes. Altiok derives formulas for matching a C_2 to $\boldsymbol{\mu}_3$ for a given point process with $scv > 1$, and identifies necessary and sufficient conditions for the fitted parameters of the C_2 to specify a legitimate distribution. Whitt [120] also develops a three-moment matching technique to fit point processes with $scv > 1$ to H_2's, comparing the quality of matching the point process over a short interval (referred to as the "stationary-interval method," originally attributed to Kuehn [66]) versus matching the behavior over a relatively long time interval (the "asymptotic method").

Additional three-moment techniques using Ph subclasses are developed by Johnson and Taaffe [59], who identify the feasible set of $\boldsymbol{\mu}_3$ that can be matched with a mixture of two Erlangs of common order (MECO-2). In this paper they derive formulas for the mixing probability p and respective rates λ_1, λ_2 for the E_{m_T}'s in the MECO-2 (for feasible order m_T) to match $\boldsymbol{\mu}_3$. Johnson and Taaffe expand on this method, using a nonlinear technique to fit Coxians and mixtures of Erlangs possibly not of common order [61], and investigate the effect of these techniques on the shapes of the density functions they attain [60]. Later they compare their MECO method to a two-moment method that uses H_2 distributions with balanced means [62].

More recently, Osogami and Harchol-Balter [91] use a sewing technique with Erlangs and Coxians to match $\mathbf{m}_3$ for a general distribution with a minimal order Ph distribution. Noting that the Erlang is the least variable of the Ph distributions [1], the authors provide necessary and sufficient conditions for matching $\mathbf{m}_3$ with Coxian distributions [90].

Bobbio and Telek [18] survey methods for fitting an Acyclic Ph distribution of order m_T (APH_{m_T}) to a set of benchmark distributions. A Ph distribution is *acyclic* if there exists an ordering of the transient phases such that $\mathbf{A}_1$ under that ordering is upper-triangular. They cite a previous Bobbio paper [14] on using maximum likelihood (ML) methods to estimate the parameters of the canonical representation of a fitted APH distribution. Bobbio et al. [15, 16, 17] develop techniques for fitting the parameters of discrete and continuous APH_{m_T} distributions to $\boldsymbol{\mu}_3$ of general distributions, while Telek and Heindl [112] focus on fitting APH_2.

In a paper on general continuous distributions, van de Liefvoort [115] provides an algorithm to specify the rational Laplace-Stieltjes transform (LST) (with maximum degree n) of a distribution from moments $\mathbf{m}_{2n-1}$. Those distributions with rational LST are known as the Matrix Exponential (ME) distributions. Ph distributions are a subset of the ME distributions.

One limitation of the rational LST technique is that it impossible to know if the set of moments correspond to a feasible ME distribution until its corresponding density is computed. Horváth and Telek [53] build on van de Liefvoort's result [115] and utilize APH_{m_T} in an attempt to overcome this limitation and target more than three moments. They propose a one-phase reduction technique, where at each step the APH_k (for $k \leq m_T$) is replaced by an APH_{k-1} possibly superposed with an $E_1(\lambda)$.

Other fitting-related work focuses on general distributions with heavy tails (i.e., distributions whose tails decay slower than exponentially). Feldman and Whitt [32] develop a technique for matching H_{m_T} distributions to heavy-tailed distributions with completely monotone density functions (such as certain Weibull and Pareto distributions); for a survey of heavy-tailed related literature, see [32]. Notice that, to date, most heavy-tailed fitting techniques are minor adaptations of the Feldman and Whitt method. Horváth and Telek [51] study the quality of several of these approaches.

A number of papers are devoted to using ML methods and the expectation-maximization (EM) algorithm to estimate parameters of Ph distributions from data. A key benefit of the EM algorithm is that it works when data are incomplete or there are missing values; for background on the EM algorithm, see [28, 123]. Asmussen et al. [8] use the EM algorithm to estimate parameters for a general Ph distribution and later for a mixture of $E_{m_T}(\lambda)$ distributions [6]. Thümmler et al. [113] also utilize mixtures of $E_{m_T}(\lambda)$ distributions to fit real and simulated Internet trace data, while El Abdouni Khayari et al. [64] use the EM algorithm to fit real trace data with hyperexponentials. Fackrell [31] develops an ML technique for determining when the fitted parameters in a rational LST correspond to a legitimate ME distribution. Riska et al. [95] use the EM algorithm to fit mixtures of Ph distributions when the histogram of the data indicates long tails.

3.2 Computer Software for Fitting Ph-Renewal Processes

Several of the papers described in Section 3.1 have been complemented with computer software. Johnson's [57] and Schmickler's [106] work on using mixtures of E_{m_T} distributions to target $\boldsymbol{\mu}_3$ has led to MEFIT and MEDA, respectively. EMPHT [89] (and its successor, EMpht) employs the EM algorithm in estimating parameters of a general Ph distribution, fitting the Ph either to data or to one of a predefined set of distributions. MLAPH [14], as per its name, uses ML techniques to fit parameters in the canonical form of an APH distribution, while PHFit [52] separates fitting techniques for the body and tail of the target distribution, using APH distributions for the body and the method of Feldman and Whitt [32] for the tail. Recently, Pérez and Riaño [93] present jPhase, with component jPhaseFit that utilizes both ML techniques for fitting Ph distributions to data and APH distributions for matching moments. For discussion on the comparative quality of several of these applications, see [69].

3.3 Evaluation of Fitting with Ph Renewal Processes

In this section we have (primarily) reviewed techniques to match the first two or three marginal moments of renewal point processes using specific families of Ph renewal processes. Based on our survey, we feel that efforts to capture these characteristics have been successful, and given values for $\mathbf{m}_3$ (or equivalently $\boldsymbol{\mu}_3$), there exist several techniques that will specify a Ph renewal process that sufficiently approximates the original process; we recommend the MECO-2 from Johnson and Taaffe and the APH techniques from Bobbio et al. because they can match any feasible triple of first three interval moments using simple formulas.

4 Non-Renewal Processes: Fitting MAPs

Real-world studies of systems in manufacturing and telecommunication networks have brought to light that standard assumptions regarding independence of interarrival times actually may be inappropriate. Therefore, more realistic models need to involve processes with nonnegligible dependence structures (i.e., nonzero autocovariance and autocorrelation) as well as non-exponentially distributed interarrival times [7].

In this section we review efforts to fit nonrenewal processes with MAPs. We first discuss techniques to capture dependence with general MAPs, following that with a discussion on the use of BMAPs and Markov-modulated Poisson processes (MMPPs). Although our focus is fitting properties (such as moments and covariance measures), we briefly cite papers that employ algorithms to estimate parameters from data. Some analytical models that result in MAP departure processes are also briefly reviewed, and the section concludes with our recommendations from amongst the cited fitting techniques.

4.1 General MAPs

Most general MAP-fitting methods involve taking superpositions and mixtures of the fundamental building blocks (i.e., exponential distributions), but in such a way as to capture dependence within the model.

Several papers cite techniques for specifying parameters of a MAP(2) to accomplish this. The BMAP representation for the MAP(2) is

$$\boldsymbol{\upsilon} = (\upsilon_1, \upsilon_2)^\top, \ \mathbf{A}_1 = \begin{pmatrix} 0 & a_1 \\ a_2 & 0 \end{pmatrix}, \text{ and } \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_2 & \alpha_2 \end{pmatrix},$$

with probabilities $\{a_1, a_2, \alpha_1, \alpha_2\} \in [0,1]^4$, and rates $\upsilon_1, \upsilon_2 \geq 0$. Thus, the MAP(2) is characterized by six free parameters.

We can use (5) to show that the autocorrelation sequence $\{\rho_k,\ k \geq 1\}$ for the MAP(2) is geometric; that is, $\rho_k = c_\rho \xi^k$, for $k \geq 1$, where both the parameter ξ and coefficient c_ρ are functions of the MAP(2) parameters (presented in Appendix A). The parameter ξ is utilized in both MAP(2)-fitting techniques described below.

Diamond and Alfa [30] provide the most general fitting technique for the MAP(2), extending Altiok [2] and Whitt [120] in matching $\mathbf{m}_3$ to also target ρ_1 for a nonrenewal interval process. The authors provide feasibility conditions on the MAP(2) parameters to achieve particular values for ρ_1 (in terms of the parameter ξ); these conditions generally include restrictions on the feasible *scv* of the marginal distribution that can be achieved. They provide algorithms for specifying the BMAP representation when the feasibility conditions are met.

To validate their technique, the authors model the departure process from a queue and then examine the moments of the resulting queue length when that departure process serves as the arrival stream to another queue. Their method leads to accurate approximations for the first three moments of the queue length when there are no restrictions on ξ and *scv*. However, if $scv < 1$ and $\xi > 0$, the minimum achievable ρ_1 is -0.037. Also, they conclude that the MAP approximation for the model is only a slight improvement over the renewal approximation (i.e., when $\alpha_2 = 1 - \alpha_1$). They hypothesize that using MAPs of larger order will allow them to target more significant levels of dependence.

Special cases of the MAP(2) are worth citing; they result when specific values are selected for the probability parameters a_1, a_2, α_1, and α_2. One such case is the MMPP(2); it is specified by $\alpha_1 = \alpha_2 = 1$. We discuss the MMPP(2) in Section 4.2. When either $a_1 = 0$ or $a_2 = 0$ (but not both), the marginal distribution of the MAP(2) is APH_2, and the resulting process is referred to as an AMAP(2).

Recently, Heindl et al. [45] utilize AMAP(2)'s to provide matching techniques for both hyperexponential (i.e., $scv > 1$) and hypoexponential (i.e., $scv < 1$) marginals, improving on an earlier Heindl result [44] where only H_2 marginals could be specified.

An important difference between the Diamond and Alfa technique and the Heindl et al. technique is that the representation in the latter also involves a free parameter $\eta \in [0, 1]$, selected by the modeler; the range of feasible ξ that can be achieved is then dependent on both the choice of η and the *scv* for the marginal distribution. Heindl et al. define feasible bounds for ξ in both the hyperexponential and hypoexponential domains, noting that, although the former domain is more flexible, in neither can the full range of ρ_1 be achieved (limitations are most apparent when the target $scv < 1$ and $\rho_1 < 0$). For reference, the BMAP representation of Heindl et al.'s AMAP(2) technique is provided in Appendix B.

A related two-step EM algorithm for first specifying the marginal distribution and then ρ_1 while fitting MAP(2)'s is described in [55]; the algorithm utilizes nonlinear optimization to specify α_1 and α_2, and its success is heavily dependent on the choice of initial values. The technique in [45] also extends earlier Heindl et al. papers [46, 47] that utilize Marie's technique [78] when $scv > 0.5$. The authors' goal is to assess the quality of the fitting technique for use in network decomposition, noting that the decomposition may be sensitive to m_3 and ξ and, thus, the two-moment fitting technique (for renewal processes) first utilized in Whitt's Queueing Network Analyzer (QNA) [121] may be insufficient.

Also in the area of network decomposition, Mitchell and van de Liefvoort [82] use sequences of correlated ME(2) distributions (with invariant marginals) in targeting both marginal and dependence moments of the departure process from a $G/G/1/N$ queue. The idea of using correlated ME distributions is developed by Mitchell [80] and extends an earlier paper [81] that investigates matching only marginal information.

Casale et al. [25] utilize Kronecker products (rather than sums) in the superposition of MAP(2)'s within a network traffic model. They provide theorems connecting the moments of the marginal distribution with the eigenvalues of $[\mathbf{U}(\mathbf{A}_1 - \mathbf{I})]^{-1}$ for the superposed process. By requiring $\mathbf{A}_1 = \mathbf{0}$ for all but one of the component processes, the authors claim they can target both hyperexponential and hypoexponential distributions. The focus of their efforts is fitting trace data; the KPCToolbox [24]—a package of Matlab scripts—has been designed to this end.

Another technique for modeling network flow comes from Bitran and Dasu [12]; the authors develop Super-Erlang (SE) chains, which they consider to be nonrenewal analogs of Erlang chains. Effectively, they start with $E_{m_T}(\lambda)$ and expand each phase j (for $j = 1, 2, \ldots, m_T$) to include several subphases (each labeled by the phase level j and a subphase index). One-step transitions in the SE chain are labeled as either unmarked or marked: unmarked transitions move the chain forward one phase level (i.e., j to $j+1$), while marked transitions move the chain backwards (i.e., j to h, where $h \leq j$). Notice that for the SE chain, $N(t)$ counts the number of marked transitions by time $t \geq 0$, and G is the distribution of times between marked transitions. The fitting technique involves targeting m_1 and c_∞^2 of the marked process and then setting the remaining SE chain parameters to match *scv*.

The authors validate their model by investigating performance measures at a queue (such as the queue length distribution and *scv* of the departure process) whose arrival stream is the superposition of renewal processes. The method approximates the superposition of low variable (i.e., $scv < 1$) renewal processes well, but cannot be utilized if any component renewal process has $scv > 1$. Further, the fitting method itself is highly complicated, with a recursive numerical procedure at its center.

In another paper that utilizes Erlang distributions, Johnson [58] extends the earlier Johnson and Taaffe work on MECO-2's [59] to create the Markov-MECO. Letting $E_n(\lambda_1)$, $E_n(\lambda_2)$ denote the two Erlang distributions (of feasible order n) in the MECO-2 marginal distribution (where the mixing probability p is assigned to $E_n(\lambda_1)$), the author introduces dependence parameters $p_{im} \equiv \Pr\{X_2 \sim E_n(\lambda_m) \,|\, X_1 \sim E_n(\lambda_i)\}$, for $i,\, m = 1, 2$. This explains the "Markov" in Markov-MECO: which Erlang the current interarrival time is from is only dependent on which Erlang generated the previous interarrival time. Notice $m_T = 2n$ since the chain can sojourn in any of n phases in either Erlang; without loss of generality, we let phases $\{1, 2, \ldots, n\}$ correspond to $E_n(\lambda_1)$ and phases $\{n+1, n+2, \ldots, 2n\}$ correspond to $E_n(\lambda_2)$. Then the BMAP representation for the Markov-MECO is

$$v_j = \begin{cases} \lambda_1, & \text{if } j \leq n, \\ \lambda_2, & \text{if } j \geq n+1, \end{cases} \quad (\mathbf{A}_1)_{jh} = \begin{cases} 1, & \text{if } h = j+1,\, j < n, \\ 1, & \text{if } h = j+1,\, j \geq n+1, \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{and } (\boldsymbol{\alpha})_{jh} = \begin{cases} 1 - p_{12}, & \text{if } (j,h) = (n,1), \\ p_{12}, & \text{if } (j,h) = (n, n+1), \\ p_{21}, & \text{if } (j,h) = (2n, 1), \\ 1 - p_{21}, & \text{if } (j,h) = (2n, n+1), \\ 0, & \text{otherwise,} \end{cases}$$

for $j, h = 1, 2, \ldots, 2n$. For the Markov-MECO to have MECO-2 marginals, the relationship $p_{12} = p_{21}(1-p)/p$ must hold. Thus, adding the Markovian structure to the model entails the addition of a single free parameter, p_{21}. Johnson further shows ρ_1 can be expressed as a

1-to-1 function of p_{21}, thus specifying the value of p_{21} that yields a given value for ρ_1.

However, two limitations arise for the Johnson model. First, the autocovariance function decays geometrically (with rate $1 - p_{21}/p$). Plugging this into (1) we find

$$c_\infty^2 = scv\left(1 + \frac{2p}{p_{21}}\rho_1\right).$$

Therefore, targeting a specific value of either ρ_1 or c_∞^2 specifies the value of the other; thus, only one can be matched by the transition parameter p_{21}. The second limitation is that not all values of ρ_1 can be matched. The author shows that $p_{21} \in [0, \min\{1, p/(1-p)\}]$, and that as p_{21} approaches the upper limit of this range, both ρ_1 and c_∞^2 approach finite lower limits. She suggests that this limitation can be overcome by increasing the value of the common order n, and thus the full range of ρ_1 can be matched. However, no proof of this conjecture is offered.

4.2 Markov-Modulated Poisson Processes (MMPPs)

This section provides an overview of MMPP literature, describing their use in fitting general nonrenewal processes to superpositions of renewal and nonrenewal processes, as well as the application of the EM algorithm in estimating the MMPP parameters.

The MMPP(m_T) is a special case of MAP where initial probability matrix $\boldsymbol{\alpha} = \mathbf{I}$; its BMAP representation has m_T^2 free parameters. MMPPs have become an important tool in fitting nonrenewal processes due to their analytical tractability and parsimonious representation. With the advent of the Internet and the interest in modeling Asynchronous Transfer Mode (ATM) performance, the MMPP has gained popularity due to its ability to model the correlation structure of packet streams [35]. The MMPP(2) has been the focus of the bulk of the literature.

Due to its 2-state representation, the MMPP(2) is often referred to as the Switched Poisson process (SPP). The SPP is a special case of MAP(2); its BMAP representation has four free parameters: rates v_1 and v_2 and probabilities a_1 and a_2. Notice we can connect the BMAP representation for a SPP to another frequently-cited representation in which the SPP is characterized by transition rates r_1 and r_2 and arrival rates λ_1 and λ_2 [35]: $r_j = \upsilon_j a_j$, $\lambda_j = \upsilon_j(1 - a_j)$, for $j = 1, 2$. An important case of SPP is the Interrupted Poisson Process (IPP), which results when either $a_1 = 1$ or $a_2 = 1$. The IPP is used to model ON/OFF traffic sources, as arrivals are turned "off" when the underlying CTMC for the IPP is in that phase j such that $a_j = 1$ (where $j = 1$ or $j = 2$).

Two important properties of the SPP are utilized in papers cited here. First, the superposition of a Poisson process and a SPP can be represented as a SPP. Specifically, if the Poisson process has rate υ_p, the parameters of the superposed SPP are

$$a_1^{(s)} = \frac{a_1 \upsilon_1}{\upsilon_1 + \upsilon_p},\ a_2^{(s)} = \frac{a_2 \upsilon_2}{\upsilon_2 + \upsilon_p},\ \upsilon_1^{(s)} = \upsilon_1 + \upsilon_p,\ \upsilon_2^{(s)} = \upsilon_2 + \upsilon_p,$$

where a_1, a_2, υ_1, and υ_2 are the parameters of the component SPP. Second, the superposition of z identical SPP's can be represented as a MMPP$(z + 1)$.

4.2.1 Fitting the SPP: Uses and Limitations

The SPP is a useful tool for fitting nonrenewal processes as its four parameters can be used to match four features of the original process: e.g., $\mathbf{m}_3$ and a single dependence measure. A key restriction, though, on using the SPP is that its marginal distribution has *scv* > 1, and the SPP may be a poor fit for processes with low variability (i.e., *scv* < 1). Since IP traffic is often found to be more variable than Poisson, the SPP is frequently utilized in this branch of the literature.

One form of IP traffic is the superposition of ATM packet streams. Stationary SPPs are frequently used as tools to model this traffic, with fitting techniques that specify the required parameters to target properties of superposed ATM count or interval processes. The earliest such technique is attributed to Heffes [41], who provides formulas for specifying a SPP given $\mathbf{m}_3$ and an asymptotic time constant, τ_c, analogous to c_∞^2 for the interval process. Utilizing the shorthand

$$\varphi = 1 + \frac{\mu_3}{2}\left[\mu_3 - \sqrt{4 + \mu_3^2}\right],$$

Heffes derives explicit formulas for the SPP parameters in terms of these descriptors:

$$\upsilon_1 = [\tau_c(1+\varphi)]^{-1} + m_1 + \sqrt{m_2'/\varphi},\ a_1 = \frac{[\tau_c(1+\varphi)]^{-1}}{\upsilon_1},$$

$$\upsilon_2 = \tau_c^{-1}\left[1-(1+\varphi)^{-1}\right] + m_1 - \sqrt{m_2'\varphi},\ a_2 = \frac{\tau_c^{-1}\left[1-(1+\varphi)^{-1}\right]}{\upsilon_2},$$

and investigates the quality of his fitting technique by modeling arrivals to a $SPP/M/s(/K)$ node (for both $s < \infty$ and $s = \infty$).

Several other techniques for targeting SPP properties are worth mentioning. Heffes and Lucantoni [42] examine counts of superposed ATM streams, providing formulas for SPP parameters to target two asymptotic measures (the long-run average arrival rate, equal to m_1^{-1}, and I_∞) and two time-dependent measures ($I(t_1)$ and $\mathbb{E}\{[N(t_2) - \mathbb{E}\{N(t_2)\}]^3\}$), calculated at arbitrary times t_1, $t_2 \in (0,\infty)$ selected by the modeler. Nagarajan et al. [83] use the first three Heffes and Lucantoni descriptors in their SPP fitting technique, replacing the third centralized count moment with $I(t_2)$; the selection of finite time t_2 here depends on the traffic load at that time. Gusella [39] targets $\boldsymbol{\mu}_2$, I_∞, and $I(t_1)$, such that the choice here of t_1 depends on *scv* of the targeted process. Rossiter [99] uses the same first three descriptors as Gusella, replacing time-dependent measure $I(t_1)$ with the asymptotic dependence measure $\lim_{t\to\infty} \mathrm{Cov}\{N(t), N(2t) - N(t)\}$. Ferng and Chang [33, 34] target $\mathbf{m}_3$ and ρ_1 of the stationary departure process from a $BMAP/G/1$ node as they model network flow.

Approaches for validating these fitting technique vary by author. Heffes and Lucantoni examine performance measures at a $SPP/G/1$ node (where the superposed ATM arrival process is fitted by a SPP), while Gusella compares the moments and IDC curve of the fitted SPP to those of the original process. In both techniques, accurate results are achieved, although the results are heavily dependent on the choices of the finite time values t_1, t_2. Also, Heffes and Lucantoni note that the SPP has too small an order to effectively capture long tails. Ferng and Chang examine both the fitted traffic descriptors and the expected delay at downstream nodes (versus simulation), and found the results to be generally satisfactory.

Formulas for specifying the SPP parameters in the Heffes and Lucantoni, Gusella, and Ferng and Chang techniques are found in Appendices C, D, and E, respectively. An additional contribution of the Heffes and Lucantoni paper is the set of SPP count moments as explicit functions of SPP parameters; these expressions have been utilized in several papers (e.g., see [43]).

Frequently, simple models for IP traffic arriving to a multiplexer are produced by aggregating the various levels of video and voice sources into two states based on whether the arrival load (i.e., rate) for a particular level is either greater (overloaded) or lower (underloaded) than the multiplexer's capacity. The two aggregated states are then considered the phases (of the underlying CTMC) of a SPP, and techniques are provided to specify the SPP parameters to target descriptors of the IP traffic.

Skelley et al. [110] use SPPs to model the superposition of variable bit rate (VBR) video traffic streams; their aggregation is based on a histogram representation of the bit-rates of each of the individual traffic steams. Kang et al. [63] aggregate arrival counts (during fixed time windows of length w); they claim that superposed ATM streams may have $scv < 1$, and fit this data with a MAP(3) (extending a SPP by adding an additional phase to the SPP underloaded state) to capture this. Wang et al. [116] approximate a superposed traffic stream (consisting of voice, video and data sources) to a multiplexer, modeling the video and voice sources as an aggregated SPP and the data as a batch Poisson process (with an exogenously determined packet size distribution).

Both Skelley et al. and Kang et al. examine loss probability in a finite-buffer ATM multiplexer (the former approximates it in validating their model, while the latter uses it as a target measure to fit). For a survey comparing Skelley et al. to other papers in this section, see [107]. The quality of the Kang et al. technique is highly dependent on the window length w; if w is either too small or too large, then time windows may be categorized incorrectly (e.g., as overloaded rather than underloaded). The authors here suggest extending their technique to a MAP(m_T) (for $m_T > 3$) to capture lower levels of the superposed stream's

scv.

Wang et al. model the multiplexer as a $BMMPP/D/1$ node, assessing the quality of the technique by investigating average system time versus simulation. They compare their technique to an earlier one from Baiocchi et al. [9], which includes a similar aggregation assumption but requires calculating eigenvalues to determine the parameters of the fitted SPP. Wang et al. claim their technique is thus less complex and also provides an exact fit (as opposed to the asymptotic match provided in Baiocchi et al.).

However, the performance of both of these techniques is expected to degrade as the load on the system increases, since the superposed arrival process is burstier than the fitted SPP. To adjust for this, Wang et al. suggest over-weighting the overloaded state. They report more accurate results for time in system versus the Baiocchi et al. model, although both techniques underestimate simulation results in the presence of high server utilization.

Several papers seek alternatives to using SPPs, citing limitations in the range of marginal moments or autocorrelations that can be targeted by the SPP. Lee et al. [71] suggest that either a generalized IPP (GIPP) or a generalized interrupted Bernoulli process (GIBP) could be used to match the moments and autocovariance of interdeparture times as an improvement over standard IPP models. The GIPP is an IPP where the "on" and "off" times are generally distributed (i.e., not exponential); the GIBP is a GIPP where the general distribution is discrete. However, the authors concede that their GIPP/GIBP model can match only marginal or dependence properties of the original process, but not both.

Heyman and Lucantoni [50] also move beyond the SPP, developing the LAMBDA algorithm to fit the parameters of a discrete MMPP(m_T) (for $m_T > 2$) to a set of arrival count data. The authors claim the SPP is insufficient to model highly bursty data (i.e., more than two phases would be required). In LAMBDA, the authors split the data across a sequence of time windows, estimating the arrival rate on each window. They find the rates υ_j of the minimum order MMPP(m_T) such that every sample rate is contained in $\upsilon_j \pm 2\sqrt{\upsilon_j}$, for some $j = 1, 2, \ldots, m_T$. In this fashion, each window is associated with some phase j, and the transition probabilities in $\mathbf{A}_1$ are approximated by examining the phase transitions between consecutive windows.

The authors also use the LAMBDA algorithm to derive approximate representations of large state MMPPs by smaller order MMPPs. They note that state reduction is key in modeling because the order of a superposition of MMPPs is the product of the orders of each of its components; we elaborate on this result in the next section. The reduction technique is shown to be quite successful, as they are able to approximate, for example, the superposition of four MMPP(21)'s (over 194,000 total states) with a single MMPP(41). This is a similar idea to one proposed by Sitaraman [109], where a large order Birth-Death Modulated Poisson process (BDMPP)—a MMPP where the underlying CTMC is a birth-death process—is approximated by the superposition of SPPs and Poisson processes.

4.2.2 Superposing SPPs and Other Simplifications

Several techniques developed to match the characteristics of a nonrenewal process involve fitting the superposition of SPPs. There are two explanations for why this idea is useful: First, the superposition of MMPPs is also a MMPP [76]. If the order in the ℓ^{th} MMPP is $m_T^{(\ell)}$, for $\ell = 1, 2, \ldots, z$, then the order of the composite MMPP($m_T^{(T)}$) is $m_T^{(T)} = \prod_{\ell=1}^{z} m_T^{(\ell)}$. However, a special case of this superposition occurs when the z MMPPs are identical SPPs; as stated in Section 4.2, this superposition can be represented as a MMPP($z+1$). If the parameters of the component SPP are υ_1, υ_2, a_1, and a_2, then the BMAP representation for the MMPP($z+1$), representing the superposition of z such SPPs is

$$\upsilon_j^{(s)} = (j-1)\upsilon_1 + (z-j+1)\upsilon_2, \ (\mathbf{A}_1)_{jh} = \begin{cases} (j-1)\upsilon_1 a_1/\upsilon_j^{(s)}, & \text{if } h = j-1, \\ (z-j+1)\upsilon_2 a_2/\upsilon_j^{(s)}, & \text{if } h = j+1, \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

for $j, h = 1, 2, \ldots, z+1$, while $\boldsymbol{\alpha} = \mathbf{I}$. Thus, to target properties of a nonrenewal process with the superposition of identical SPPs requires specifying only the quantity z of SPPs and the four SPP parameters.

The second reason this superposition of identical SPPs is frequently used is that IP traffic has been shown to exhibit self-similarity and long range dependence (LRD) [72]. Since this superposition can be represented as in (7), we can use (5) to express ρ_k, for a sequence of lags $\{k_1, k_2, \ldots, k_d\}$ (for some $d \in \mathbb{Z}^+$), as functions of the SPP parameters and the quantities z and d. Hence, components of the superposed fitted process can be determined to target autocorrelations of the original process over multiple time-lags.

One paper to utilize these ideas is Andersen and Nielsen [4]. Each component SPP in their technique is expressed as the superposition of an IPP and a Poisson process; the parameters in the superposition are set to target m_1, ρ_1, and an asymptotic approximation of the autocovariance of the original counting process. Yoshihara et al. [124] propose a similar technique, targeting the exact variance of the superposed process as opposed to the asymptotic autocovariance targeted by Andersen and Nielsen. The authors utilize linear algebraic queueing theory (for background, see [74]) to determine the rates and non-linear optimization to approximate the transition probabilities in the component SPPs.

The quality of both techniques here is heavily dependent on choices for z and d. The quality of the Andersen and Nielsen technique is also dependent on the particular choice of form for the asymptotic approximation of the autocovariance function, while the range of variance that can be targeted in Yoshihara et al. is bounded. Finally, both sets of authors note their respective technique accurately captures properties of the counting process itself, but is insufficient to model nodal properties when the process feeds a queueing node.

Shah-Heydari and Le-Ngoc [108] use the superposition of identical SPPs to model count data from an arbitrary ATM stream, using the IDC curve to establish the parameters of the component SPP. This a data-fitting technique, and several of the parameters are found by minimizing the difference between the fitted pdf and the empirical pdf.

Moving beyond the superposition solely of SPPs, Salvador et al. [102, 103] use the superposition of a single MMPP(m_T) and z SPPs (not necessarily identical) to target properties of network IP traffic data. The authors separately use the SPPs to target autocovariance properties of the traffic (on z time lags) and the MMPP(m_T) to target its marginal properties. This method is also a data fitting technique which uses an approximated empirical covariance function and pdf. The superposed process is then tested on various telecommunications traces and the authors find the results satisfactory in approximating queueing behavior. One limitation here is that the superposed process has a very large order (i.e., $2^z m_T$), while a second limitation is that the output of the fitting process is generated as the solution to a set of nonlinear equations.

For a further comparison of some of the techniques described in this section, see [104].

4.2.3 Maximum-Likelihood Estimation

Meier-Hellstern [79] was the first to use ML techniques in fitting SPPs to time-series data in an effort to model processes found in telecommunication networks. In her paper, she solves for adjusted parameters from the complete likelihood function and creates a 1-to-1 correspondence between this solution and the SPP parameters. She notes that the likelihood function is unimodal, simplifying the task of computing the initial probability vector. Meier-Hellstern concedes that her model performs poorly if the data to be fit appears to be Poisson in nature; thus, the modeler must check the "Poisson-ness" of the data. Also, phases with too few arrivals may be overlooked and the estimate of the hidden phase distribution may have too few phase changes.

The dominant citation for application of ML to the general MMPP model is Rydén [100]. In this paper, the author surveys existing fitting techniques and proves the consistency of the ML estimator. He also develops a technique for using EM to estimate MMPP parameters, but cannot extend his model beyond the SPP case. Rydén's conclusion that the analytical solutions traditionally derived from ML techniques cannot be achieved in MMPP estimation has sparked work that develops numerical techniques for establishing MMPP parameters.

One such paper is Lindgren and Holst [73], who develop methods to estimate SPP parameters in a model such that the observed variable (i.e., arrival count or interarrival time) is dependent on both the current and previous state of the hidden variable (i.e., phase). However, the model here only achieves a solution when the components of the matrix product $\mathbf{UA}_1$ are small, and the authors concede that the recursion technique may need to be carefully controlled in its early stages to guarantee convergence.

Ge et al. [36] apply the 'k-means algorithm' from Deng and Mark [29] to establish an initial value for their application of the EM algorithm to the MMPP parameter problem. They find success in comparing their approximated process to a simulated MMPP(m_T) arrival process with predicted parameters, but have difficulty matching particularly small and large interarrival times. The authors also acknowledge that their fitted MMPPs may produce uncorrelated data. Nunes and Pacheco [87] also extend Deng and Mark's technique to allow for multiple arrivals in a small interval of time. The authors choose this time discretization technique as they claim rates are better estimated from small intervals, while quality estimation of transition probabilities require longer intervals.

Buchholz [22] develops an EM algorithm for fitting a MAP to real trace data by adapting a technique from Wei et al.[118] that uses initial portions of the trace to approximate conditional probabilities for being in unobservable states (i.e., phases of the fitted underlying CTMC). Buchholz's technique utilizes randomization, identifying a maximum rate from the data to use in approximating transition probabilities. As expected, the efficiency and quality of the application of EM here are heavily dependent on the value of this maximum rate. Riska et al. [98] also fit IP traffic using the EM algorithm, modeling a web server as a $MAP/Ph/1$ node. They utilize hidden Markov models in their approach, first identifying dependence in the arrival process, and then using existing techniques for fitting a Ph distribution to the interarrival data.

Recently, Okamura et al. [88] present an EM algorithm for estimating Markov-modulated compound Poisson processes (MMCPPs) which result from a MMPP combining compound Poisson processes; for background on the MMCPP, see [26]. The authors provide pseudocode for estimating the MMCPP when the intended output is multivariate normal. Their technique is dependent on the initial value of the maximization step in the EM algorithm (i.e., the M-step), and the computational intensity may be heavy if $[\mathbf{U}(\mathbf{A}_1 - \mathbf{I})]$ for the fitted process is stiff.

4.3 BMAPs: Fitting Batch Arrivals

To date, methods to fit MAPs with batch arrivals (i.e., BMAPs) to nonrenewal processes have focused on directly estimating the BMAP matrices from data using ML techniques including the EM algorithm. The general assumption behind these papers is that the data to be fit are *incomplete*; that is, the interarrival times and batch sizes (for example) are observable, but the phases of arrivals are not.

The two papers cited here differ from the remainder of the papers on matching nonrenewal processes as they take batch size into account. In Klemm et al. [65], the batch size corresponds to packet length, while in Breuer [21], the author fits a series of arrivals that occur in batches of size greater than one. We explore this below.

Klemm et al. [65] study interarrival time and volume distributions in the IP traffic found on a dial-up connection at a university site. The authors notice that by associating "rewards" (i.e., batch sizes) with arrival times, the BMAP is a superior model to either Poisson or MMPP models of IP traffic. They apply the EM algorithm to the observed data, and describe the effectiveness of their procedure by calculating $\boldsymbol{\mu}_4$ for the data rates of the measured traffic over various time scales.

Breuer [21] also develops a technique for fitting BMAP distributions by applying a simple alteration to the classical EM algorithm. The author cites his paper as the only one focused on using EM to fit BMAPs to empirical time series. The application of EM is broken into two parts: first, interarrival times are used to estimate the components of $\mathbf{A}_1$ and $\boldsymbol{v}$, after which discriminant analysis is performed on the incomplete data set (i.e., identifying unobservable phases at observable arrival instants) to estimate $\mathbf{A}_2$ and $\boldsymbol{\alpha}$. In his model, Breuer assumes the number of arrival phases is fixed, but refers the reader to Jewell [56] where the minimum number of phases is determined iteratively.

4.4 Analytical Models of the Departure Process from a $MAP/MSP/1(/K)$ Node

It is known that the stationary departure process from a $MAP/MSP/1$ node (where MSP indicates a service process characterized by a MAP) is not renewal in general; an exception is the case of the $M/M/1$ node. Bean et al. [11] is one of many papers to note this. Utilizing a description of the node size as a quasi-birth-death process (QBD) [86], this departure process can be characterized using MAP representation [13] if we allow the underlying CTMC to have infinite state space. Although exact, this result is impractical, as the departure process may serve as the arrival process to another node in a network and hence be impossible to incorporate into analytical models. Recent papers focus on approximating the departure MAP by truncating the infinite CTMC, with the necessary goal of maintaining as much of the true marginal and autocovariance information of the departure process as possible.

In an early paper on this topic, Sadre et al. [101] propose a technique for approximating the departure process from the $MAP/MSP/1$ node by a finite MAP, encompassing models from Green [37, 38], Haverkort [40], and Kumaran et al. [68] where either the service process (in Green) or both processes (in Haverkort and Kumaran et al.) are uncorrelated. Sadre et al. [101] propose a technique to identify a truncation point for the space of the underlying CTMC, aggregating phases with larger indices into a single phase; this technique is an extension of [10] in which the queue length is truncated to yield an approximation of the departure process by a MAP with a finite state-space. Sadre et al. [101] also propose techniques for identifying multiple truncation points, which allows for matching multiple autocorrelation targets; however, their results show that improvements from this do not always justify the increased complexity of the model with multiple truncations.

Heindl and Telek [48] investigate tandem networks of $\cdot/Ph/1(/K)$ nodes (with one external MAP arrival stream), providing MAP approximations for the departure process during a busy period. Their technique involves using the DTMC of the QBD process (describing the queue size) embedded in a semi-Markov process (SMP), and then providing a MAP representation for the SMP describing the output process. Notice that this requires calculating distributions for the idle time of the server, conditional on whether the previous busy period consisted of a single service or multiple services.

Recently, Heindl et al. [49] utilize ETAQA [27, 96] for aggregating states in the infinite MAP departure process from the $MAP/MSP/1$ node. In ETAQA, the QBD queueing process is truncated and its generator matrix is specified using techniques introduced by Latouche and Ramaswami [70]. Heindl et al. compare the complexity of their model to Sadre et al. [101], and note their technique is more efficient when the only goal of the analysis is to describe an output MAP; however, if performance measures are sought for downstream nodes, then the two techniques have a similar efficiency. ETAQA is implemented in the modeling tool MAMSolver [97].

Several of the truncation techniques described here have been utilized in network decomposition. Notice the resulting processes from splitting a MAP (e.g. due to Markovian routing) or superposing MAPs (e.g., from multiple departure processes feeding a single node) are also MAPs. Thus, these techniques—when successfully utilized in specifying the MAP representation of the truncated departure process—lead to MAP representations for the split or superposed arrival process at a downstream $\cdot/MSP/1$ node.

4.5 Minimal MAP Representations

As we have seen, most MAP fitting techniques utilize special structures for the $\mathbf{A}_1$ and $\boldsymbol{\alpha}$ matrices. A MAP(m_T) is characterized by $m_T(2m_T-1)$ free parameters and, therefore, is often over-parameterized in terms of targeting a few specific properties of a general point process. An open question in MAP characterization is in finding minimal BMAP representations (i.e., MAPs with the correct properties that utilize a minimal number of non-zero parameter values). Along these lines, Bodrog et al. [19] discuss the relationship between AMAP(2)'s and MAP(2)'s, while Telek and Horváth [54] extend van de Liefvoort's result [115] on converting distributional moments into rational LST's, and attempt to specify a minimal MAP representation from there. For further discussion on the current status of this topic, see [20].

4.6 Evaluation of Fitting with MAPs

In this section we have surveyed several techniques for specifying MAPs to target properties of nonrenewal point processes. Many of the papers cited here are data-fitting techniques that specify the MAP based on histograms or from results of ML methods. These papers do a sufficient job of fitting data but cannot be extended to matching descriptors (i.e., marginal moments and dependence measures).

Those techniques most suitable for targeting descriptors are the AMAP(2), the Markov-MECO model, and several of the MMPP papers, including those from Heffes, Lucantoni, and their co-authors. Although their techniques accurately target marginal properties of the original process, upon extending the target to dependence measures they each have limitations. Often they target only a single dependence measure at a time (so either a short or long range dependence measure may be matched, but not both) or the achievable range of autocorrelation is limited. The model from Andersen and Nielsen improves on this by targeting several time-lags, but their technique provides only asymptotic approximations for the parameters in their model. Unlike the renewal-fitting problem, discussed in Section 3, the problem of finding a technique to accurately target several dependence measures while matching marginal properties appears to still be open.

5 Summary and Further Research

In this paper we have provided a survey of tools that have been developed to approximate general stationary point processes in a Markovian framework to make models more analytically tractable. We have provided an overview of techniques to match characteristics of renewal and nonrenewal processes, with a focus on the latter and the efforts made to capture the dependence present in many of these point processes.

Work continues to be done in this area, as MAPs (and their special cases such as MMPPs) remain the most effective tool for modeling processes in telecommunications systems and related areas. From here we may expect to see further tweaking of the aforementioned models in an effort to improve the range and quality of what is captured. The idea that which characteristics of a point process are important to match appears to be problem-dependent leaves the door open for further efforts.

One research area where providing an accurate approximation of a general nonrenewal process with a MAP might play an important role is in modeling internode traffic flow in nonstationary queueing networks. Notice that any matching technique designed to accomplish this must be able to specify Markovian processes that are fairly flexible, as the traffic flow to be approximated may be more variable than Poisson on some portions of the time horizon and less variable on others; additionally, it may exhibit extreme levels of positive and negative autocorrelation which occur at multiple lags. Further, the approximated Markovian process must be specified to match not only descriptors of the traffic flow itself, but also yield accurate approximations for congestion measures at nodes that the traffic flow feeds. At present, there is no single matching technique that can meet all of these potential requirements.

Acknowledgments

The authors thank Mike Taaffe for helpful discussions. This work is supported by National Science Foundation Grant DMII-0521857.

References

[1] D. Aldous and L. Shepp. The least variable phase type distribution is Erlang. *Communications in Statistics–Stochastic Models*, 3(3):467–473, 1987.

[2] T. Altiok. On the phase-type approximations of general distributions. *IIE Transactions*, 17(2):110–116, 1985.

[3] A. T. Andersen, M. F. Neuts, and B. F. Nielsen. On the time reversal of Markovian arrival processes. *Communications in Statistics - Stochastic Models*, 20(2):237–260, 2004.

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

[5] S. Asmussen. *Applied Probability and Queues.* John Wiley & Sons, New York, 1987.

[6] S. Asmussen. Phase-type distributions and related point processes: Fitting and recent advances. In *Matrix-Analytic Methods in Stochastic Models, Lecture Notes In Pure and Applied Mathematics*, pages 137–149. Marcel Dekker, Inc., 1997.

[7] S. Asmussen. Matrix-analytic models and their analysis. *Scandinavian J. of Statistics*, 27:193–226, 2000.

[8] S. Asmussen, O. Nerman, and M. Olson. Fitting phase type distributions via the EM Algorithm. *Scandinavian J. of Statistics*, 23:419–441, 1996.

[9] A. Baiocchi, N. B. Melazzi, M. Listanti, A. Roveri, and R. Winkler. Loss performance analysis of an ATM multiplexer loaded with high-speed on-off sources. *IEEE Journal on Selected Areas in Communications*, 9(3):388–393, Apr 1991.

[10] N. G. Bean, D. A. Green, and P. G. Taylor. Approximations to the output processes of $MAP/M/1$ queues. In *Advances in Matrix Analytic Methods for Stochastic Models*, pages 151–170. Second International Workshop on Matrix-Analytic Methods, 1998.

[11] N. G. Bean, D. A. Green, and P. G. Taylor. The output process of an $MMPP/M/1$ queue. *J. Appl. Probab.*, 35(4):998–1002, 1998.

[12] G. R. Bitran and S. Dasu. Approximating nonrenewal processes by Markov chains: Use of Super-Erlang (SE) chains. *Operations Research*, 41(5):903–923, 1993.

[13] G. R. Bitran and S. Dasu. Analysis of the $\sum Ph_i/Ph/1$ queue. *Operations Research*, 42(1):158–174, 1994.

[14] A. Bobbio and A. Cumani. ML estimation of the parameters of a Ph distribution in triangular canonical form. In G. Serazzi G. Balbo, editor, *Computer Performance Evaluation*, pages 33–46. Elsevier,Amsterdam, 1992.

[15] A. Bobbio, A. Horváth, M. Scarpa, and M. Telek. Acyclic discrete phase-type distributions: Properties and a parameter estimation algorithm. *Performance Evaluation*, 54(1):1–32, 2003.

[16] A. Bobbio, A. Horváth, and M. Telek. The scale factor: A new degree of freedom in phase-type approximation. *Performance Evaluation*, 56:121–144, 2004.

[17] A. Bobbio, A. Horváth, and M. Telek. Matching three moments with minimal acyclic phase type distributions. *Stochastic Models*, 21:303–326, 2005.

[18] A. Bobbio and M. Telek. A benchmark for Ph estimation algorithms: Results for acyclic-Ph. *Communications in Statistics.–Stochastic Models*, 10:661–677, 1994.

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

[20] L. Bodrog, A. Heindl, G. Horváth, M. Telek, and A. Horváth. Current results and open questions on Ph and MAP characterization. In D. Bini, B. Meini, V. Ramaswami, M.-A. Remiche, and P. G. Taylor, editors, *Numerical Methods for Structured Markov Chains*, volume 07461 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2008.

[21] 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.

[22] P. Buchholz. An EM-algorithm for MAP fitting from real traffic data. In P. Kemper and W. H. Sanders, editors, *Computer Performance Evaluation / TOOLS*, volume 2794 of *Lecture Notes in Computer Science*, pages 218–236. Springer, 2003.

[23] W. Bux and U. Herzog. The phase concept: Approximation of measured data and performance analysis. In K.M. Chandy and M. Reiser, editors, *Computer Performance*, pages 23–38. North-Holland, New York, 1977.

[24] G. Casale, E. Z. Zhang, and E. Smirni. KPC-Toolbox: Simple yet effective trace fitting using Markovian arrival processes. To appear in *QEST'08*.

[25] G. Casale, E. Z. Zhang, and E. Smirni. Interarrival times characterization and fitting for Markovian traffic analysis. In D. Bini, B. Meini, V. Ramaswami, M.-A. Remiche, and P. G. Taylor, editors, *Numerical Methods for Structured Markov Chains*, volume 07461 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2008.

[26] R. Chakka and T. van Do. The $MM\sum_{k=1}^{K} CPP_k/GE/c/L$ G-queue with heterogeneous servers: Steady state solution and an application to performance evaluation. *Performance Evaluation*, 64(3):191–209, 2007.

[27] G. Ciardo and E. Smirni. ETAQA: An efficient technique for the analysis of QBD-processes by aggregation. *Performance Evaluation*, 36-37(1-4):71–93, 1999.

[28] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM Algorithm. *J. of Royal Statistical Society, Series B*, 39:1–38, 1977.

[29] L. Deng and J. W. Mark. Parameter estimation for Markov-modulated Poisson processes via the EM Algorithm with time-discretization. *Telecommunications Systems*, 1:321–338, 1993.

[30] J. E. Diamond and A. S. Alfa. On approximating higher order MAPs with MAPs of order two. *Queueing Systems*, 34:269–288, 2000.

[31] M. Fackrell. Fitting with matrix-exponential distributions. *Stochastic Models*, 21:377–400, 2005.

[32] A. Feldman and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31:245–279, 1998.

[33] H.-W. Ferng and J.-F. Chang. Connection-wise end-to-end performance analysis of queuing networks with MMPP inputs. *Performance Evaluation*, 43(1):39–62, 2001.

[34] H.-W. Ferng and J.-F. Chang. Departure processes of $BMAP/G/1$ queues. *Queueing Syst. Theory Appl.*, 39(2-3):109–135, 2001.

[35] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18(2):149–171, 1993.

[36] H. Ge, U. Harder, and P. G. Harrison. Parameter estimation for MMPPs using the EM algorithm. In *Proceedings of UKPEW 2003*, pages 293–306, 2003.

[37] D. Green. *Departure Processes from $MAP/PH/1$ Queues.* PhD thesis, University of Adelaide, 1999.

[38] D. Green. Lag correlations of approximating departure processes of $MAP/PH/1$ queues. In *Proceedings of the Third International Conference on Matrix-Analytic Methods in Stochastic Models*, pages 135–151, 2000.

[39] R. Gusella. Characterizing the variability of arrival processes with indexes of dispersion. *IEEE J. on Selected Areas in Communications*, 9(2):203–211, 1991.

[40] B. R. Haverkort. Approximate analysis of networks of $PH/PH/1/K$ queues with customer losses: Test results. *Annals of Operations Research*, 79:271–291, 1998.

[41] H. Heffes. A class of data traffic processes: Covariance function characterization and related queueing results. *Bell System Technical Journal*, 59(6):897–929, 1980.

[42] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. on Selected Areas in Communications, Special Issue on Network Performance Evaluation*, 4:856–868, 1986.

[43] A. Heindl. Decomposition of general tandem queueing networks with MMPP input. *Performance Evaluation*, 44(1-4):5–23, 2001.

[44] A. Heindl. Inverse characterization of hyperexponential MAP(2)s. In *Proc. 11th Int. Conference on Analytical and Stochastic Modelling Techniques and Applications*, pages 183–189, 2004.

[45] A. Heindl, G. Horváth, and K. Gross. Explicit inverse characterizations of acyclic MAPs of second order. In András Horváth and Miklós Telek, editors, *EPEW*, volume 4054 of *Lecture Notes in Computer Science*, pages 108–122. Springer, 2006.

[46] A. Heindl, K. Mitchell, and A. van de Liefvoort. The correlation region of second-order MAPs with application to queueing network decomposition. In *Computer Performance Evaluation / TOOLS*, pages 237–254, 2003.

[47] A. Heindl, K. Mitchell, and A. van de Liefvoort. Correlation bounds for second-order MAPs with application to queueing network decomposition. *Performance Evaluation*, 63(6):553–577, 2006.

[48] A. Heindl and M. Telek. MAP-based decomposition of tandem networks of $\cdot/PH/1(/K)$ queues with MAP input. In *MMB*, pages 179–194, 2001.

[49] A. Heindl, Q. Zhang, and E. Smirni. ETAQA truncation models for the $MAP/MAP/1$ departure process. In *QEST*, pages 100–109. IEEE Computer Society, 2004.

[50] D. P. Heyman and D. M. Lucantoni. Modeling multiple IP traffic streams with rate limits. *IEEE/ACM Transactions on Networking*, 11(6):948–958, 2003.

[51] A. Horváth and M. Telek. Approximating heavy tailed behavior with phase type distributions. In *Advances in Matrix-Analytic Methods for Stochastic Models*, Notable Publications, pages 191–214. 2000.

[52] A. Horváth and M. Telek. Phfit: A general phase-type fitting tool. In *Proceedings of Tools 2002*, pages 82–91, 2002.

[53] A. Horváth and M. Telek. Matching more than three moments with acyclic phase type distributions. *Stochastic Models*, 23(2):167–194, 2007.

[54] G. Horváth and M. Telek. A minimal representation of Markov arrival processes and a moments matching method. *Performance Evaluation*, 64(9–12):1153–1168, Aug. 2007.

[55] G. Horváth, M. Telek, and P. Buchholz. A MAP fitting approach with independent approximation of the inter-arrival time distribution and the lag correlation. In *QEST*, pages 124–133. IEEE Computer Society, 2005.

[56] N. P. Jewell. Mixtures of exponential distributions. *Annals of Statistics*, 10(2):479–484, 1982.

[57] M. A. Johnson. Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and Erlang distributions. *ORSA Journal on Computing*, 5(1):69–83, 1993.

[58] M. A. Johnson. Markov MECO: A simple Markovian model for approximating nonrenewal arrival processes. *Communications in Statistics–Stochastic Models*, 14(1&2):419–442, 1998.

[59] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Mixtures of Erlang distributions of common order. *Communications in Statistics–Stochastic Models*, 5:711–743, 1989.

[60] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Density function shapes. *Communications in Statistics–Stochastic Models*, 6:283–306, 1990.

[61] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Nonlinear programming approaches. *Communications in Statistics–Stochastic Models*, 6:259–281, 1990.

[62] M. A. Johnson and M. R. Taaffe. An investigation of phase-distribution moment matching algorithms for use in queueing models. *Queueing Systems*, 8:129–147, 1991.

[63] S. H. Kang, Y. H. Kim, D. K. Sung, and B. D. Choi. An application of Markovian arrival process (MAP) to modeling superposed ATM cell streams. *IEEE Transactions on Communications*, 50(4):633–642, 2002.

[64] R. El Abdouni Khayari, R. Sadre, and B. R. Haverkort. Fitting world-wide web request traces with the EM-algorithm. *Performance Evaluation*, 52(2-3):175–191, 2003.

[65] A. Klemm, C. Lindemann, and M. Lohmann. Traffic modeling of IP networks using the batch Markovian arrival process. In *Proceedings of Tools 2002*, pages 92–110, 2002.

[66] P. Kuehn. Approximate analysis of general queuing networks by decomposition. *IEEE Transactions on Communications*, 27(1):113–126, Jan 1979.

[67] V. G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall, Ltd., London, UK, 1995.

[68] J. Kumaran, K. Mitchell, and A. van de Liefvoort. Characterization of the departure process from an $ME/ME/1$ queue. *Operations Research*, 38(2):173–191, 2004.

[69] A. Lang and J. L. Arthur. Parameter approximation for phase-type distributions. In *Matrix-Analytic Methods in Stochastic Models, Lecture Notes In Pure and Applied Mathematics*, pages 266–274. Marcel Dekker, Inc., 1996.

[70] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, Philadelphia, 1999.

[71] Y. D. Lee, A. van de Liefvoort, and V. L. Wallace. Modeling correlated traffic with a generalized IPP. *Performance Evaluation*, 40(1-3):99–114, 2000.

[72] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2(1):1–15, 1994.

[73] G. Lindgren and U. Holst. Recursive estimation of parameters in Markov-modulated Poisson processes. *IEEE Transactions on Communications*, 43(11):2812–2820, 1995.

[74] L. Lipsky. *Queueing Theory: A Linear Algebraic Approach*. MacMillan, New York, 1992.

[75] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics–Stochastic Models*, 7(1):1–46, 1991.

[76] D. M. Lucantoni. The $BMAP/G/1$ queue: A tutorial. In *Performance Evaluation of Computer and Communication Systems, Joint Tutorial Papers of Performance '93 and Sigmetrics '93*, pages 330–358, London, UK, 1993. Springer-Verlag.

[77] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A single server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22:676–705, 1990.

[78] R. Marie. Calculating equilibrium probabilities for $\lambda(n)/c_k/1/N$ queues. In *Proceedings of the 1980 International Symposium on Computer Performance Modelling, Measurement and Evaluation*, pages 117–125, 1980.

[79] K. S. Meier-Hellstern. A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European J. of Operational Research*, 29:370–377, 1987.

[80] K. Mitchell. Constructing a correlated sequence of matrix exponentials with invariant first-order properties. *Operations Research Letters*, 28(1):27–34, 2001.

[81] K. Mitchell, K. Sohraby, A. Van de Liefvoort, and J. Place. Approximation models of wireless cellular networks using moment matching. *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, 1:189–197, 2000.

[82] K. Mitchell and A. van de Liefvoort. Approximation models of feed-forward $G/G/1/N$ queueing networks with correlated arrivals. *Performance Evaluation*, 51(2-4):137–152, 2003.

[83] R. Nagarajan, J. F. Kurose, and D. F. Towsley. Approximation techniques for computing packet loss in finite-buffered voice multiplexers. *IEEE Journal on Selected Areas in Communications*, 9(3):368–377, 1991.

[84] B. L. Nelson and M. R. Taaffe. The $MAP_t/Ph_t/\infty$ queueing system and multiclass $[MAP_t/Ph_t/\infty]^K$ queueing network. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2006.

[85] M. F. Neuts. A versatile Markovian point process. *J. of Applied Probability*, 16(4):764–779, 1979.

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

[87] C. Nunes and A. Pacheco. Parametric estimation in MMPP(2) using time discretization. In *Proceedings of the 2nd Internation Symposium on Semi-Markov Models: Theory and Applications*, 1998.

[88] H. Okamura, Y. Kamahara, and T. Dohi. Estimating Markov-modulated compound Poisson processes. In *ValueTools '07: Proceedings of the 2nd international conference on Performance evaluation methodologies and tools*, pages 1–8, ICST, Brussels, Belgium, Belgium, 2007. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[89] M. Olsson. The EMpht-programme. Technical report, Department of Mathematics, Chalmers University of Technology, 1998.

[90] T. Osogami and M. Harchol-Balter. Necessary and sufficient conditions for representing general distributions by Coxians. Technical report, CMU-CS-02-178, School of Computer Science, Carnegie Mellon University, 2002.

[91] T. Osogami and M. Harchol-Balter. A closed-form solution for mapping general distributions to minimal Ph distributions. Technical report, CMU-CS-03-114, School of Computer Science, Carnegie Mellon University, 2003.

[92] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[93] J. F. Pérez and G. Riaño. jPhase: An object-oriented tool for modeling phase-type distributions. In *SMCtools '06: Proceeding from the 2006 workshop on Tools for solving structured Markov chains*, page 5, New York, NY, USA, 2006. ACM.

[94] V. Ramaswami. The $N/G/1$ queue and its detailed analysis. *Advances in Applied Probability*, 12(1):222–261, 1980.

[95] A. Riska, V. Diev, and E. Smirni. An EM-based technique for approximating long-tailed data sets with Ph distributions. *Performance Evaluation*, 55(1&2):147–164, 2004.

[96] A. Riska and E. Smirni. Exact aggregate solutions for $M/G/1$-type Markov processes. *SIGMETRICS Performance Evaluation Rev.*, 30(1):86–96, 2002.

[97] A. Riska and E. Smirni. MAMSolver: A matrix analytic methods tool. In *TOOLS '02: Proceedings of the 12th International Conference on Computer Performance Evaluation, Modelling Techniques and Tools*, pages 205–211, London, UK, 2002. Springer-Verlag.

[98] A. Riska, M. Squillante, S. Yu, Z. Liu, and L. Zhang. Matrix-analytic analysis of a $MAP/PH/1$ queue fitted to web server data. In G. Latouche and P. Taylor, editors, *Matrix-analytic Methods: Theory and Applications*, Dagstuhl Seminar Proceedings, pages 333–356. World Scientific, 2002.

[99] M. H. Rossiter. *Characterizing a Random Point Process by a Switched Poisson Process.* PhD thesis, Monash University, Melbourne, 1989.

[100] T. Rydén. Parameter estimation for Markov modulated Poisson processes. *Communications in Statistics–Stochastic Models*, 10(4):795–829, 1994.

[101] R. Sadre, B. R. Haverkort, and A. Ost. An efficient and accurate decomposition method for open finite- and infinite-buffer queueing networks. In *Proc. 3rd Int. Workshop on Numerical Solution of Markov Chains*, pages 1–20. Zaragosa University Press, 1999.

[102] P. Salvador, A. Nogueira, R. Valadas, and A. Pacheco. Multi-time-scale traffic modeling using Markovian and L-systems models. In *Universal Multiservice Networks*, Lecture Notes in Computer Science, pages 297–306. Springer, Berlin / Heidelberg, 2004.

[103] P. Salvador, R. Valadas, and A. Pacheco. Multiscale fitting procedure using Markov-modulated Poisson processes. *Telecommunications Systems*, 23(1&2):123–148, 2003.

[104] P. S. Salvador, A. N. Nogueira, and R. Valadas. Modelling local area network traffic with Markovian traffic models. In *Proc Conf. on Telecommunications - ConfTele, Figueira da Foz, Portugal*, 2001.

[105] C. Sauer and K. Chandy. Approximate analysis of central server models. *IBM J. of Research and Development*, 19:301–313, 1975.

[106] L. Schmickler. MEDA: Mixed Erlang distributions as phase-type representations of empirical distribution functions. *Communications in Statistics–Stochastic Models*, 8:131–156, 1992.

[107] S. Shah-Heydari and T. Le-Ngoc. MMPP modeling of aggregated ATM traffic. *Canadian Conference on Electrical and Computer Engineering (CCECE'98)*, Waterloo, Canada:129–132, 1998.

[108] S. Shah-Heydari and T. Le-Ngoc. MMPP models for multimedia traffic. *Telecommunications Systems*, 15:273–293, 2000.

[109] H. Sitaraman. Approximation of some Markov modulated Poisson processes. *ORSA J. on Computing*, 3(1):12–22, 1991.

[110] P. Skelly, M. Schwartz, and S. Dixit. A histogram-based model for video traffic behavior in an ATM multiplexer. *Transactions on Networking*, 1:446–458, 1993.

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

[112] M. Telek and A. Heindl. Matching moments for acyclic discrete and continuous phase-type distributions of second order. *International J. of Simulation*, 3(3-4):47–57, 2003.

[113] A. Thümmler, P. Buchholz, and M. Telek. A novel approach for fitting probability distributions to real trace data with the EM algorithm. In *DSN '05: Proceedings of the 2005 International Conference on Dependable Systems and Networks*, pages 712–721, Washington, DC, USA, 2005. IEEE Computer Society.

[114] H. C. Tijms. *Stochastic Models: An Algorithmic Approach.* John Wiley & Sons, Inc, Chichester, England, 1994.

[115] A. van de Liefvoort. The moment problem for continuous distributions, Working Paper CM-1990-02. Technical report, Univ. of Missouri, 1990.

[116] S. S. Wang and J. A. Silvester. An approximate model for performance evaluation of real-time multimedia communication systems. *Performance Evaluation*, 22(3):239–256, 1995.

[117] A. J. Weerstra. Using matrix-geometric methods to enhance the QNA method for solving large queueing metworks. Master's thesis, University of Twente, 1994.

[118] W. Wei, B. Wang, and D. Towsley. Continuous-time hidden Markov models for network performance evaluation. *Performance Evaluation*, 49(1-4):129–146, 2002.

[119] W. Whitt. Approximating a point process by a renewal process: The view through a queue, an indirect approach. *Management Science*, 27:619–634, 1981.

[120] W. Whitt. Approximating a point process by a renewal process, I: Two basic methods. *Operations Research*, 30:125–147, 1982.

[121] W. Whitt. The Queueing Network Analyzer. *Bell System Technical Journal*, 62(9):2779–2815, Nov. 1983.

[122] W. Whitt. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Labs Technical J.*, 63(1):163–175, 1984.

[123] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

[124] T. Yoshihara, S. Kasahara, and Y. Takahashi. Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telecommunication Systems*, 17:185–211, 2001.

Appendices

A MAP(2): Formula for the lag-k autocorrelation

We provide the explicit expression for ρ_k for the MAP(2), for $k \geq 1$. We use shorthand notation

$$\kappa_1 = \frac{(1-a_1)(1-\alpha_1) + a_1\alpha_2(1-a_2)}{1-a_1a_2},\ \kappa_2 = \frac{(1-a_2)(1-\alpha_2) + a_2\alpha_1(1-a_1)}{1-a_1a_2}.$$

From (5), we find $\rho_k = c_\rho \xi^k$, such that

$$\begin{aligned} \xi &= 1 - \kappa_1 - \kappa_2, \\ c_\rho &= \frac{(\kappa_1+\kappa_2)\left[\upsilon_1\kappa_2(\kappa_2 a_1 + \kappa_1) - \upsilon_2\kappa_1(\kappa_2 + \kappa_1 a_2)\right]\left[\upsilon_2(1-a_2) - \upsilon_1(1-a_1)\right]}{d_1+d_2}, \end{aligned}$$

where

$$\begin{aligned} d_1 &= \upsilon_1(\kappa_2 a_1 + \kappa_1)\left[(\kappa_1 + 2\kappa_2)(\upsilon_2 a_2 + \upsilon_1) - \kappa_2(\upsilon_2 + \upsilon_1 a_1)\right], \\ d_2 &= \upsilon_2(\kappa_2 + \kappa_1 a_2)\left[(2\kappa_1 + \kappa_2)(\upsilon_2 + \upsilon_1 a_1) - \kappa_1(\upsilon_2 a_2 + \upsilon_1)\right], \end{aligned}$$

for $k \geq 1$.

B AMAP(2) Fitting: Heindl et al. [45]

We provide formulas to target $\mathbf{m}_3$ and ρ_1 with an AMAP(2) (given free parameter $\eta \in [0,1]$). We use shorthand notation

$$h_2 = \frac{m_2}{2m_1^2} - 1,\ h_3 = \frac{2m_3m_1 - 3m_2^2}{12m_1^4},\ h_4 = h_3 + h_2^2 - h_2,\ h_5 = \sqrt{h_4^2 + 4h_2^3},$$

$$h_6 = \frac{(1-\eta)(2h_2\xi + h_4 - h_5) + \xi(h_4+h_5) - (h_4 - h_5)}{(1-\eta)(2h_2 + h_4 - h_5) + 2h_5},$$

$$h_7 = \frac{(\xi-1)(h_4-h_5)}{(1-\eta)(2h_2 + h_4 - h_5) + 2h_5},$$

where parameter $\xi = (2h_2+1)\rho_1/h_2$. If $h_2 > 0$ (i.e., *scv* > 1), then $\mathbf{m}_3$ may only be matched if additionally $h_3 > 0$. If this holds, the fitted BMAP parameters for the AMAP(2) are

$$\upsilon_1 = \frac{2h_2 + h_4 - h_5}{2m_1h_3},\ \upsilon_2 = \frac{2h_2 + h_4 + h_5}{2m_1h_3},\ a_1 = 1-\eta,\ a_2 = 0,\ \alpha_1 = \frac{h_6}{\eta},\ \alpha_2 = 1 - h_7. \quad \text{(A.1)}$$

If $h_2 < 0$ (i.e., $scv < 1$), then $h_3 < 0$ must hold. If so, the fitted BMAP parameters are the same as in (A.1), except the sign in front of h_5 must be switched in each place that shorthand appears.

C SPP Fitting: Heffes and Lucantoni [42]

We provide formulas to target m_1, I_∞, $I(t_1)$, and $f_3(t_2) \equiv \mathbb{E}\{(N(t_2)^3\}$ with a SPP (given times t_1, $t_2 > 0$). Notice that time t_1 must be selected such that $I(t_1) > 1$.

We define two terms that will be useful here: d_1, which solves

$$d_1 = \frac{1}{t_1}\left(\frac{I_\infty - 1}{I_\infty - I(t_1)}\right)\left(1 - e^{-d_1 t_1}\right), \tag{A.2}$$

and C, which satisfies

$$\begin{aligned} f_3(t_2) \;&=\; \left(\frac{t_2}{m_1}\right)^3 + \frac{3\,(I_\infty - 1)\,t_2^2}{m_1^2} + \frac{3\,(I_\infty - 1)\,t_2}{d_1 m_1}\left(\frac{C}{d_1} - m_1^{-1}\right) \\ &\quad + \frac{3\,(I_\infty - 1)\,t_2 e^{-d_1 t_2}}{d_1^2 m_1}\left(C + \frac{d_1}{m_1}\right) - \frac{6C\,(I_\infty - 1)}{d_1^3 m_1}\left(1 - e^{-d_1 t_2}\right). \end{aligned}$$

If $C = 0$, then we define $h_1 = h_2 = d_1/2$ and

$$\ell_1 = m_1^{-1} + \frac{d_1}{2}\,(I_\infty - 1)\,,\;\; \ell_2 = m_1^{-1} - \frac{d_1}{2}\,(I_\infty - 1)\,,$$

while if $C \neq 0$, we define

$$d_2 = \frac{(I_\infty - 1)\,d_1^3}{2C^2 m_1},$$

and

$$h_1 = \frac{d_1}{2}\left(1 + \frac{1}{\sqrt{4d_2 + 1}}\right),\;\; h_2 = d_1 - h_1,\;\; \ell_2 = m_1^{-1} - \frac{Ch_2}{d_1\,(h_1 - h_2)},\;\; \ell_1 = \frac{C}{h_1 - h_2} + \ell_2.$$

Then the fitted SPP parameters are $\upsilon_j = h_j + \ell_j$ and $a_j = h_j/\upsilon_j$, for $j = 1, 2$.

D SPP Fitting: Gusella [39]

We provide formulas to target m_1, scv, I_∞, and $I(t_1)$ with a SPP (given $t_1 > 0$). Notice that time t_1 must be selected such that $I(t_1) > 1$.

We utilize d_1, as in (A.2), and find ℓ_2 which solves

$$scv = \frac{2m_1\ell_2^2 + \ell_2\left[2m_1d_1 + m_1d_1\left(I_\infty - 1\right) - 2\right] - 2d_1I_\infty}{2m_1\ell_2^2 + \ell_2\left[2m_1d_1 - m_1d_1\left(I_\infty - 1\right) - 2\right] - 2d_1}. \tag{A.3}$$

Notice that typically there will be two solutions for ℓ_2 in (A.3). Then, defining

$$h_1 = \frac{m_1d_1^2\left(I_\infty - 1\right)}{2 + m_1d_1\left(I_\infty - 1\right) - 4m_1\ell_2 + 2m_1^2\ell_2^2},$$

$$h_2 = \frac{2d_1\left(m_1\ell_2^2 - 1\right)^2}{2 + m_1d_1\left(I_\infty - 1\right) - 4m_1\ell_2 + 2m_1^2\ell_2^2},$$

$$\ell_1 = \frac{2 + m_1d_1\left(I_\infty - 1\right) - 2m_1\ell_2}{2m_1 - 2m_1^2\ell_2},$$

leads to the fitted SPP parameters $\upsilon_j = h_j + \ell_j$ and $a_j = h_j/\upsilon_j$, for $j = 1, 2$.

E SPP Fitting: Ferng and Chang [33]

We provide formulas to target $\mathbf{m}_3$ and ρ_1 with a SPP. We define several terms that will be useful here: first, $d_1 = m_1^{-1}$,

$$d_2 = \frac{\rho_1\left(m_2 - m_1^2\right)}{m_1\left[\left(m_2 - 2m_1^2\right)/2 - \rho_1\left(m_2 - m_1^2\right)\right]},\ d_3 = \frac{\left[\left(m_2 - 2m_1^2\right)\left(d_1 + d_2\right) + 2d_2m_1^2\right]}{2d_1m_1^2},$$

$$d_4 = \frac{6\left(d_1d_3 - d_2\right)}{m_3d_1\left(d_1 + d_2\right)^2 - 3m_2d_1\left(d_1 + d_2\right) - 6\left(d_3 + d_3^2\right)},$$

$d_5 = d_1d_4$, $d_6 = d_2d_4$, and $d_7 = d_1 + d_3d_4$. From these, we define

$$\ell_1 = \frac{d_7 + \sqrt{d_7^2 - 4d_6}}{2},\ \ell_2 = \frac{d_7 - \sqrt{d_7^2 - 4d_6}}{2}.$$

If $\ell_1 < \ell_2$, then we reverse their assignments. Then

$$h_1 = \frac{d_4\ell_1 - d_5}{\ell_1 - \ell_2},\ h_2 = \frac{d_5 - d_4\ell_1}{\ell_1 - \ell_2},$$

and the fitted SPP parameters are $\upsilon_j = h_j + \ell_j$ and $a_j = h_j/\upsilon_j$, for $j = 1, 2$.