Copula Analysis of Temporal Dependence Structure in Markov Modulated Poisson Process and Its Applications

FANG DONG, KUI WU, and VENKATESH SRINIVASAN, University of Victoria

The Markov Modulated Poisson Process (MMPP) has been extensively studied in random process theory and widely applied in various applications involving Poisson arrivals whose rate varies following a Markov process. Despite the rich literature on MMPP, very little is known on its intricate temporal dependence structure. No exact solution is available so far to capture the *functional* temporal dependence of MMPP at the stationary state over slotted times.

This article tackles the above challenges with copula analysis. It not only presents a novel analytical framework to capture the temporal dependence of MMPP but also provides the exact copula-based solutions for single MMPP as well as the aggregate of independent MMPP. This theoretical contribution discloses functional dependence structure of MMPP. It also lays the foundation for many applications that rely on the temporal dependence of MMPP for adaptive control or predictive resource provisioning. We demonstrate case studies, with real-world trace data as well as simulation, to illustrate the practical significance of our analytical results.

CCS Concepts: • **Computing methodologies → Model development and analysis**; • **Theory of computation → Theory and algorithms for application domains;**

Additional Key Words and Phrases: Copula analysis, markov modulated poisson process, traffic prediction

ACM Reference Format:
Fang Dong, Kui Wu, and Venkatesh Srinivasan. 2017. Copula analysis of temporal dependence structure in markov modulated poisson process and its applications. ACM Trans. Model. Perform. Eval. Comput. Syst. 2, 3, Article 14 (June 2017), 28 pages.
DOI: http://dx.doi.org/10.1145/3089254

Nomenclature

Q	The transition rate matrix for a CTMC
Λ	The vector of Poisson arrival rates
Π	The stationary distribution for a CTMC
$P(t)$	The transition matrix for a CTMC after time t
ρ_τ	Kendall's tau
ρ_s	Spearman's rho
ρ_t^+	The upper tail dependence coefficient
ρ_t^-	The lower tail dependence coefficient
ρ	Pearson's correlation coefficient
A_i	The arrival count in ith time slot of single trace
A_i^l $(l > 1)$	The arrival count in ith time slot of l aggregate trace
$G_j(x)$	The marginal distribution of A_i on the condition that associated CTMC is in state j

This work was supported by the Natural Sciences and Engineering Research Council of Canada.
Authors' addresses: F. Dong, K. Wu, and V. Srinivasan, Department of Computer Science, University of Victoria, 3800 Finnerty Rd, Victoria, BC, Canada, V8P 5C2; emails: {fdong, wkui, srinivas}@uvic.ca.

DOI: http://dx.doi.org/10.1145/3089254

$\mathcal{G}(x)$	The vector $\mathcal{G}(x) = [G_1(x), G_2(x), \ldots, G_m(x)]$
$M_l(x)$	The marginal distribution for l-IHMMPP ($l \geq 1$)
$C_l(u, v)$	The copula for l-IHMMPP ($l \geq 1$)
$\nabla C_l(u, v)$	The copula gradient for l-IHMMPP ($l \geq 1$)
$\mathcal{O}^l$	The l-IHMMPP copula matrix ($l \geq 1$)
$\mathcal{R}^l$	The l-IHMMPP copula gradient matrix ($l \geq 1$)
$C_{\{l;i'\}}(u, v)$	The copula for i'-step dependence in l-IHMMPP ($l \geq 1$)

1. INTRODUCTION

Markov modulated Poisson process (MMPP) is the doubly stochastic Poisson process whose arrival rate is modulated by an irreducible continuous time Markov chain (CTMC) independent with the arrival process [Fischer and Meier-Hellstern 1993]. Specifically, the arrival process is a Poisson process with arrival rate λ_j whenever the CTMC is in state j. MMPP was first proposed by Yechiali and Naor to model non-homogeneous Poisson arrival process in queueing systems [Yechiali and Naor 1971]. Compared with traditional Poisson process, MMPP allows the arrival rates to vary from time to time, making the model more flexible. Besides, MMPP is effective to capture burst arrivals and sudden changes in arrivals, since it can integrate significantly different rates into one model. All these benefits make MMPP a widely applied model for the arrival processes in networks [Choi et al. 2008; Gharehajlu et al. 2015], for the processes that show pattern changes [Csóka and Polec 2015], and for burst events detection [Scott 2000; Ihler et al. 2007].

The good properties and the broad applications of MMPP are all on the basis of the temporal dependence carried by MMPP. Essentially, the dependence/correlation among inter-arrival times is the main difference between MMPP and Poisson process [Begain et al. 2012]. In the model of MMPP, the inter-arrival times are not independent. The dependence between inter-arrival times comes from CTMC that modulates the state switches over time. With the dependence structure of MMPP, we can better understand the process and predict its trend. For example, when we model or detect traffic of networks with MMPP, the temporal dependence of MMPP can be the objective to match with that of the real traffic trace. Another example is to model web traffic or traffic in cloud with MMPP [Rajabi and Wong 2014, 2012; Pacheco-Sanchez et al. 2011]. In this case, resource provisioning based on MMPP arrivals is the problem of interest. The capability of predicting arrivals based on temporal dependence structures is critical in designing the resource provisioning policy.

Existing theoretical studies of MMPP mainly fall into two categories. One track of studies is to use MMPP as the input of the queueing system and study the queueing performance. Current representative works include [Choi et al. 2008; Rajabi and Wong 2014]. The other track of studies is to develop algorithms to estimate the parameters of MMPP. Recent developments cover the algorithms of fitting MMPP to IP traffic traces [Bali and Frost 2007], the expectation-maximization (EM) based algorithms to learn MMPP as a type of Markovian Arrival Process [Okamura et al. 2009], the algorithms to learn MMPP through the detection of change points along with the arrival rates estimation [Burkatovskaya et al. 2015], and the online learning algorithms by modeling MMPP as a Hidden Markov Model [Chis and Harrison 2015]. All these learning algorithms are either based on the arrival times or the number of arrivals within every unit of time (arrival counts). Despite the abundant existing theoretical results on MMPP, there still is a large gap in the formal analysis of the dependence structure of MMPP in the literature. This gap is reflected in the following aspects.

First, the temporal dependence of MMPP is not well understood. The existing results related to the MMPP temporal dependence are all on the basis of covariance/autocorrelation. Neuts derived covariance between arrival counts over any two time slots for stationary MMPP in 1989 [Neuts 1989], which is still the strongest result known. This covariance result is not sufficient for many applications. To begin with, the covariance is not easy to compute due to the matrix exponential and matrix inverse involved (especially when the number of states becomes large). For two-state stationary MMPP, the closed-form of the covariance between arrival counts was given by Andersen and Nielsen [1998]. For multi-state MMPPs, their covariances are usually obtained approximately by statistical counting on simulated traces [Pacheco-Sanchez et al. 2011; Chis and Harrison 2015; Andersen and Nielsen 1998]. Furthermore, the covariance or the autocorrelation is only capable of measuring the *linear* dependence degree over time. However, the MMPP network traffic may contain temporal dependence more complex than linear dependence. Through a detailed example given in Section 3.2, it is clear that the covariance only captures MMPP dependence structure partially and is far from reflecting its whole dependence structure. This motivates us to search for the exact and functional temporal dependence structure of MMPPs.

Second, there is no analysis on the temporal dependence in the superposition of MMPPs, that is, the aggregation of multiple flows, each modeled as an MMPP [Heyman and Lucantoni 2003]. Although it has been proved that the superposition of MMPP is still an MMPP [Fischer and Meier-Hellstern 1993], analysing the superposition of MMPP becomes intractable in real applications due to the exponential increase of the number of states. For instance, the superposition of two 20-states MMPP is computationally expensive to solve [Heyman and Lucantoni 2003]. In other words, simply treating the superposition of MMPP as one MMPP of higher number of states would not work in practice. The temporal dependence of superposition of MMPPs thus requires a different analytical method.

Copula, an advanced dependence measure that links marginals into joint distributions, is ideal for modeling the temporal dependence of MMPP. First, copula can be constructed theoretically based on the analysis on marginals and joint distributions of the observations in MMPP. Second, copula is capable of capturing all the characteristics from dependence structures. Beyond the linear dependence, it characterizes functional dependence structure and carries abundant dependence information. Third, with the help of copula, it is easy to avoid the explosion of the number of parameters when modeling the superposition of MMPPs, and it is computationally tractable to calculate the temporal dependence of superposed MMPPs. Finally, the invariant property of copula keeps the dependence measure stable even when MMPP trace changes functionally.

Using copula analysis, this article builds the theoretical foundation for the temporal dependence structure of MMPP and makes the following contributions:

(1) It uses a new dependence measure, copula, to analyse the dependence structure of MMPP traffic. The copula-based dependence reveals richer information of temporal dependence and is more powerful than the commonly used measures, covariance and correlation.
(2) It gives the exact form of temporal dependence of MMPP with arbitrary number of states. This is the first theoretical result on the functional temporal dependence of multi-state MMPP. The result could be applied for trend prediction or designing learning algorithms for parameter estimation.
(3) It offers a solution to deal with superposition of homogeneous MMPPs. In this case, we not only develop a recursive algorithm to compute the exact values of the temporal dependence but also introduce a speed-up method that adopts parametric copulas for the approximation of theoretical results.
(4) It demonstrates an application that uses copula-based MMPP dependence for traffic prediction.

2. RELATED WORK

The Markov Modulated Poisson Process (MMPP) was first applied in the network domain in 1971 [Yechiali and Naor 1971]. Since then, tremendous research efforts have been devoted to MMPP. Early theoretical results and applications of MMPP were outlined in the review by Fischer and Meier-Hellstern [1993] and references therein. In brief, the review includes the theoretical results of the characterization of MMPP, the statistical moments of MMPP arrivals, and the superposition of independent MMPPs. Afterwards, MMPP was further studied as the arrival input of queueing systems. Furthermore, various learning algorithms for parameter estimation of MMPP were proposed. In the following, we summarize the research progress on MMPP after the review article by Fischer and Meier-Hellstern [1993].

MMPP has been broadly applied as arrival input for queueing system. Baiocchi and Blefari-Melazzi derived the mean waiting time of MMPP/G/1/K queue on steady state [Baiocchi and Blefari-Melazzi 1993]. Du analysed the loss probability of MMPP(3)/M(3)/1/0 queue, where the service rate is modulated by the same CTMC that modulates the arrivals [Du 1995]. Choi et al. considered the MMPP/G/1/K queue with arrival rate dependent on queue length, and derived the loss probability, the mean queue length, and the mean waiting time of the model [Choi et al. 2008]. When the input flow is a superposition of MMPPs, the performance analysis of the queueing system is faced with the challenge that the number of states explodes exponentially. To simplify the computation, Giacomazzi analysed the envelope of MMPP arrivals and integrated the MMPP envelope into the framework of stochastic network calculus for performance analysis [Giacomazzi 2007, 2009].

The parameter estimation problem of MMPP has been studied for several decades. According to different fitting objectives, the traditional estimation algorithms can be mainly categorized into two groups: the maximum likelihood estimation (MLE) algorithms and the moment-based algorithms. The former type was shown to achieve consistent results [Rydén 1994]. The MLE-based algorithms were implemented via expectation-maximization (EM) algorithm in Rydén [1996]. Rydén's EM algorithm for estimating MMPP in Rydén [1996] was further enhanced to ease the difficulty of calculating integrals [Roberts et al. 2006; Elliott and Malcolm 2008] and to estimate parameters from observations of either arrival times or arrival counts [Breuer and Kume 2010; Elliott and Malcolm 2008].

The moment-based algorithms learn MMPP parameters by finding the moments, such as marginal moments and autocovariance. Compared with the MLE-based algorithms, the moment-based algorithms are usually fast and emphasize more on emulating specific dependence structures of real traces. For instance, moment-based algorithms were broadly used to emulate the self-similarity or long-range dependence (LRD) of network traffic [Andersen and Nielsen 1997, 1998; Yoshihara et al. 2001; Kasahara 2001; Salvador et al. 2003; Shao et al. 2005]. The superposition of two-state MMPPs was shown to be capable of modeling the self-similarity [Andersen and Nielsen 1997]. They constructed a high-dimensional MMPP with superposition of 2-MMPPs, because the moment of a two-state MMPP is easy to compute. Following this idea, the superposition of 2-MMPPs has been used to model the self-similarity or LRD of network traffic traces by matching their asymptotic covariances [Andersen and Nielsen 1998] or exact variances over different time scales [Kasahara 2001; Salvador et al. 2003; Shao et al. 2005; Yoshihara et al. 2001]. The learned MMPP parameters were integrated into queueing theory to predict the queueing performance.

In addition to the above two main categories, other fitting algorithms have also been developed. Algorithms were developed to fit IP traces into discrete MMPP by assuming that the Poisson arrivals of each state fall into certain range of variation [Bali and

Frost 2007; Heyman and Lucantoni 2003]. A Bayesian learning algorithm based on the posterior probability was developed to model and detect the bursty events [Ihler et al. 2007]. The most recent algorithm learns MMPP parameters by first detecting the points of state switching and then estimating the arrival rates at the corresponding state [Burkatovskaya et al. 2015].

Among all the literature, the research that is most related to our work is summarized as follows. MMPP and other Markovian arrival processes were generalized into the versatile Markov point process in Neuts [1989]. The covariance between arrival counts was derived for the versatile Markov point process. The closed form of the covariance of 2-state MMPP was given in Andersen and Nielsen [1998]. The covariance was further derived into an asymptotic form and used for learning parameters. In Pacheco-Sanchez et al. [2011], Chis and Harrison [2015], and Andersen and Nielsen [1998], covariance was the evaluation metric for goodness of fit test for MMPP, and it was computed empirically from simulated trace of fitted MMPP. Different from the above works, our article derives the theoretical results on temporal dependence of MMPP in terms of copula, which represents functional dependence.

In the case of superposition of MMPPs, its mathematical form has been given in Fischer and Meier-Hellstern [1993], but the parameter computation of superposed MMPPs is complex due to the explosion of state number. To reduce the computational complexity, recent efforts have been made to reduce the number of states and obtain an approximate solution [Heyman and Lucantoni 2003; Yu and Zhou 2006]. Our article, in contrast, focuses on the exact and tractable solution of temporal dependence in the superposition of homogeneous MMPPs.

Copula models have been broadly used in the domain of financial analysis, for multivariate dependence modeling [Beil 2013; Embrechts et al. 2003] as well as for time series modeling [Rémillard et al. 2012; Patton 2012]. In the recent decade, copula has also been used for dependence modeling in the domain of computer networks [Dong et al. 2015, 2016; Neuhäuser et al. 2013; Markovich 2008, 2010; Avrachenkov et al. 2015]. To the best of our knowledge, this article is the first that takes advantage of copula models to tackle the challenging analysis of the temporal dependence of MMPP.

3. PRELIMINARIES

3.1. Markov Modulated Poisson Process

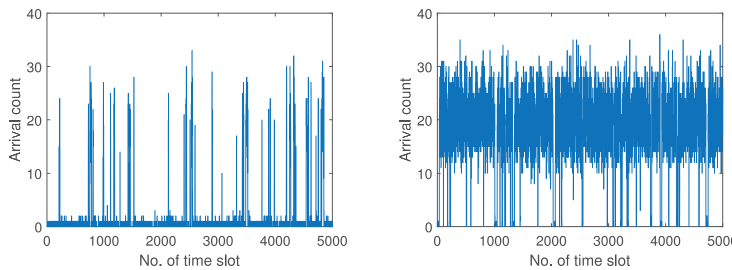
We introduce the definition and key concepts of MMPP.

Definition 3.1. A **Markov-modulated Poisson Process** (MMPP) [Fischer and Meier-Hellstern 1993] is constructed by varying the arrival rate of a Poisson process according to an m-state irreducible continuous-time Markov chain (CTMC). In particular, when the Markov Chain is in state j, the arrivals follow a Poisson process of rate λ_j. Therefore, an MMPP can be parameterized by the transition rate matrix Q [Ross 2003] of CTMC and the m Poisson arrival rates, $\Lambda = (\lambda_1, \dots, \lambda_m)$.

We thus denote an MMPP by parameters (Q, Λ).

Definition 3.2. **Environment-stationarity** of an MMPP [Fischer and Meier-Hellstern 1993]: An MMPP (Q, Λ) is considered to be environment-stationary if its associated CTMC with parameter Q is stationary.

For an environment-stationary MMPP, the stationary distribution of the states, $\Pi = (\pi_1, \dots, \pi_m)$, is determined by solving the equation $\Pi Q = 0$. In this article, we **only consider environment-stationary MMPP.**



(a) Trace 1 (b) Trace 2

Fig. 1. Arrival counts of the two traces.

Definition 3.3 (*Superposition of independent homogeneous MMPPs*). An MMPP is denoted as **l-IHMMPP** if it is a superposition of l independent homogeneous MMPPs all with the same parameters (Q, Λ).

We consider the superposition of homogeneous MMPPs. The constituent MMPPs of an l-IHMMPP all have the same dependence structure. This nice feature makes it possible to design a fast algorithm to compute the dependence of an l-IHMMPP based on the dependence inside the constituent MMPPs. Modeling the temporal dependence in the superposition of non-homogeneous MMPPs is left as an open challenge.

Since l-IHMMPP is still an MMPP, to distinguish regular MMPP with l-IHMMPP, in this article MMPP not created from superposition is termed as single MMPP by default.

3.2. Why Do Existing Results Not Suffice?

The strongest result so far that discloses the temporal dependence of MMPP is from Neuts [1989], where the covariance or the autocorrelation of arrival counts over different time slots is given. Covariance, however, is only capable of capturing linear dependence. MMPP trace may contain temporal dependence much more complex than linear dependence. To illustrate the pitfalls of covariance, we consider two MMPPs with their parameters as (Q_1, Λ_1) and (Q_2, Λ_2) as follows:

$$Q_1 = \begin{pmatrix} -0.1 & 0.1 \\ 1 & -1 \end{pmatrix}, \Lambda_1 = (2, 200); \qquad Q_2 = \begin{pmatrix} -0.1 & 0.1 \\ 1 & -1 \end{pmatrix}, \Lambda_2 = (200, 2).$$

We simulate traces from these two MMPPs: Trace 1 is from MMPP (Q_1, Λ_1); Trace 2 is from MMPP (Q_2, Λ_2). The traces are analysed by arrival counts. Specifically, the number of arrivals in ith timeslot is denoted as $A_i (i \in \mathbb{N})$. The arrival counts of the two traces are shown in Figure 1. From the figure, the traces from the two MMPPs are very different. To study their dependence, we first analyse the covariances of two MMPPs and then visualize their temporal dependence by the joint distribution of successive arrival counts.

The theoretical form of the covariance of a two-state MMPP is given by Equation (3) in Section 2 of Andersen and Nielsen [1998]. Based on the given covariance function, the covariances between A_i and $A_{i+i'} (i' \in \mathbb{N},$ is the time lag) of the two MMPPs in this example can be shown to be theoretically the same. In Figure 2, we use the green plots to show that the theoretical covariances of the two MMPPs (from the theoretical analysis with Equation (3) of Andersen and Nielsen [1998]) are all the same over different time lags. We also plot the empirical covariances from the simulated traces. The covariances of two traces are close, though they vary slightly from the theoretical results. So in terms of covariance, the two MMPPs show the same dependence structure.

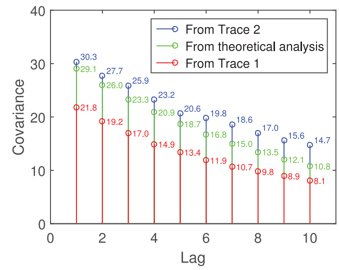
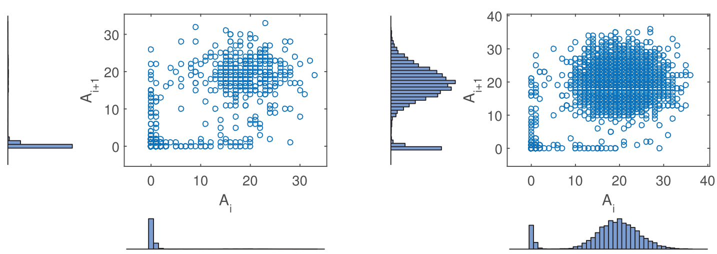
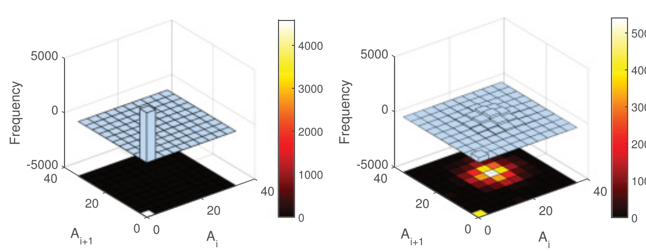


Fig. 2. Covariances of two MMPPs over different time lags.



(a) Trace 1 (b) Trace 2

Fig. 3. Scatter plot with marginal histograms of A_i and A_{i+1} in two traces.



(a) Trace 1 (b) Trace 2

Fig. 4. Bivariate frequency histogram (upper layer) with its heat map (lower layer).

To obtain the full view of the dependence structure, we visualize the joint behaviour of A_i and A_{i+1} by the scatter plots with marginal histograms in Figure 3 and their bivariate frequency histograms with heat map in Figure 4. From the two figures, we can observe that the joint behaviour of two successive arrival counts is quite different in the two MMPPs. Therefore, it is clear that the two MMPPs have different temporal dependence between A_i and A_{i+1}.

In the above simple example, the two MMPPs have the same covariance theoretically. However they generate traces with significantly different temporal dependence structure. Therefore, covariance, only measuring partial information from dependence, is not sufficient to represent MMPP dependence. This motivates us to seek a better dependence structure to characterize temporal dependence beyond linear scope when modeling network traffic with MMPP. We tackle this challenge with copula analysis.

3.3. Copulas

A copula is a function that links univariate marginals to their multivariate distribution. The definition of two-copula is:

Definition 3.4 (*Copula*). A two-dimensional copula is a function C having the following properties [Nelson 2006]:

(1) Its domain is $[0, 1]^2$;
(2) C is 2-increasing, that is, for every $u_1, u_2, v_1, v_2 \in [0, 1]$ and $u_1 \leq u_2, v_1 \leq v_2$, we have $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$.
(3) $C(u, 0) = C(0, v) = 0$, $C(u, 1) = u$, $C(1, v) = v$, for every $u, v \in [0, 1]$.

The function is called a subcopula if it has the second and the third properties of copula, but its domain is $b_1 \times b_2$, where b_1, b_2 are subsets of $[0, 1]$ containing 1 and 0.

THEOREM 3.5 (SKLAR'S THEOREM) [NELSON 2006]. *Let H be a joint distribution function with marginals F_X and F_Y, then there exists a copula C such that for all x and y,*

$$H(x, y) = Pr(X \leq x, Y \leq y) = C(F_X(x), F_Y(y)).$$

If the marginals F_X and F_Y are continuous, then copula C is unique; otherwise, C is uniquely determined on the range of the marginals. Sklar's theorem is the core of the copula theory. On one hand, it shows how the copula connects marginals with joint distribution. This property is especially useful, since the joint distribution of random variables is hard to find directly in many applications [Bína and Jiroušek 2013; Nelson 2006]. In this situation, integration of a copula model and marginals makes it easy to understand the joint behaviour. On the other hand, Sklar's theorem implies that copula, as a dependence measure, is entirely separated from both marginals and joint distribution. The dependence in terms of copula is stable when the marginals change functionally. This beautiful feature is termed as the invariant property of copulas [Nelson 2006] and is formally stated in the following theorem:

THEOREM 3.6 (THE INVARIANT PROPERTY OF COPULAS) [NELSON 2006]. *Let X and Y be continuous random variables with copula C_{XY}. If α_1 and α_2 are strictly increasing functions on the range of X and the range of Y, respectively, then $C_{\alpha_1(X)\alpha_2(Y)} = C_{XY}$. In other words, C_{XY} is invariant under strictly increasing transformations of X and Y.*

There are mainly two methods to build a copula model. The first method is to construct a theoretical copula for the problem at hand. The inversion method belongs to this category.

THEOREM 3.7 (INVERSION METHOD) [NELSON 2006]. *Let H be a joint distribution function with marginals F_X and F_Y. Let F_X^{-1} and F_Y^{-1} be the inverse function of F_X and F_Y. Then the copula between X and Y can be constructed as*

$$C(u, v) = H\big(F_X^{-1}(u), F_Y^{-1}(v)\big) \quad \forall u, v,$$

such that

$$H(x, y) = C(F_X(x), F_Y(y)) \quad \forall x, y.$$

The inversion method essentially constructs copula from the inverse of Sklar's theorem. The constructed copula from inversion method is the theoretical copula between X and Y that satisfies Sklar's theorem. The inversion method leads to a unique copula when the marginals are continuous, and leads to a unique subcopula when the marginals are not continuous. The unique subcopula can be easily extended to a copula via various ways, for instance, bilinear interpolation. Thus, a subcopula shares most properties of a copula. In the following context, we do not differentiate between subcopula and copula, because their difference does not impact the analytical results in our application context.

The other method to build a copula model is to fit real data into known parametric copulas. A large variety of parametric copulas are available for parametric copula modeling, for instance, Gaussian copula, Student's copula, Clayton copula, Frank copula, and Gumbel copula. Since the theoretical copula is not always easy to derive, parametric copula modeling has become popular in practice [Trivedi and Zimmer 2007; Genest et al. 2009; Azam and Pitt 2014; Chen et al. 2009].

Copula-based dependence has three main associated dependence measures. They are Kendall's tau ρ_τ, Spearman's rho ρ_s, and tail dependence ρ_t. Let (X_i, Y_i), (X_j, Y_j) and (X_k, Y_k) denote different observations from a vector (X, Y) of continuous random variables with copula between them as $C(u, v)$. ρ_τ and ρ_s can be calculated from either data or copula with Equations (1) and (2), respectively. These two dependence values measure the concordance between random variables. Larger values represent larger concordance. Tail dependence comprises of upper tail dependence as shown in Equation (3) and lower tail dependence as shown in Equation (4). The tail dependence shows the probability that two random variables achieve extremely large (or small) values simultaneously:

$$\rho_\tau = Pr((X_i - X_j)(Y_i - Y_j) > 0) - Pr((X_i - X_j)(Y_i - Y_j) < 0) = 4\int_0^1\int_0^1 C(u, v)dC(u, v) - 1, \tag{1}$$

$$\rho_S = 3(Pr((X_i - X_j)(Y_i - Y_k) > 0) - Pr((X_i - X_j)(Y_i - Y_k) < 0)) = 12\int_0^1\int_0^1 C(u, v)dudv - 3, \tag{2}$$

$$\rho_t^+ = \lim_{u\to 1} Pr\big(X > F_X^{-1}(u)\big|Y > F_Y^{-1}(u)\big) = \lim_{u\to 1}\frac{1 - 2u + C(u, u)}{1 - u}, \tag{3}$$

$$\rho_t^- = \lim_{u\to 0} Pr\big(X < F_X^{-1}(u)\big|Y < F_Y^{-1}(u)\big) = \lim_{u\to 0}\frac{C(u, u)}{u}. \tag{4}$$

With all these copula-based dependence measures, the temporal dependence of MMPP can be clearly illustrated. Since ρ_τ, ρ_s and ρ_t are all dependence measurements in terms of copula, they all follow the invariant property and remain stable under increasing functional transformation.

To recap, this section briefly introduces copulas, and explains why copula is promising to model the temporal dependence of MMPP: copula of both single MMPP and l-IHMMPP can be constructed theoretically or by parametric copula modeling; the invariant property keeps copula structure stable under increasing functional transformation; its associated dependence measures are rich to show dependence from different aspects. In this article, we will study both theoretical copula (Section 4) and parametric copula of MMPPs (Section 5).

4. THEORETICAL COPULA ANALYSIS FOR MMPPS

4.1. Theoretical Copula for Single MMPP

We first study an m-state MMPP with parameters (Q, Λ). To ease the application of copula model in our context, we divide the time into equal-sized small intervals, called time slots. We assume the length of each time slot is Δ, which is short enough such that the state transition of MMPP within one time slot is negligible.[1] In other words, we assume that the state transitions occur only at the boundaries of time slots. This approximation has been used in previous research, for example, in Okamura et al. [2009].

Denote the sequence of time slots as $I_1, I_2, \ldots, I_n$, the state in slot I_i as S_i, and the number of arrivals in I_i as A_i. The sequence $\{A_i\}$ is also called arrival counts as introduced in Section 1. Denote the transition matrix by $P(t) = [p_{j_1 j_2}(t)]$, where $p_{j_1 j_2}(t)$ is the probability that the CTMC switches from state j_1 to state j_2 after time t. $P(t) = e^{Qt}$ can be calculated with numerical methods such as those introduced in Chapter 6.8 of Ross [2003]. As Δ is small, $P(\Delta)$ relates to Q matrix in the following way:

$$p_{j_1 j_2}(\Delta) = 1 + q_{j_1 j_2}\Delta + o(\Delta), \quad j_1 = j_2; \qquad p_{j_1 j_2}(\Delta) = q_{j_1 j_2}\Delta + o(\Delta), \quad j_1 \neq j_2.$$

Therefore, by a simple calculation, $P(\Delta) \approx E + Q\Delta$, where E is an identity matrix. In the following, we derive the marginal distribution function of single MMPP as M_1 and the copula for single MMPP as C_1.

THEOREM 4.1. *The marginal distribution of A_i is*

$$M_1(x) \equiv Pr(A_i \leq x) = \sum_{j=1}^{m} \pi_j G_j(x), \tag{5}$$

where $G_j(x) \equiv Pr(A_i \leq x | S_i = j) = e^{-\lambda_j \Delta} \sum_{k=0}^{k=x} \frac{(\lambda_j \Delta)^k}{k!}$

PROOF.

$$M_1(x) \equiv Pr(A_i \leq x) = \sum_{j=1}^{m} Pr(A_i \leq x | S_i = j) Pr(S_i = j) = \sum_{j=1}^{m} \pi_j G_j(x). \quad \square$$

THEOREM 4.2 (MMPP COPULA). *The copula of A_i and A_{i+1} can be calculated as*

$$C_1(u, v) = \mathcal{G}\big(M_1^{-1}(u)\big) diag(\Pi) P(\Delta) \mathcal{G}\big(M_1^{-1}(v)\big)^T, \tag{6}$$

where

—$\mathcal{G}(x) \equiv [G_1(x), \ldots, G_m(x)]$ *is a vector,*
—M_1^{-1} *is the inverse function of* M_1 *defined by Equation (5),*
—$diag(\Pi)$ *is a square diagonal matrix with the elements of vector* Π *on the main diagonal,*
—$\mathcal{G}(M_1^{-1}(v))^T$ *is the transpose of* $\mathcal{G}(M_1^{-1}(v))$

[1]To keep this assumption valid, we recommend the length of time slot to be no larger than the smallest average time of MMPP staying on one state, that is, $\Delta \leq \frac{1}{\max_i |q_{ii}|}$. Under this condition, few state transitions will happen in one time slot, so the arrival rate in one time slot is (approximately) stable. Experiments [Okamura et al. 2009] have showed that the parameter estimation based on arrival counts becomes inaccurate when $\Delta > \frac{1}{\max_i |q_{ii}|}$, indicating that a large value of Δ would make the arrival counts lack enough information to retrieve the MMPP.

Proof. The joint distribution of A_i and A_{i+1} is derived as

$$
\begin{aligned}
&H(x, y) \equiv Pr(A_i \le x, A_{i+1} \le y)\\
&= \sum_{z=0}^{x} Pr(A_i = z) Pr(A_{i+1} \le y | A_i = z)\\
&= \sum_{z=0}^{x}\sum_{j_2=0}^{m}\sum_{j_1=0}^{m} Pr(A_i = z) Pr(S_i = j_1 | A_i = z) Pr(S_{i+1} = j_2 | S_i = j_1) Pr(A_{i+1} \le y | S_{i+1} = j_2)\\
&= \sum_{z=0}^{x}\sum_{j_2=0}^{m}\sum_{j_1=0}^{m} Pr(S_i = j_1) Pr(A_i = z | S_i = j_1) Pr(S_{i+1} = j_2 | S_i = j_1) Pr(A_{i+1} \le y | S_{i+1} = j_2)\\
&= \sum_{j_2=0}^{m}\sum_{j_1=0}^{m} Pr(S_i = j_1) Pr(A_i \le x | S_i = j_1) Pr(S_{i+1} = j_2 | S_i = j_1) Pr(A_{i+1} \le y | S_{i+1} = j_2)\\
&= \sum_{j_2=0}^{m}\sum_{j_1=0}^{m} G_{j_2}(y) p_{j_1 j_2}(\Delta) G_{j_1}(x) \pi_{j_1}\\
&= \mathcal{G}(x) diag(\Pi) P(\Delta) \mathcal{G}(y)^T.
\end{aligned}
$$

With the inverse method based on Theorem 3.7, the copula between A_i and A_{i+1} is constructed as

$$
\begin{aligned}
C_1(u, v) &= H\big(M_1^{-1}(u), M_1^{-1}(v)\big)\\
&= \mathcal{G}\big(M_1^{-1}(u)\big) diag(\Pi) P(\Delta) \mathcal{G}\big(M_1^{-1}(v)\big)^T. \quad \square
\end{aligned}
$$

The copula C_1 in Theorem 4.2 is the theoretical copula that models the dependence of arrival counts of single MMPP in two consecutive time slots. We refer to this theoretical copula as *MMPP copula*.

4.2. Theoretical Copula for *l*-IHMMPP

In many network systems, such as Internet core routers, the incoming traffic may be an aggregation (superposition) of multiple independent homogeneous MMPPs (i.e., l-IHMMPP). We denote the arrival counts of l-IHMMPP within each time slot Δ as A_i^l. We define the marginal distribution of A_i^l as M_l, the copula between A_i^l and A_{i+1}^l as C_l. It is hard to derive the theoretical copula for l-IHMMPP with the inversion method when l is large, due to the difficulty of obtaining the joint distribution of its arrival counts. To tackle this difficulty, we develop a recursive and efficient algorithm to calculate the copula values. Although the solution is not expressed in closed-form, the algorithmic approach indeed finds the exact solution.

The following theorems are required in our algorithm.

Theorem 4.3. *For l-IHMMPP, the marginal of A_i^l is*

$$
\begin{aligned}
M_l(x) = \sum_{l_1+\cdots+l_m=l} & \binom{l_1}{l}\binom{l_2}{l-l_1}\cdots\binom{l_m}{l-l_1-\cdots-l_{m-1}}\\
& * \pi_1^{l_1}\pi_2^{l_2}\cdots\pi_m^{l_m} * Po((l_1\lambda_1+\cdots+l_m\lambda_m)\Delta, x),
\end{aligned} \tag{7}
$$

where $\binom{k}{n}$ is the combinatorial number of choosing k from n, and $Po(\lambda, x)$ represents the Poisson cumulative distribution of value x, with parameter λ.

PROOF. Assume the number of MMPP in State j is l_j ($j = 1, 2, \ldots, m$). The probability that the l-IHMMPP is at the above allocation of states is $\binom{l_1}{l}\binom{l_2}{l-l_1}\cdots\binom{l_m}{l-l_1-\cdots-l_{m-1}} * \pi_1^{l_1}\pi_2^{l_2}\cdots\pi_m^{l_m}$. Since the superposition of two Poisson processes with rate λ_1 and λ_2, respectively, is a Poisson process with rate $\lambda_1+\lambda_2$, the aggregate arrivals within one time slot of l-IHMMPP under the assumed state combination follow Poisson distribution with parameter $(l_1\lambda_1 + \cdots + l_m\lambda_m)\Delta$. Adding up all the possible allocations of states leads to the marginal form in Theorem 4.3. □

THEOREM 4.4. *The copula for A_i^l and A_{i+1}^l in l-IHMMPP ($l \geq 2$) can be calculated recursively:*

$$C_l(M_l(x), M_l(y)) = \sum_{x'=0}^{x}\sum_{y'=0}^{y} \nabla C_1(M_1(x'), M_1(y')) * C_{l-1}(M_{l-1}(x-x'), M_{l-1}(y-y')), \quad (8)$$

where $\nabla C_1(M_1(x'), M_1(y'))$ is called MMPP-copula gradient, which can be calculated as

$$\begin{aligned}\nabla C_1(M_1(x'), M_1(y')) \;=\; & C_1(M_1(x'), M_1(y')) + C_1(M_1(x'-1), M_1(y'-1)) \\ & - C_1(M_1(x'), M_1(y'-1)) - C_1(M_1(x'-1), M_1(y')). \quad (9)\end{aligned}$$

PROOF. Since the constituent MMPPs are mutually independent, the arrivals of l-IHMMPP can be divided into arrivals of $(l-1)$-IHMMPP and arrivals of a single MMPP, that is, $A_i^l = A_i^{l-1} + A_i$. Following this idea, we have

$$\begin{aligned}& C_l(M_l(x), M_l(y)) \\ &= Pr\left(A_i^l \leq x, A_{i+1}^l \leq y\right) \\ &= \sum_{x'=0}^{x}\sum_{y'=0}^{y} Pr\left(A_i^l \leq x, A_{i+1}^l \leq y \middle| A_i = x', A_{i+1} = y'\right) * Pr(A_i = x', A_{i+1} = y') \\ &= \sum_{x'=0}^{x}\sum_{y'=0}^{y} Pr\left(A_i^{l-1} \leq x - x', A_{i+1}^{l-1} \leq y - y'\right) * Pr(A_i = x', A_{i+1} = y') \\ &= \sum_{x'=0}^{x}\sum_{y'=0}^{y} C_{l-1}(M_{l-1}(x-x'), M_{l-1}(y-y')) * Pr(A_i = x', A_{i+1} = y').\end{aligned}$$

Since the arrival counts are non-negative integer random variables, for all non-negative integers x' and y', we have

$$\begin{aligned}& Pr(A_i = x', A_{i+1} = y') \\ &= Pr(A_i \leq x', A_{i+1} \leq y') + Pr(A_i \leq x'-1, A_{i+1} \leq y'-1) \\ &\quad - Pr(A_i \leq x', A_{i+1} \leq y'-1) - Pr(A_i \leq x'-1, A_{i+1} \leq y') \\ &= C_1(M_1(x'), M_1(y')) + C_1(M_1(x'-1), M_1(y'-1)) \\ &\quad - C_1(M_1(x'), M_1(y'-1)) - C_1(M_1(x'-1), M_1(y')) \\ &= \nabla C_1(M_1(x'), M_1(y')),\end{aligned}$$

and $C_1(M_1(x'), M_1(y')) = 0$ if any of x' and y' is negative.

By replacing $Pr(A_i = x', A_{i+1} = y')$ with $\nabla C_1(M_1(x'), M_1(y'))$, we can calculate the copula C_l as

$$C_l(M_l(x), M_l(y)) = \sum_{x'=0}^{x}\sum_{y'=0}^{y} \nabla C_1(M_1(x'), M_1(y')) * C_{l-1}(M_{l-1}(x-x'), M_{l-1}(y-y')). \quad □$$

Table I. Definition of Two Matrices

	Number in row x and column y
$[\mathcal{O}^l]_{\hat{a}\times\hat{a}}$	$\mathcal{O}^l_{xy} \equiv C_l(M_l(x-1), M_l(y-1))$
$[\mathcal{R}^l]_{\hat{a}\times\hat{a}}$	$\mathcal{R}^l_{xy} \equiv \nabla C_l(M_l(x-1), M_l(y-1))$

Theorem 4.4 actually defines a recursive function. Even with the recursive function in Theorem 4.4, it is still difficult to obtain the closed-form copula C_l. Fortunately, the value of C_l can be calculated numerically with a recursive algorithm. To reduce the running time of the algorithm, we narrow down the range of A^l_i from its infinite domain to finite range with an upper threshold $\hat{a}$. In other words, although the range of A^l_i is on the whole non-negative integer domain, we only need to compute the copula values $C_l(M_l(x), M_l(y))$ for $x < \hat{a}$ and $y < \hat{a}$. The selection of $\hat{a}$ is application dependent and can be set appropriately based on observations. Narrowing down the range makes the computation feasible and still fulfils the demand for real applications, because the arrival counts in real traffic flows always fall within a limited range.

On the range $[0, \hat{a})$, we define two matrices with dimension $\hat{a} \times \hat{a}$ in Table I to represent the values of copula C_l and its gradient ∇C_l, respectively. The matrix $\mathcal{O}^l$ is a look-up table for copula values of l-IHMMPP. We call it as *l-IHMMPP copula matrix*. Similarly, the matrix $\mathcal{R}^l$ is called *l-IHMMPP copula gradient matrix*. The notation $[\mathcal{O}^l]_{\hat{a}\times\hat{a}}$ and $[\mathcal{R}^l]_{\hat{a}\times\hat{a}}$ are used to emphasize the matrices' dimension. We also use a notation $[\mathcal{O}^l]_{x\times y}$ to represent the submatrix of $[\mathcal{O}^l]_{\hat{a}\times\hat{a}}$ with its first x rows and first y columns. In addition, $[\mathcal{R}^l]_{x\times y}$ is used similarly for $[\mathcal{R}^l]_{\hat{a}\times\hat{a}}$.

Algorithm 1 computes the l-IHMMPP copula matrix $\mathcal{O}^l$ and MMPP copula gradient matrix $\mathcal{R}^1$, with parameters of constituent MMPP (Q, Λ). The time complexity of Algorithm 1 is $O(\hat{a} \times \hat{a} \times l)$. A recursive procedure, CpaMatCal, which implements both Theorems 4.2 and 4.4, is embedded in the algorithm. Some details of this procedure are described below:

Remark 4.5. In line 6, the elements of the matrix $[\mathcal{O}^1]_{\hat{a}\times\hat{a}}$ are computed based on the MMPP-copula in Theorem 4.2 by

$$\mathcal{O}^1_{xy} = C_1(M_1(x-1), M_1(y-1)) = \mathcal{G}(x-1)diag(\Pi)P(\Delta)\mathcal{G}(y-1)^T.$$

Remark 4.6. In line 7, the elements of the matrix $[\mathcal{R}^1]_{\hat{a}\times\hat{a}}$ are calculated from $[\mathcal{O}^1]_{\hat{a}\times\hat{a}}$ in the following way:

(1) $\mathcal{R}^1_{xy} = \mathcal{O}^1_{xy}$ if $x = 1$ and $y = 1$,
(2) $\mathcal{R}^1_{xy} = \mathcal{O}^1_{xy} - \mathcal{O}^1_{(x-1)y}$ if $x \neq 1$ and $y = 1$,
(3) $\mathcal{R}^1_{xy} = \mathcal{O}^1_{xy} - \mathcal{O}^1_{x(y-1)}$ if $x = 1$ and $y \neq 1$,
(4) $\mathcal{R}^1_{xy} = \mathcal{O}^1_{xy} + \mathcal{O}^1_{(x-1)(y-1)} - \mathcal{O}^1_{x(y-1)} - \mathcal{O}^1_{(x-1)y}$ if $x \neq 1$ and $y \neq 1$.

Integrating Theorem 4.3 with Theorem 4.4, we have the following theorem for l-IHMMPP copula:

THEOREM 4.7 (l-IHMMPP COPULA). *For any u, v in $[0, 1]$, if $M_l^{-1}(u) < \hat{a}$ and $M_l^{-1}(v) < \hat{a}$,*

$$C_l(u, v) = \mathcal{O}^l_{M_l^{-1}(u)M_l^{-1}(v)}. \tag{10}$$

Using the analysis in this section, we can calculate the copula for l-IHMMPP as shown in Theorem 4.7. Although mathematically it is not in closed-form, the copula values can be computed numerically. Therefore, the marginal M_l analysed in Theorem 4.3 and the l-IHMMPP copula matrix $\mathcal{O}^l$ together can be regarded as the theoretical copula results for l-IHMMPP.

ALGORITHM 1: An Algorithm to Compute the l-IHMMPP Copula Matrix $\mathcal{O}^l$ and MMPP Copula Gradient Matrix $\mathcal{R}^1$

Require: MMPP parameters Λ, Q, the value of l, the upper threshold $\hat{a}$
Ensure: $[\mathcal{O}^l]_{\hat{a}\times\hat{a}}$ and $[\mathcal{R}^1]_{\hat{a}\times\hat{a}}$
$([\mathcal{O}^l]_{\hat{a}\times\hat{a}}, [\mathcal{R}^1]_{\hat{a}\times\hat{a}}) \leftarrow$ CpaMatCal(Λ, Q, l, $\hat{a}$)
return $[\mathcal{O}^l]_{\hat{a}\times\hat{a}}$ and $[\mathcal{R}^1]_{\hat{a}\times\hat{a}}$

procedure CpaMatCal(Λ, Q, l, $\hat{a}$)
    // Base Case
    if $l == 1$ **then**
        Compute the matrix $[\mathcal{O}^1]_{\hat{a}\times\hat{a}}$ based on the MMPP-copula in Theorem 4.2
        Compute the matrix $[\mathcal{R}^1]_{\hat{a}\times\hat{a}}$ from $[\mathcal{O}^1]_{\hat{a}\times\hat{a}}$
        return $[\mathcal{O}^1]_{\hat{a}\times\hat{a}}$ and $[\mathcal{R}^1]_{\hat{a}\times\hat{a}}$
    end if
    // Inductive Step
    $([\mathcal{O}^{l-1}]_{\hat{a}\times\hat{a}}, [\mathcal{R}^1]_{\hat{a}\times\hat{a}}) \leftarrow$ CpaMatCal(Λ, Q, $l-1$, $\hat{a}$)
    for $x \leftarrow 1, \hat{a}$ **do**
        for $y \leftarrow 1, \hat{a}$ **do**
            Rotate matrix $[\mathcal{R}^1]_{x\times y}$ 180 degree clockwise to be $[\mathcal{R}']_{x\times y}$
            Calculate Hadamard product of $[\mathcal{O}^{l-1}]_{x\times y}$ and $[\mathcal{R}']_{x\times y}$ as $[\mathcal{T}]_{x\times y}$
            $\mathcal{O}^l_{xy} \leftarrow$ sum of all elements in matrix $[\mathcal{T}]_{x\times y}$
        end for
    end for
    return $[\mathcal{O}^l]_{\hat{a}\times\hat{a}}$ and $[\mathcal{R}^1]_{\hat{a}\times\hat{a}}$
end procedure

4.3. Multi-Step Theoretical Copulas for MMPPs

With the above analysis, the theoretical copula between A_i and A_{i+1} in single MMPP and that in l-IHMMPP are constructed. They can be extended to a more general version for multi-step dependence, that is, the dependence between A_i and $A_{i+i'}(i' \in \mathbb{Z}_{>1})$. Theorems 4.8 and 4.9 give the multi-step MMPP copula and multi-step l-IHMMPP copula, respectively.

Theorem 4.8 (Multi-step MMPP copula). *The copula of A_i and $A_{i+i'}$ in a single MMPP can be calculated as:*

$$C_{\{1;i'\}}(u, v) = \mathcal{G}\big(M_1^{-1}(u)\big)diag(\Pi)P(i'\Delta)\mathcal{G}\big(M_1^{-1}(v)\big)^T. \tag{11}$$

Theorem 4.8 is an extension of Theorem 4.2. For brevity, its proof is omitted, because the proof is the same as the proof of Theorem 4.2 except for replacing A_{i+1} with $A_{i+i'}$.

Theorem 4.9 (Multi-step l-IHMMPP copula). *The copula of A_i^l and $A_{i+i'}^l$ in l-IHMMPP can be constructed by integrating the multi-step MMPP copula in Theorem 4.8 into the recursive method in Theorem 4.4. Specifically, all copulas C_l in the recursive method are replaced by multi-step copulas $C_{\{l;i'\}}$.*

5. PARAMETRIC COPULA MODELING FOR MMPPS

In this section, we construct parametric copulas for single MMPP as well as l-IHMMPP. The parametric copulas we investigate include the following three Archimedean copulas [Nelson 2006]:

(1) Clayton copula,

$$C_\theta(u, v) = [\max\{u^{-\theta} + v^{-\theta} - 1, 0\}]^{-1/\theta}, \quad \theta \in [-1, \infty) \setminus \{0\};$$

(2) Frank copula,

$$C_\theta(u, v) = -\frac{1}{\theta}\log\left[1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1}\right], \quad \theta \in [-\infty, \infty) \setminus \{0\};$$

(3) Gumbel copula,

$$C_\theta(u, v) = \exp[-((-\log u)^\theta + (-\log v)^\theta)^{1/\theta}], \quad \theta \in [1, \infty).$$

We use these three parametric copulas, denoted as C_θ in context, to model MMPPs, due to two reasons: they are all one-parameter copula, and they capture different types of tail dependence effectively. In the aspect of dependence, the parameters θ of the three copula are directly related to the copula-based dependence measures—Kendall's tau ρ_τ and Spearman's rho ρ_s [Aas 2004]. So the parameter itself shows the degree of the dependence between random variables. The tail dependence features of the three copulas are distinct with each other: Clayton copula is to model lower tail dependence with upper tail dependence as 0, that is, $\rho_t^+ = 0$; Frank copula is to model symmetric upper and lower tail dependence, that is, $\rho_t^+ = \rho_t^-$; Gumbel copula is to model upper tail dependence with lower tail dependence as 0, that is, $\rho_t^- = 0$. Overall, the above three copulas are good candidates to potentially capture the dependence structure in MMPP.

There are three main steps to conduct the parametric copula modeling:

(1) Compute the tail dependence by definitions in Equations (3) and (4). The upper tail dependence, as the limit of a function when u approaches 1, can be approximated by evaluating a function value at u, where u is close to 1 [Van Oordt and Zhou 2012], say 0.99. Similarly, the lower tail dependence can be approximated by evaluating the function value at u where u is close to 0, say 0.01:

$$\begin{aligned}\rho_t^+ &\approx Pr\big(X > F_X^{-1}(u) \big| Y > F_Y^{-1}(u)\big)\big|_{u=0.99} \approx \frac{1 - 2u + C(u, u)}{1 - u}\Big|_{u=0.99},\\ \rho_t^- &\approx Pr\big(X < F_X^{-1}(u) \big| Y < F_Y^{-1}(u)\big)\big|_{u=0.01} \approx \frac{C(u, u)}{u}\Big|_{u=0.01}.\end{aligned} \tag{12}$$

(2) Choose the parametric copula for MMPP modeling based on tail dependence calculated from the theoretical copula for MMPPs:
   (a) choose Clayton copula if $\rho_t^+ \approx 0$ and $\rho_t^- > 0$;
   (b) choose Frank copula if $\rho_t^+ \approx \rho_t^-$;
   (c) choose Gumbel copula if $\rho_t^+ > 0$ and $\rho_t^- \approx 0$.

(3) Determine the value of the parameter θ for the chosen parametric copula. In detail, the parameter θ is learned by fitting the MMPP trace data into the chosen copula with the maximum likelihood estimation method [Bouyé et al. 2000].

6. AN APPLICATION: TRAFFIC PREDICTION BASED ON COPULAS FOR MMPP

So far, we have studied the temporal dependence structure of MMPPs captured with copulas. This contribution can benefit many applications. To list some examples, we can use the temporal dependence structure to predict future traffic; we can use it for parameter estimation of MMPP traffic; we can also use it to collaboratively auto-scale service rates in cloud to support composite services, where the aggregated work flow of tasks could be modeled with MMPP [Dong et al. 2016]. Taking parameter estimation as an example, an estimation method of MMPP could be developed based on joint behaviour of successive arrival counts. The parameter of MMPP can be determined by fitting marginal distributions of arrival counts and fitting copula between arrival counts separately. The two-step fitting not only models the full dependence structure and statistical property of arrival counts well, but also decreases the computational overhead, because it avoids the complex calculation of all MMPP parameters in one step.

In this article, we apply MMPP copula for traffic prediction. With experiments in the following section, we show that prediction of MMPP traffic with copula-based dependence model outperforms classical prediction methods, such as autoregressive model and linear predictive coding, because the copula-based model is stable and carries rich functional dependence. The problem of traffic prediction can be posed in different forms. In this article, we focus on estimating the future arrival count $A_{i+i'}$ based on the current observation of arrival count A_i. The prediction is made by maximizing the conditional probability $Pr(A_{i+i'}|A_i)$, that is, $\hat{A}_{i+i'} = \text{argmax}_y\, Pr(A_{i+i'} = y|A_i = x)$. When $i' = 1$, the prediction is made one-step forward; when $i' > 1$, the prediction is made multi-step forward. In this section, we introduce the one-step prediction method with both theoretical copulas and parametric copulas, followed by a discussion for multi-step prediction.

6.1. One-Step Prediction Based on Theoretical Copulas

With MMPP-copula C_1 for single MMPP and theoretical copula C_l for l-IHMMPP, Theorem 6.1 can be used to predict future arrivals.

THEOREM 6.1. *(1) Consider a MMPP having its MMPP-copula C_1. If $A_i = x$ is the current observation from the arrival process and if the prediction is made by maximizing the conditional probability $Pr(A_{i+1}|A_i)$, then the predicted arrival count $\hat{A}_{i+1}$ is*

$$\hat{A}_{i+1} = \underset{y}{\text{argmax}}\, \nabla C_1(M_1(x), M_1(y)). \tag{13}$$

(2) Consider l-IHMMPP having theoretical copula C_l. If $A_i^l = x$ is the current observation from the arrival process and if the prediction is made by maximizing the conditional probability $Pr(A_{i+1}^l|A_i^l)$, then the predicted arrival count $\hat{A}_{i+1}^l$ is

$$\hat{A}_{i+1}^l = \underset{y}{\text{argmax}}\, \nabla C_l(M_l(x), M_l(y)). \tag{14}$$

PROOF. We only prove part (1), since part (2) can be proved in the same way. Since the prediction is made by maximizing the conditional probability $Pr(A_{i+1}|A_i)$, we have

$$\begin{aligned}\hat{A}_{i+1} &= \underset{y}{\text{argmax}}\, Pr(A_{i+1} = y|A_i = x)\\ &= \underset{y}{\text{argmax}}\, \frac{Pr(A_i = x, A_{i+1} = y)}{Pr(A_i = x)}\\ &= \underset{y}{\text{argmax}}\, \frac{\nabla C_1(M_1(x), M_1(y))}{Pr(A_i = x)}\\ &= \underset{y}{\text{argmax}}\, \nabla C_1(M_1(x), M_1(y)). \quad \square\end{aligned}$$

As shown in Section 4.2, the value of ∇C_1 function is represented by the gradient matrix $\mathcal{R}^1$. With this numerical transformation by definition in Table I, the predicted arrival count is

$$\hat{A}_{i+1} = \underset{y}{\text{argmax}}\, \mathcal{R}^1_{(x+1)(y+1)} = \left(\underset{y}{\text{argmax}}\, \mathcal{R}^1_{(x+1)y}\right) - 1.$$

Similarly, for the prediction of l-IHMMPP, we can compute the gradient matrix $\mathcal{R}^l$ from $\mathcal{O}^l$ following the rule described in Remark 4.6.

While it is extremely difficult to theoretically prove the uniqueness of *argmax* value in Equations (13) and (14), in practice, we only search in a limited range of arrival counts, and hence the *argmax* value is unique most of time. When the *argmax* value is not unique, although unlikely and never seen in our later case studies, we recommend taking the average value as the predicted value for robustness consideration.

6.2. One-Step Prediction Based on Parametric Copulas

Given a single MMPP or l-IHMMPP trace, a parametric copula C_θ is learned according to Section 5. The marginal values of A_i and A_i^l are discrete. However, the parametric copula C_θ is continuous in the domain of [0, 1]. Due to this reason, we first study the prediction problem on stochastic process with continuous marginals (Theorem 6.2) and then extend its usage for discrete distributions (Theorem 6.3).

THEOREM 6.2. *Consider a stochastic process B_i that has a parametric copula C_θ between B_i and B_{i+1}, continuous marginal distribution F, and marginal probability density function (pdf) f. We have the following conditional pdf:*

$$f(B_{i+1} = y | B_i = x) = c_\theta(F(x), F(y)) f(y), \tag{15}$$

where $c_\theta(u, v) = \frac{\partial}{\partial u}\frac{\partial}{\partial v} C_\theta(u, v)$ is called the copula density function.

For discrete marginals, we revise Theorem 6.2 by relating the probability density function(pdf) in continuous distribution to the probability mass function (pmf) in discrete distribution.

THEOREM 6.3. *Consider a stochastic process B_i that has a parametric copula C_θ between B_i and B_{i+1}, discrete marginal distribution F, and marginal probability mass function (pmf) p. We have the following conditional pmf:*

$$p(B_{i+1} = y | B_i = x) = c_\theta(F(x), F(y)) p(y). \tag{16}$$

The proofs of Theorems 6.2 and 6.3 are straightforward using techniques similar to those in Azam and Pitt [2014]. With Theorem 6.3, prediction based on parametric copula on MMPP arrival counts is given by Theorem 6.4.

THEOREM 6.4. *Consider a single MMPP that has its parametric copula C_θ. If $A_i = x$ is the current observation from the arrival process and if the prediction is made by maximizing the conditional probability $Pr(A_{i+1}|A_i)$, then the predicted future arrival count $\hat{A}_{i+1}$ is*

$$\begin{aligned} \hat{A}_{i+1} &= \operatorname*{argmax}_y c_\theta(M_1(x), M_1(y)) p(y) \\ &= \operatorname*{argmax}_y c_\theta(M_1(x), M_1(y))(M_1(y) - M_1(y-1)); \end{aligned} \tag{17}$$

In the case of prediction on l-IHMMPP, the predicted future arrival count $\hat{A}^l_{i+1}$ is

$$\hat{A}^l_{i+1} = \operatorname*{argmax}_y c_\theta(M_l(x), M_l(y))(M_l(y) - M_l(y-1)). \tag{18}$$

6.3. Multi-Step Prediction Based on Copulas

The multi-step prediction can be made by replacing the theoretical copulas in Theorem 6.1 with multi-step theoretical copulas, as indicated in the following theorem.

THEOREM 6.5. *(1) Consider a MMPP having its multi-step MMPP-copula $C_{\{1;i'\}}$. If $A_i = x$ is the current observation from the arrival process, then the predicted arrival count $\hat{A}_{i+i'}$ is*

$$\hat{A}_{i+i'} = \operatorname*{argmax}_y \nabla C_{\{1;i'\}}(M_1(x), M_1(y)). \tag{19}$$

(2) Consider l-IHMMPP having multi-step l-IHMMPP copula $C_{\{l;i'\}}$. If $A_i^l = x$ is the current observation from the arrival process, then the predicted arrival count $\hat{A}_{i+i'}^l$ is

$$\hat{A}_{i+i'}^l = \operatorname*{argmax}_y \nabla C_{\{l;i'\}}(M_l(x), M_l(y)). \tag{20}$$

To make multi-step prediction with parametric copulas, the parametric copula of A_i and $A_{i+i'}$ needs to be fitted empirically. Then the prediction is made according to Theorem 6.4.

7. EXPERIMENTAL EVALUATION

We conduct experiments to show how the copula model could help traffic prediction. In the evaluation, we first give a broad view of the methods to evaluate the performance of copula-based prediction. We then show case studies on single MMPP as well as l-IHMMPP.

7.1. Evaluation Methods

We have proposed both theoretical copula modeling and parametric copula modeling for traffic prediction in Section 6. To evaluate the two prediction models, we implement two classic prediction models, autoregressive model (AR(1)) and linear predictive coding (LPC(1)), for comparison. Note that the first order AR model and the first order LPC model are used here for a fair comparison, because our copula-based prediction model is first order in the sense that only dependence between two successive arrival counts is considered each time.

Given one observation A_i, the multi-step prediction based on AR(1) will be made as

$$\begin{aligned}
\hat{A}_{i+1} &= c + \varphi A_i + \epsilon_{i+1},\\
\hat{A}_{i+2} &= c + \varphi \hat{A}_{i+1} + \epsilon_{i+2},\\
&\cdots\\
\hat{A}_{i+i'} &= c + \varphi \hat{A}_{i+i'-1} + \epsilon_{i+i'},
\end{aligned}$$

where c is a constant, φ is the parameter of the AR(1) model, and ϵ_t is white noise.

With LPC(1), the multi-step prediction from A_i is made as

$$\begin{aligned}
\hat{A}_{i+1} &= \gamma A_i,\\
\hat{A}_{i+2} &= \gamma \hat{A}_{i+1},\\
&\cdots\\
\hat{A}_{i+i'} &= \gamma \hat{A}_{i+i'-1},
\end{aligned}$$

where γ is the parameter of LPC(1) model, and is directly determined by autocorrelation of A_i.

As a purely linear predictor, LPC(1) model is to predict data only based on the dependence information in terms of autocorrelation. Thus it is set as the benchmark predictor to show how functional dependence modeling with copulas improves over linear dependence. We also compare copula-based prediction with AR(1) model, since AR(1) model is the popular statistical method for prediction.

When applying any of the prediction models on a traffic trace, the trace is divided into two parts, the training set and the testing set. The training set comes from the first certain percentage of trace data, and the rest of the trace constitutes the testing set. For example, if the training percentage is 50%, the first half of the trace will be used to train a model, and the second half will be used to test prediction accuracy. The prediction accuracy is measured by root-mean-square error (RMSE) across the test set, defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{A}_i - A_i)^2}, \tag{21}$$

where A_i is arrival counts from test set at timeslot i, $\hat{A}_i$ denotes the corresponding predicted value, and n is the total number of time slots in the testing period.

For a prediction model, its average RMSE (aRMSE) over different experiment scenarios represents its overall performance on MMPP traffic trace prediction. Its performance improvement ratio (IMP RATIO) over benchmark model (LPC(1)) is defined in Equation (22). The larger the value is, the more the predictor improves over LPC(1) model:

$$\text{IMP RATIO} = \frac{\mathrm{aRMSE}_{\mathrm{benchmark}} - \mathrm{aRMSE}}{\mathrm{aRMSE}_{\mathrm{benchmark}}} * 100\%. \tag{22}$$

7.2. Case Study on Single MMPP from Real-World Trace

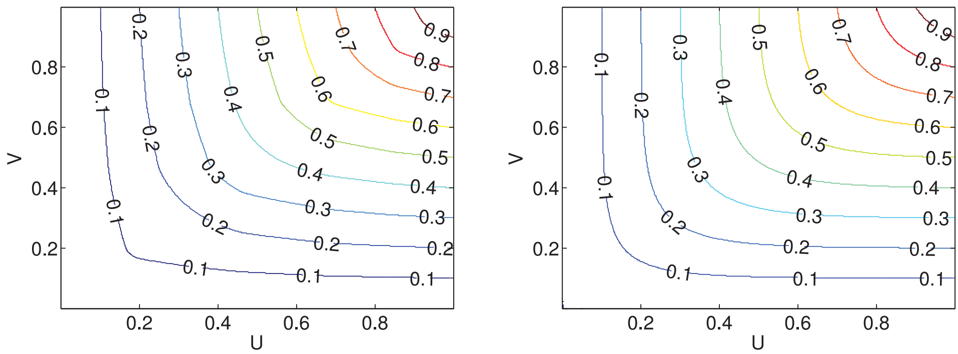
BCpAug89 trace, one of Bellcore traces,[2] records the exact arrival times of 1,000,000 packets on an Ethernet at Bellcore Morristown Research and Engineering facility. Previous research has shown that the trace is well characterized by MMPP [Andersen and Nielsen 1998; Nogueira et al. 2003; Muscariello et al. 2005]. We analyse the trace in terms of arrival counts every second, that is, the length of time slot is set as $\Delta = 1$ (second), and A_i denotes the number of arrivals in ith second. With learning algorithm proposed in Heyman and Lucantoni [2003], this trace is modeled by a 12-state MMPP with parameters $(\boldsymbol{Q}, \Lambda)$ as shown in Equation (23). In the case study, we will apply copula to model its dependence structure and predict the trace flow. We will also vary the trace by a functional transformation to show that the copula-based dependence model is much more stable than other models.

$$\boldsymbol{Q} = \left(\begin{array}{cccccccccccc}
-0.857 & 0.286 & 0.428 & 0.143 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.067 & -0.900 & 0.267 & 0.233 & 0.233 & 0.067 & 0.033 & 0 & 0 & 0 & 0 & 0 \\
0.023 & 0.078 & -0.837 & 0.336 & 0.203 & 0.103 & 0.078 & 0 & 0.016 & 0 & 0 & 0 \\
0 & 0.026 & 0.140 & -0.722 & 0.274 & 0.153 & 0.085 & 0.030 & 0.007 & 0.007 & 0 & 0 \\
0.002 & 0.008 & 0.051 & 0.173 & -0.651 & 0.244 & 0.122 & 0.041 & 0.006 & 0.002 & 0.002 & 0 \\
0 & 0.001 & 0.027 & 0.074 & 0.173 & -0.696 & 0.303 & 0.094 & 0.014 & 0.009 & 0.001 & 0 \\
0 & 0.001 & 0.004 & 0.019 & 0.099 & 0.233 & -0.617 & 0.200 & 0.048 & 0.012 & 0.001 & 0 \\
0 & 0 & 0.008 & 0.023 & 0.049 & 0.184 & 0.409 & -0.775 & 0.084 & 0.015 & 0.003 & 0 \\
0 & 0 & 0.008 & 0.015 & 0.015 & 0.120 & 0.301 & 0.218 & -0.805 & 0.113 & 0.015 & 0 \\
0 & 0.020 & 0 & 0 & 0.059 & 0.059 & 0.235 & 0.078 & 0.275 & -0.824 & 0.098 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.077 & 0.231 & 0.231 & 0.154 & 0.077 & -0.847 & 0.077 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1
\end{array}\right),$$

$$\Lambda = (782.069, 674.207, 574.345, 482.483, 398.621, 322.759, 254.897, 195.035, 143.173, 99.311, 63.449, 35.587). \tag{23}$$

7.2.1. One-Step Prediction on BCpAug89 Trace. Given the learned MMPP parameter $(\boldsymbol{Q}, \Lambda)$, we first construct copula for MMPP theoretically and empirically. With theoretical analysis in Section 4.1, the MMPP-copula for learned MMPP is computed from the parameters $(\boldsymbol{Q}, \Lambda)$ based on Theorem 4.2. The contour of the computed MMPP-copula is shown in Figure 5(a). Dependence measures, including Kendall's tau ρ_τ, Spearman's rho ρ_s, tail dependence ρ_t^+ and ρ_t^-, and Pearson correlation coefficient ρ, between A_i and A_{i+1} are analysed in Table II. Theoretical results of ρ_τ, ρ_s, and ρ_t are calculated from copula with Equations (1)–(4). Pearson coefficient is calculated based on the analysis in Neuts [1989]. Except Pearson coefficient, all other dependence measures can be obtained via copula, indicating that copula includes rich information about dependence

[2]The Bellcore traces are available on the website http://ita.ee.lbl.gov/html/contrib/BC.html.



(a) Theoretical MMPP-copula

(b) Parametric copula with the training percentage as 80%

Fig. 5. Copula contours for MMPP learned from BCpAug89 trace.

Table II. Dependence Measures of BCpAug89 Trace from Theoretical Analysis and Empirical Analysis

	ρ_τ	ρ_s	$\rho_t^+\vert_{u=0.99}$	$\rho_t^-\vert_{u=0.01}$	ρ
Theoretical	0.4788	0.6150	0.4067	0.3359	0.7555
Empirical	0.4212	0.5897	0.3935	0.3248	0.6149

structure. In addition, the comparison between theoretical and empirical dependence measures shows that copula accurately captures the trace dependence.

Empirically, a parametric copula can be also chosen to model temporal dependence between A_i and A_{i+1}. As shown in Table II, ρ_t^+ is close to ρ_t^-, we thus choose Frank copula to model the BCpAug89 trace. The parameter of Frank copula is determined by fitting a training set from the trace. With different training percentage of data, the parameter of Frank copula will be determined accordingly. Figure 5(b) shows the contour of parametric copula trained from 80% percentage of data in BCpAug89 trace. Figures 5(a) and 5(b) have close contour shape. The similarity of two copulas can be quantified by the discrete L_2 norm distance over the size of discrete lattice [Durrleman et al. 2000], which is 0.0173 in our case. This value is close to those in the experiments of Durrleman et al. [2000] when selecting two similar copulas, indicating that the parametric copula trained empirically from data accords well with the theoretical copula from analysis.

With the copulas constructed from the training set of BCpAug89 trace, one-step prediction is conducted on its testing set. Figure 6 shows at a glance the prediction with theoretical copula on the last 20% arrivals of BCpAug89 trace. To obtain multiple prediction results for the aRMSE measurement, we adjust the training percentage from 50% to 90%. The prediction accuracy of copulas in measure of RMSE is shown in detail and compared with AR(1) model and LPC(1) model in Table III. From the table, we can infer that both MMPP-copula and parametric copula characterize the temporal dependence of BCpAug89 trace well, leading to a good prediction. Copula-based predictions, including theoretical copula model and parametric copula model, have more than 10% improvement ratio over the LPC(1) model, showing the advantage of functional dependence modeling (such as copulas) over linear dependence measurement (such as autocorrelation). Copula-based predictions achieve accuracy similar to the classical AR(1) model, showing that copula captures the dependence of real-world MMPP trace effectively, which in turn helps the prediction. In addition, copula-based predictions have other benefits compared to AR(1) model, as shown in the next section.

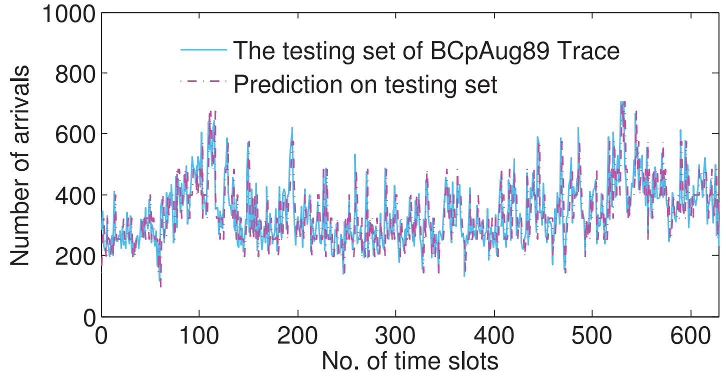


Fig. 6. Prediction with theoretical copula on the testing set (last 20%) of BCpAug89 trace.

Table III. One-Step Prediction RMSE on BC-pAug89 Trace with Different Training Percentages

Training Percentage	Theoretical Copula	Parametric Copula	AR(1)	LPC(1)
50%	94.2411	88.9070	92.2850	110.3130
60%	90.5974	87.5581	88.8550	106.1805
70%	93.1982	91.7414	90.8745	108.5895
80%	92.3244	88.1251	90.2618	105.2930
90%	94.4256	93.2204	92.9318	108.0254
aRMSE	92.95734	89.9104	91.04162	107.6803
IMP RATIO	13.67%	16.50%	15.45%	—

Table IV. Dependence Measures of the Associated Trace from Theoretical Analysis and Empirical Analysis

	ρ_τ	ρ_s	$\rho_t^+\|_{u=0.99}$	$\rho_t^-\|_{u=0.01}$	ρ
Theoretical	0.4788	0.6150	0.4067	0.3359	—
Empirical	0.4212	0.5897	0.3935	0.3248	0.5916

7.2.2. The Stability of Copula-Based Model. Nowadays, a network flow may pass through many middleboxes, which may transform the traffic with some (potentially unknown) functions. In some scenarios, we may need to consider another counting process closely associated with the incoming traffic, for example, the number of CPU resources or the size of cache space that should be (dynamically) allocated for processing the traffic. In these cases, the traffic is transformed with some functions or the new counting process can be viewed as the traffic transformed with a function. In the following, we study a new process $D_i = \log(A_i)$ as an example. We note that the same conclusion could be drawn with other transformation functions. We call D_i an *associated* trace.

With the invariant property of copulas, the temporal dependence between D_i and D_{i+1} in terms of copula remains the same as that between A_i and A_{i+1}. The measures ρ_τ, ρ_s and ρ_t among trace D_i will also have the same theoretical results, since all of them could be derived with copula. However, since Pearson correlation does not satisfy the invariant property with the above transformation, ρ of D_i is not theoretically tractable and thus needs to be calculated from empirical statistics. Table IV shows the measures of trace D_i. Comparing Table II and Table IV, we can see that copula-based dependencies are all the same while Pearson correlation varies, indicating that copula is much more stable than Pearson correlation.

Taking the above advantage, we do not need to rebuild the dependence model when it comes to the prediction of D_i with copula, because the same copula model for A_i can be applied and the marginal function of D_i can be obtained from A_i by $M_D(x) = M_A(2^x)$.

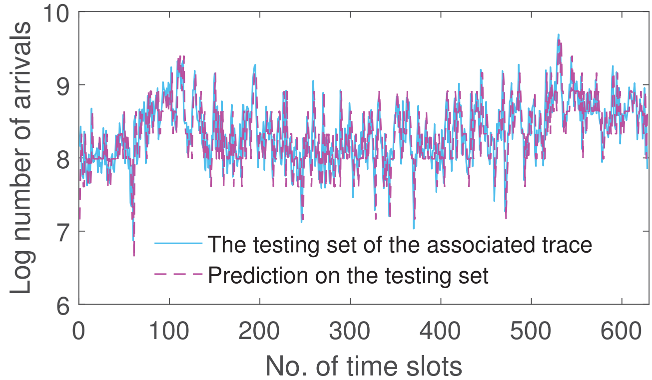


Fig. 7. Prediction with theoretical copula on the testing set (last 20%) of the associated trace.

Table V. One-Step Prediction RMSE on the Associated Trace with Different Training Percentages

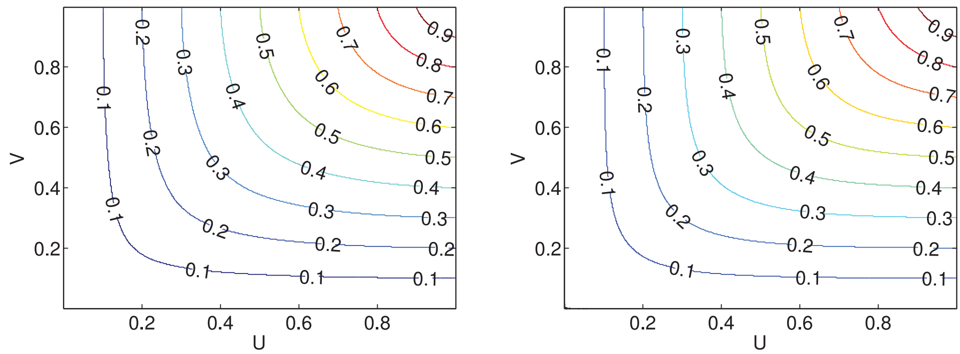
Training Percentage	Theoretical Copula	Parametric Copula	AR(1)	LPC(1)
50%	0.4788	0.4344	0.5918	0.6953
60%	0.4653	0.4286	0.5868	0.6982
70%	0.4710	0.4393	0.5863	0.7137
80%	0.3955	0.3659	0.5471	0.6765
90%	0.3780	0.3655	0.5324	0.7390
aRMSE	0.4377	0.4068	0.5689	0.7045
IMP RATIO	37.87%	42.26%	19.26%	—

Therefore, all copula models in Section 7.2.1 can be applied directly to predict D_{i+1} given the D_i value. Figure 7 shows the prediction on the last 20% of the associated trace by using the same copula model of A_i. Nevertheless, without rebuilding the dependence model, the AR(1) and LPC(1) models for A_i applied to D_i will lead to poor prediction performance. It is worth noting that rebuilding a new model for D_i may be non-trivial due to the potentially unknown transformation function and the need of collecting and recording historical data of D_i.

To test prediction performance without rebuilding a model, we apply the trained models (i.e., copula, AR(1), and LPC(1)) from A_i to predict the associated trace D_i. The one-step prediction RMSEs on D_i of four methods are listed in Table V. Both theoretical and parametric copulas outperform AR(1) and LPC(1) significantly. The results indicate that copula-based prediction is much more stable in the presence of traffic transformation, and both AR(1) and LPC(1) cannot capture the dependence in the associated trace accurately without a re-modeling process. Copulas take advantage of the invariant property to avoid the re-modeling process whenever an increasing functional transformation is imposed on the original traffic, leading to its much better performance over other models.

7.3. Case Study on *I*-IHMMPP Trace with Simulation

In real world, the availability of superposition MMPP traffic traces is limited, because it is not easy to identify them with proper fitting and goodness-testing methods. So, we generate superposed MMPP traces by simulation. We consider a scenario that there are three independent sources sending the traffic flows, with features similar to BCpAug89 trace, to one destination. The flow to the destination can be simulated as the aggregation trace of three independent MMPP traces generated by parameters (Q, Λ) shown in Equation (23). The simulation lasts for $7,200$s. The generated trace as



(a) Theoretical 3-IHMMPP copula

(b) Parametric copula for 3-IHMMPP with the training percentage as 80%

Fig. 8. One-step copula contours for 3-IHMMPP.

Table VI. Dependence Measures of the Simulated 3-IHMMPP Trace from Theoretical Analysis and Empirical Analysis

	ρ_τ	ρ_s	$\rho_t^+\|_{u=0.99}$	$\rho_t^-\|_{u=0.01}$	ρ
Theoretical	0.5681	0.7329	0.3367	0.2484	—
Empirical	0.5500	0.7370	0.2857	0.2875	0.7566

described above is named as the *simulated* 3-*IHMMPP trace*. We analysed it in terms of arrival counts every second, and the arrival count in ith second is denoted as A_i^3. On the simulated 3-IHMMPP trace, we perform both one-step prediction and two-step prediction and compare copula models with others.

7.3.1. One-Step Prediction on the Simulated 3-IHMMPP Trace. When constructing 3-IHMMPP copula matrix for the simulated 3-IHMMPP trace, the threshold is set as $\hat{a} = 2000$ according to observation of samples. Given parameters (Q, Λ) in Equation (23), the theoretical copula of 3-IHMMPP is calculated based on Theorems 4.3, 4.4, and 4.7. The contour of the theoretical copula is shown in Figure 8(a). Note that even though we only compute the theoretical copula for $A_i^3 \leq 2000$, it is almost the complete copula, because the threshold is large enough to make $C_3(M_3(\hat{a}), M_3(\hat{a})) = 0.99998 \approx 1$, meaning that the probability for arrival counts to go beyond the threshold is extremely small. With the theoretical copula C_3 constructed, dependence measures are calculated accordingly and compared with empirical results from trace data. Table VI show the results, which indicate the accuracy of copula in modeling the trace. The theoretical Pearson correlation ρ is missing, since its value on aggregate MMPP is extremely hard to calculate when the underlying MMPP has a large number of states.

Since the trace has similar upper and lower tail dependences as shown in Table VI, we choose Frank copula as the parametric copula. It indicates that the l-IHMMPP inherits the tail dependence features from single MMPP. The parameter of Frank copula is fitted according to different training set. Figure 8(b) shows the contour of the parametric copula trained from first 80% data of the simulated 3-IHMMPP trace. Contours in Figures 8(a) and 8(b) are very similar. Their discrete L_2 norm distance [Durrleman et al. 2000] is 0.0100, which is small enough to justify the similarity of two copulas according to the results in Durrleman et al. [2000].

With both theoretical copula C_3 and parametric copula C_θ, we perform one-step prediction on the simulated 3-IHMMPP trace. Figure 9 shows the prediction with theoretical 3-IHMMPP copula on the testing set of the last 20% arrival counts.

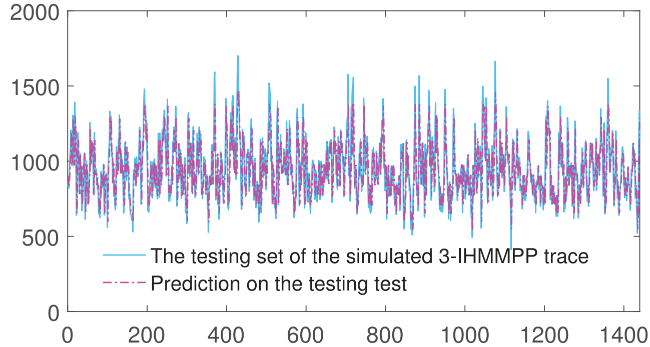
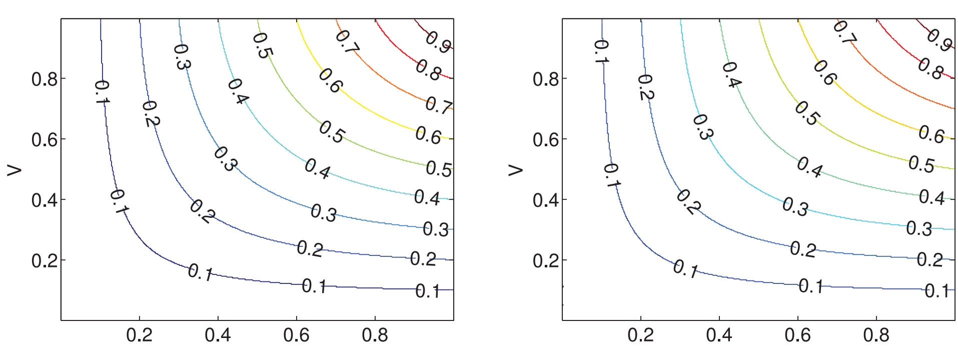


Fig. 9. Prediction with theoretical 3-IHMMPP copula on the testing set (last 20%) of the simulated 3-IHMMPP trace.

Table VII. One-Step Prediction RMSE on the Simulated 3-IHMMPP Trace with Different Training Percentage

Training Percentage	Theoretical Copula	Parametric Copula	AR(1)	LPC(1)
50%	133.2363	135.5447	138.2753	197.2492
60%	133.3298	137.0763	138.4576	199.7005
70%	131.5002	135.3443	136.6052	199.2604
80%	131.0944	132.6420	136.3443	198.3723
90%	127.9317	128.9178	133.2959	198.4436
aRMSE	131.4185	133.9050	136.5957	198.6052
IMP RATIO	33.82%	32.58%	31.22%	—



(a) Two-step theoretical 3-IHMMPP copula

(b) Two-step parametric copula for 3-IHMMPP with the training percentage as 80%

Fig. 10. Two-step copula contours for 3-IHMMPP.

We adjust the training percentage from 50% to 90%, and the prediction results are shown in Table VII. From the table, copula-based prediction has the highest IMP RATIO over the benchmark prediction regarding the aggregate MMPP traffic, indicating that copulas capture the temporal dependence of l-IHMMPP the best.

7.3.2. Two-Step Prediction on the Simulated 3-IHMMPP Trace. We also experiment two-step prediction on the simulated 3-IHMMPP trace. That is, with any observation A_i^3 in the test set, A_{i+2}^3 is predicted. In order to make two-step prediction, the two-step theoretical copula of 3-IHMMPP $C_{\{3;2\}}$ is constructed as shown in Figure 10(a). Based on the two-step copula, the dependence measures are given and compared with empirical

Table VIII. Two-step Dependence Measures of the Simulated 3-IHMMPP Trace from Theoretical Analysis and Empirical Analysis

	ρ_τ	ρ_s	$\rho_t^+\|_{u=0.99}$	$\rho_t^-\|_{u=0.01}$	ρ
Theoretical	0.2979	0.4189	0.1338	0.1077	—
Empirical	0.3372	0.4836	0.1690	0.1286	0.5104

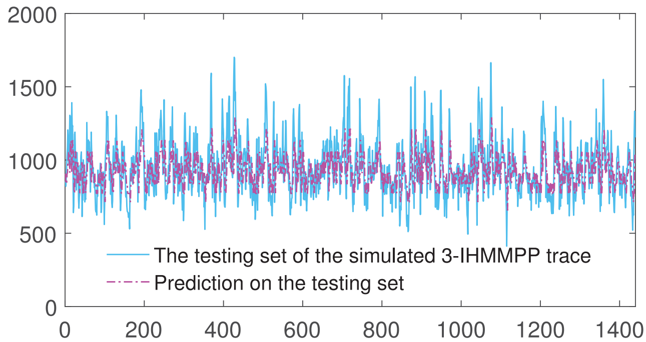


Fig. 11. Two-step prediction with theoretical $C_{\{3,2\}}$ copula on the testing set (last 20%) of the simulated 3-IHMMPP trace.

Table IX. Two-Step Prediction RMSE on the Simulated 3-IHMMPP Trace with Different Training Percentage

Training Percentage	Theoretical Copula	Parametric Copula	AR(1)	LPC(1)
50%	173.3941	174.2526	196.5692	242.6005
60%	174.9574	176.4601	198.0623	247.9247
70%	173.3858	174.7150	197.3366	250.4306
80%	173.2769	174.3444	196.9888	250.7354
90%	169.6243	170.0910	194.5932	246.3572
aRMSE	172.9277	173.9726	196.7100	247.6097
IMP RATIO	30.16%	29.74%	20.56%	—

results in Table VIII. Compared to one-step dependencies in Table VI, two-step dependencies between A_i^3 and A_{i+2}^3 are smaller, because the dependence decreases as the step increases.

Based on a training set of the simulated 3-IHMMPP trace, the parametric copula between A_i^3 and A_{i+2}^3 is trained accordingly. Figure 10(b) shows the contour of the two-step parametric copula trained from a training set consisting of the first 80% data of the simulated 3-IHMMPP trace. For two-step dependence, the theoretical copula and the parametric copula are also close to each other (their discrete L_2 norm distance [Durrleman et al. 2000] is 0.0045).

With different training percentages, two-step predictions are performed on the simulated 3-IHMMPP trace. The prediction results of applying the theoretical copula $C_{\{3;2\}}$ on the last 20% of the simulated 3-IHMMPP trace are shown in Figure 11. Prediction errors in terms of RMSE with different training percentages are shown in Table IX. Our copula models have significant improvement ratio (IMP RATIO) over benchmark model regarding the two-step predictions. Compared with AR(1), copulas also have a much better performance, indicating that copula can better characterize multi-step temporal dependence of arrival counts in MMPP.

7.4. Summary of Experiments

Both real-world traffic trace and simulated trace are used to evaluate the copula model and its application to traffic prediction. In Section 7.2.1, the trace BCpAug89 is chosen to evaluate the accuracy of using MMPP copula to model the real-world trace. Prediction on a transformed trace from BCpAug89 in Section 7.2.2 shows that copula has much better dependence characterization and makes more accurate prediction than existing models in the presence of traffic transformation. Experiments in Sections 7.3.1 and 7.3.2 show copula's good performance on one-step prediction and two-step prediction of l-IHMMPP, respectively.

From all the experiments, copula-based model has advantages over other dependence models (AR(1) and LPC(1)) in three aspects: First, it provides theoretical dependence structure of MMPP, including one-step or multi-step dependence of single MMPP and l-IHMMPP. Second, it provides more information on dependence beyond linear scope. Third, it is more stable than other models. In the presence of traffic transformation, copula-based model does not require rebuilding a new dependence model but still guarantees accuracy.

8. CONCLUSION

This article is the first that theoretically derives the intricate temporal dependence structure in MMPPs with copula analysis. It presents the theoretical solution for modeling temporal dependence in both single MMPP and superposition of homogeneous MMPPs. In addition, parametric copulas have been investigated for fast approximation of theoretical copulas. Using the theoretical and parametric copulas to predict the trend of real and simulated traffic traces, we show the value of our research in real-world applications.

REFERENCES

Kjersti Aas. 2004. Modelling the dependence structure of financial assets: A survey of four copulas. Technical Report SAMBA/22/04, Norsk Regnesentral (2004).

Allan T. Andersen and Bo Friis Nielsen. 1997. An application of superpositions of two state markovian source to the modelling of self-similar behaviour. In *Proceedings of INFOCOM'97*. IEEE, Kobe, Japan, 196–204.

Allan T. Andersen and Bo Friis Nielsen. 1998. A Markovian approach for modeling packet traffic with long-range dependence. *IEEE J. Select. Areas Commun.* 16, 5 (1998), 719–732.

Konstantin Avrachenkov, Natalia M. Markovich, and Jithin K. Sreedharan. 2015. Distribution and dependence of extremes in network sampling processes. *Computat. Soc. Netw.* 2, 1 (2015), 1.

Kazim Azam and Michael K. Pitt. 2014. Bayesian inference for a semi-parametric copula-based Markov chain. Working Paper, University of Warwick. (2014).

Andrea Baiocchi and Nicola Blefari-Melazzi. 1993. Steady-state analysis of the MMPP/G/1/K queue. *IEEE Trans. Commun.* 41, 4 (1993), 531–534.

Soshant Bali and Victor S. Frost. 2007. An algorithm for fitting MMPP to IP traffic traces. *IEEE Commun. Lett.* 11, 2 (2007), 207–209.

Khalid Begain, Gunter Bolch, and Helmut Herold. 2012. *Practical Performance Modeling: Application of the MOSEL Language*. Springer Science & Business Media, U.S.

Martina Beil. 2013. *Modeling Dependencies Among Financial Asset Returns Using Copulas*. Ph.D. Dissertation. Technische Universität München.

Vladislav Bína and Radim Jiroušek. 2013. A short note on multivariate dependence modeling. *Kybernetika* 49, 3 (2013), 420–432.

Eric Bouyé, Valdo Durrleman, Ashkan Nikeghbali, Gaël Riboulet, and Thierry Roncalli. 2000. Copulas for finance-a reading guide and some applications. *Available at SSRN 1032533* (2000).

Lothar Breuer and Alfred Kume. 2010. An EM algorithm for Markovian arrival processes observed at discrete times. In *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*. Springer, Berlin, 242–258.

Yulia Burkatovskaya, Tatiana Kabanova, and Sergey Vorobeychikov. 2015. CUSUM algorithms for parameter estimation in queueing systems with jump intensity of the arrival process. In *Information Technologies and Mathematical Modelling-Queueing Theory and Applications*. Springer, Switzerland, 275–288.

Xiaohong Chen, Wei Biao Wu, and Yanping Yi. 2009. Efficient estimation of copula-based semiparametric Markov models. *Ann. Stat.* 37, 6B (2009), 4214–4253.

Tiberiu Chis and Peter G. Harrison. 2015. Adapting hidden Markov models for online learning. *Electron. Notes Theoret. Comput. Sci.* 318 (2015), 109–127.

Doo Il Choi, Tae-Sung Kim, and Sangmin Lee. 2008. Analysis of an MMPP/G/1/K queue with queue length dependent arrival rates, and its application to preventive congestion control in telecommunication networks. *Eur. J. Operat. Res.* 187, 2 (2008), 652–659.

Tibor Csóka and Jaroslav Polec. 2015. Modeling poisson error process on wireless channels. *Int. J. Commun. Netw. Inform. Secur.* 7, 1 (2015), 1–7.

Fang Dong, Kui Wu, Venkatesh Srinivasan, and Jianping Wang. 2016. Copula analysis of latent dependency structure for collaborative auto-scaling of cloud services. In *Proceedings of ICCCN*. IEEE, 1–8.

Fang Dong, Kui Wu, and Srinivasan Venkatesh. 2015. Copula analysis for statistical network calculus. In *Proceedings of INFOCOM'15*. IEEE, Hong Kong, 1535–1543.

Qing Du. 1995. A monotonicity result for a single-server queue subject to a Markov-modulated Poisson process. *J. Appl. Probabil.* 32, 4 (1995), 1103–1111.

Valdo Durrleman, Ashkan Nikeghbali, Thierry Roncalli, and others. 2000. Which copula is the right one? Working paper, Crédit Lyonnais. (2000).

Robert J. Elliott and W. Paul Malcolm. 2008. Discrete-time expectation maximization algorithms for Markov-modulated Poisson processes. *IEEE Trans. Automat. Control* 53, 1 (2008), 247–256.

Paul Embrechts, Filip Lindskog, and Alexander McNeil. 2003. Modelling dependence with copulas and applications to risk management. *Handbook Heavy Tail. Distrib. Fin.* 8, 1 (2003), 329–384.

Wolfgang Fischer and Kathleen Meier-Hellstern. 1993. The Markov-modulated Poisson process (MMPP) cookbook. *Perform. Eval.* 18, 2 (1993), 149–171.

Christian Genest, Bruno Rémillard, and David Beaudoin. 2009. Goodness-of-fit tests for copulas: A review and a power study. *Insur.: Math. Econ.* 44, 2 (2009), 199–213.

Mahmood Mollaei Gharehajlu, Saadan Zokaei, and Yousef Darmani. 2015. Statistical analysis of different traffic types effect on QoS of wireless ad hoc networks. *J. Inform. Syst. Telecommun.* 3, 1(9) (2015), 7–15.

Paolo Giacomazzi. 2007. Statistical traffic envelopes for Markov-modulated Poisson packet sources. In *Proceedings of IEEE GLOBECOM'07*. IEEE, Washington, D.C., 2628–2633.

Paolo Giacomazzi. 2009. Closed-form analysis of end-to-end network delay with Markov-modulated Poisson and fluid traffic. *Comput. Commun.* 32, 4 (2009), 640–648.

Daniel P. Heyman and David Lucantoni. 2003. Modeling multiple IP traffic streams with rate limits. *IEEE/ACM Trans. Netw.* 11, 6 (2003), 948–958.

Alexander Ihler, Jon Hutchins, and Padhraic Smyth. 2007. Learning to detect events with markov-modulated poisson processes. *ACM Trans.n Knowl. Discov. Data* 1, 3 (2007), 13.

Shoji Kasahara. 2001. Internet traffic modeling: Markovian approach to self-similar traffic and prediction of loss probability for finite queues. *IEICE Trans. Commun.* 84, 8 (2001), 2134–2141.

Natalia Markovich. 2008. *Nonparametric Analysis of Univariate Heavy-tailed Data: Research and Practice*. Vol. 753. John Wiley & Sons, England.

Natalia M. Markovich. 2010. Modeling of dependence in a peer-to-peer video application. In *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*. ACM, 316–320.

Luca Muscariello, Marco Mellia, Michela Meo, M. Ajmone Marsan, and R. Lo Cigno. 2005. Markov models of internet traffic and a new hierarchical MMPP model. *Comput. Commun.* 28, 16 (2005), 1835–1851.

Roger B. Nelson. 2006. *An Introduction to Copulas*. Springer, New York.

David Neuhäuser, Christian Hirsch, Catherine Gloaguen, and Volker Schmidt. 2013. A parametric copula approach for modelling shortest-path trees in telecommunication networks. In *Analytical and Stochastic Modeling Techniques and Applications*. Springer, Berlin, 324–336.

Marcel F. Neuts. 1989. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Taylor & Francis, New York.

António Nogueira, Paulo Salvador, Rui Valadas, and António Pacheco. 2003. Modeling self-similar traffic through markov modulated poisson processes over multiple time scales. In *High-Speed Networks and Multimedia Communications*. Springer, Berlin, 550–560.

Hiroyuki Okamura, Tadashi Dohi, and Kishor S. Trivedi. 2009. Markovian arrival process parameter estimation with group data. *IEEE/ACM Trans. Netw.* 17, 4 (2009), 1326–1339.

Sergio Pacheco-Sanchez, Giuliano Casale, Bryan Scotney, Sally McClean, Gerard Parr, and Stephen Dawson. 2011. Markovian workload characterization for qos prediction in the cloud. In *Proceedings of the 2011 IEEE International Conference on CLOUD*. IEEE, 147–154.

Andrew Patton. 2012. Copula methods for forecasting multivariate time series. *Handbook Econ. Forecast.* 2 (2012), 899–960.

Ali Rajabi and Johnny W. Wong. 2012. MMPP characterization of web application traffic. In *Proceedings of the 2012 IEEE 20th International Symposium on MASCOTS*. IEEE, 107–114.

Ali Rajabi and Johnny W. Wong. 2014. Provisioning of computing resources for web applications under time-varying traffic. In *Proceedings of the 2014 IEEE 22nd International Symposium on MASCOTS*. IEEE, 152–157.

Bruno Rémillard, Nicolas Papageorgiou, and Frédéric Soustra. 2012. Copula-based semiparametric models for multivariate time series. *J. Multivar. Anal.* 110 (2012), 30–42.

William J. J. Roberts, Yariv Ephraim, and Elvis Dieguez. 2006. On Rydén's EM algorithm for estimating MMPPs. *IEEE Signal Process. Lett.* 13, 6 (2006), 373–376.

Sheldon M. Ross. 2003. *Introduction to Probability Models*. Academic Press, Burlington.

Tobias Rydén. 1994. Parameter estimation for Markov modulated Poisson processes. *Stoch. Models* 10, 4 (1994), 795–829.

Tobias Rydén. 1996. An EM algorithm for estimation in Markov-modulated Poisson processes. *Computat. Stat. Data Anal.* 21, 4 (1996), 431–447.

Paulo Salvador, Rui Valadas, and António Pacheco. 2003. Multiscale fitting procedure using Markov modulated Poisson processes. *Telecommun. Syst.* 23, 1–2 (2003), 123–148.

Steven L. Scott. 2000. Detecting network intrusion using a Markov modulated nonhomogeneous poisson process. *Available Online* (2000). https://astro.temple.edu/~msobel/courses_files/scott-smythe.pdf.

Shou-Kuo Shao, Malla Reddy Perati, Meng-Guang Tsai, Hen-Wai Tsao, and Jingshown Wu. 2005. Generalized variance-based markovian fitting for self-similar traffic modelling. *IEICE Trans. Commun.* 88, 4 (2005), 1493–1502.

Pravin K. Trivedi and David M. Zimmer. 2007. *Copula Modeling: An Introduction for Practitioners*. Now Publishers Inc., Boston.

Maarten R. C. Van Oordt and Chen Zhou. 2012. The simple econometrics of tail dependence. *Econ. Lett.* 116, 3 (2012), 371–373.

Ury Yechiali and Pinhas Naor. 1971. Queuing problems with heterogeneous arrivals and service. *Operat. Res.* 19, 3 (1971), 722–734.

Tadafumi Yoshihara, Shoji Kasahara, and Yutaka Takahashi. 2001. Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telecommun. Syst.* 17, 1–2 (2001), 185–211.

Ming Yu and Mengchu Zhou. 2006. A model reduction method for traffic described by MMPP with unknown rate limit. *IEEE Commun. Lett.* 10, 4 (2006), 302–304.

Received April 2016; revised April 2017; accepted April 2017