

Pseudo Self-Similar Traffic: An Extended Markov Model

Z. K. Liu, V. Choulakian, J. Almhana
 University of Moncton
 Moncton, New Brunswick
 Canada E1A 3E9
 {liuz,choulav,almhanaj}@umoncton.ca

R. McGorman
 Nortel Networks
 4001 E. Chapel Hill-Nelson Hwy
 Research Triangle Park
 North Carolina USA 27709
 mcgorman@nortelnetworks.com

ABSTRACT

During the past decade, there have been a large number of papers studying high speed telecommunication network traffic modeling problem and a lot of traffic models have been proposed in the literature. Among them, the pseudo self-similar Markov models are of particular importance. In this paper, we propose an extension to the pseudo self-similar Markov models and Markov modulated Poisson process models are used. To capture the similarity property, the modulating Markov processes are assumed to have multiple-time scales, and at different time scales they are statistically similar. An extended EM algorithm is proposed to fit the model. To illustrate the usefulness of the proposed model, some preliminary experimental results are presented.

Categories and Subject Descriptors

C.2.5 [Local and Wide-Area Networks]: Internet

General Terms

Internet Traffic

Keywords

Internet traffic, self-similar process, Markov modulated Poisson process, EM algorithm

1. INTRODUCTION

In [4], Leland, Taqqu, Willinger and Wilson reported the self-similarity property of Ethernet traffic, which has had a profound effect on the area of high speed telecommunication network performance modeling. Since then, an explosion of work has ensued investigating the multifaceted nature of this phenomenon and a lot of models have been proposed. For more information about this area, we refer the reader to Willinger, Taqqu and Erramilli [11], which provides a comprehensive survey of the literature before 1996.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CNSR 2003 Conference, May 15-16, 2003, Moncton, New Brunswick, Canada.

Copyright 2003 CNSR Project 1-55131-080-5 ...\$5.00.

Among various models, the pseudo self-similar traffic model proposed by Robert and Le Boudec [8] is of special importance because of its simplicity. Their model is characterized by a discrete-time Markov chain, which requires only three parameters and can describe a wide range of self-similar behavior. They assume that the input process can be in one of n states. When in state 1, a single arrival occurs with probability 1. In all other states, no arrival occurs. The Markov chain governing the changes between states of the input process is such that transitions are made from state 1 to all states, and also from all states to state 1, but not between arbitrary states. The transition matrix of the Markov chain is given by

$$\begin{pmatrix} e_1 & \frac{1}{a} & \frac{1}{a^2} & \cdots & \frac{1}{a^{n-1}} \\ \frac{q}{a} & 1 - \frac{q}{a} & 0 & \cdots & 0 \\ \frac{q^2}{a^2} & 0 & 1 - \frac{q^2}{a^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ \frac{q^{n-1}}{a^{n-1}} & 0 & 0 & \cdots & 1 - \frac{q^{n-1}}{a^{n-1}} \end{pmatrix} \quad (1)$$

where $e_1 = 1 - 1/a - 1/a^2 - \cdots - 1/a^{n-1}$.

The above model is reported to work very well to model real systems. In this model, the underlying Markov chain has only one 'ON' state and many 'OFF' states. In this case, the duration of the 'ON' time is still geometrically distributed, which is not heavy-tailed. Since the underlying Markov chain has more than one 'OFF' state, the 'OFF' duration is not geometrically distributed, which may have long-memory. As pointed out in [6] in practice it is of more importance for the 'ON' duration to be heavy-tailed. In the literature of telecommunication network traffic modeling, many Markov-modulated Poisson process models have been studied by [7, 10, 12] and references therein. In this paper, we use Markov modulated Poisson processes to model the traffic, in which the underlying modulating processes have many 'ON' states and only one 'OFF' state. The modulating Markov process has many time scales. At different time scales, the subsystems have statistically similar characteristics. Since the state of the underlying Markov process is not observable, the proposed model is actually a hidden Markov model. So, the maximum likelihood approaches are used to fit the model. In practice, the state of the underlying Markov process is not accessible. So, we treat the sampled data as incomplete data and the EM algorithm is used to fit the system parameters. The rest of this paper is organized as follows: Section 2 gives some general description of self-similar processes. Section 3 provides a discrete-time model and gives two parameter fitting algorithms. Section

4 addresses the continuous-time model. Section 5 concludes the paper.

2. SELF-SIMILAR STOCHASTIC PROCESSES

Roughly speaking, a system is said to be self-similar if it behaves alike viewed at different time scales.

DEFINITION 2.1. *Stochastic process $X = \{X_t, t \geq 0\}$ is said to be self-similar with Hurst parameter H ($0 < H < 1$) if for all $a > 0$ and $t \geq 0$, the following holds*

$$X(t) \stackrel{d}{=} a^{-H} X(at) \quad (2)$$

that is, $X(at)$ and $a^H X(t)$ have the same distribution.

However, in practice for a given stochastic process it is very hard to check that (2) holds. So, we turn to consider the second-order properties of self-similar processes. To this end, let us give some definitions.

Function $g(x)$ is said to be slowly varying at infinity if the following holds for any $a > 0$

$$\lim_{x \rightarrow \infty} \frac{g(ax)}{g(x)} = 1. \quad (3)$$

Let $Y = \{y_t, t \geq 0\}$ be a second-order stationary process with the variance σ^2 . Its auto-correlation function $\gamma(k)$ is defined by

$$\gamma_y(k) = \frac{\text{Cov}(y_t, y_{t+k})}{\text{Var}(y_t)}. \quad (4)$$

DEFINITION 2.2. *Stochastic process Y is said to be second-order self-similar with Hurst parameter $H = 1 - \beta/2$ if*

$$\gamma_y(k) = \frac{1}{2} \left\{ (k+1)^{2H} - 2k^{2H} + (k-1)^{2H} \right\} \quad (5)$$

DEFINITION 2.3. *For stationary stochastic process Y , let its auto-correlation function $\gamma_y(k)$ be defined by (4). Process Y is said to be asymptotically second-order self-similar with Hurst parameter $H = 1 - \beta/2$ if*

$$\text{Var}(Y(m)) = g(m)m^{-\beta}$$

where $g(\cdot)$ is a slowly varying function.

Stochastic process Y is said to have long-range dependence if $\sum_{k=1}^{\infty} |\gamma_y(k)| = \infty$, or short-range dependence otherwise. From the definition of self-similarity and long-range dependence, it is easy to conclude that there are some self-similar processes that are not long-range dependent, and vice versa. However, in the case of asymptotic second-order self-similarity, by the restriction $1/2 < H < 1$ in the definitions, self-similarity and long-range dependence are equivalent.

3. DISCRETE TIME MODEL

3.1 Discrete time model description

Consider $m+1$ independent Markov chains $\{r_i(t), t = 0, 1, 2, \dots\}$, $0 \leq i \leq m$ with the same state space $\{0, 1\}$. We assume that Markov chain $r_i(t)$ has state transition matrix

$$P_i = \begin{pmatrix} 1 - p\varepsilon^i & p\varepsilon^i \\ q\varepsilon^i & 1 - q\varepsilon^i \end{pmatrix}.$$

Then, if we define an augmented stochastic process

$$\mathbf{a}_t = (r_0(t), r_1(t), \dots, r_m(t)),$$

then \mathbf{a}_t is also a Markov chain with state space $\mathcal{S} = \{0, 1\}^{(m+1)}$. For ease of manipulation, we define r_t to be the integer value of the binary vector \mathbf{a}_t . Then, $\{r_t\}$ is a Markov process with state space $\{0, 1, \dots, M\}$, where $M = 2^{m+1} - 1$. In order to define the state transition probability matrix of $\{r_t\}$, we need the notion of Kronecker product.

Let $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ and $B = (b_{kl}) \in \mathbb{R}^{s \times t}$ be two matrices. The Kronecker product of A and B , denoted by $A \otimes B$, is defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

Kronecker product has the following property (see [2])

PROPERTY 3.1. *Let A, B, C and D be four matrices with appropriate dimensions, then*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \quad (6)$$

With the above notations, it is easy to check that the transition probability matrix of $\{r_t\}$, $P \in [0, 1]^{M \times M}$, can be given by

$$P = P_0 \otimes P_1 \otimes \cdots \otimes P_m \quad (7)$$

A row probability distribution vector π , satisfying $\pi(i) \geq 0$ and $\sum_{i=0}^M \pi(i) = 1$, is said to be the stationary distribution of $\{r_t\}$ (or P) if

$$\pi P = \pi \quad (8)$$

Using the property (6), we have

$$\pi = \frac{1}{(p+q)^{m+1}} (q \ p)^{\otimes(m+1)}, \quad (9)$$

where $(q \ p)^{\otimes(m+1)}$ designates the $(m+1)^{th}$ Kronecker power of vector $(q \ p)$. In fact, by direct computation we find that the transition probability matrices P_0, P_1, \dots, P_m have the same stationary distribution $\mathbf{u} = \frac{1}{p+q} (q \ p)$, that is,

$$\mathbf{u}P_i = \mathbf{u}, i = 0, 1, \dots, m \quad (10)$$

which combined with the property (6) yields

$$\begin{aligned} \mathbf{u}^{\otimes(m+1)} P &= \overbrace{(\mathbf{u} \otimes \cdots \otimes \mathbf{u})}^{m+1} (P_0 \otimes P_1 \otimes \cdots \otimes P_m) \\ &= (\mathbf{u}P_0) \otimes \cdots \otimes (\mathbf{u}P_m) \\ &= \underbrace{\mathbf{u} \otimes \cdots \otimes \mathbf{u}}_{m+1} = \mathbf{u}^{\otimes(m+1)} \end{aligned} \quad (11)$$

which proves $\pi = \mathbf{u}^{\otimes(m+1)}$.

For each i , $0 \leq i \leq M$, let (i_0, i_1, \dots, i_m) be the binary expression of the integer i , then we define

$$\sigma(i) = i_0 + i_1 + \cdots + i_m \quad (12)$$

We assume that $\xi(i)$ is a Poisson process with arrival rate given by $\sigma(i)\lambda$ where λ is a positive constant, specifying the packet arrival rate at a specific time scale.

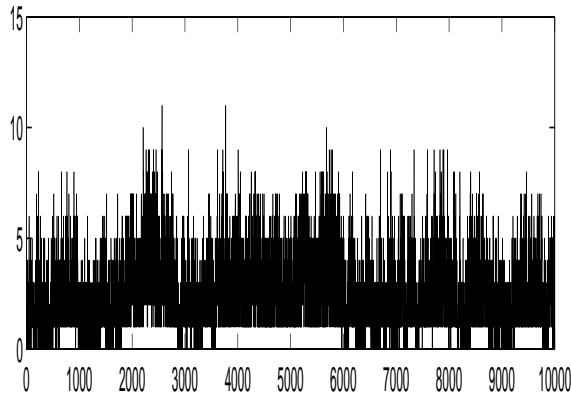


Figure 1: Simulated Traffic

Let x_t denote the number of packets arrived during the t^{th} time slot. We assume the following model for x_t

$$x_t = \begin{cases} \sigma(r_t) + \eta(r(t)), & \text{if } r(t) \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

REMARK 3.1. Note that in this model $\sigma(r_t)$ denotes the number of subsystems that are in state 'ON'. And, $\eta(r_t)$ represents the sum of $\sigma(r_t)$ independent Poisson processes with arrival rate λ .

In this model, there are five parameters: $p, q, \varepsilon, \lambda$, and m . By the Markov property, the 'ON' duration of Markov chain $r_i(t)$ has geometric distribution with parameter $q\varepsilon^i$. If we assume the arrival rate is λ , then at each burst the mean number of packets arrived is $\frac{\lambda}{q\varepsilon^i}$. Therefore, $r_i(t)$ viewed at the time scale $t\varepsilon^i$, looks like $r_0(t)$ at time scale t .

3.2 Self-Similarity Testing

In the last subsection, a pseudo self-similar process model is proposed. Next, we proceed to addressing the self-similarity of the model. For this purpose, let us first see an example.

EXAMPLE 3.1. Let us consider a system consisting of 4 levels, with parameters given as follows: $p = 0.5, q = 0.8, \varepsilon = 0.1, \lambda = 1$. With this set of data, a sample path with length 10000 is generated. The simulated sample path is plotted in Figure 1.

We are interested in viewing the system behaviors at different time scales. To this end, let us define two aggregated processes

$$y_1(t) = \sum_{k=(t-1)*10+1}^{t*10} y_t, \quad y_2(t) = \sum_{k=100(t-1)+1}^{100t} y_t$$

The first 100 terms of $y_1(t)$ and $y_2(t)$ are plotted in Figure 2.

From Figure 1 and Figure 2, we cannot declare that the simulated traffic is self-similar or has long-range dependence. To check the self-similarity, we compute the Hurst parameter H of the simulated data. If $H \in (0.5, 1)$, the process has long-range dependence. There are many approaches to estimate the Hurst parameter H (see [3] for details). Here, the aggregate variance method is utilized. The estimating procedure consists of three steps:

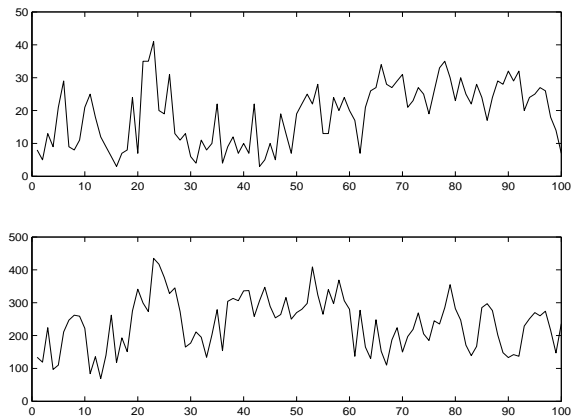


Figure 2: Aggregated Traffic

times scales	2	3	4	5
estimated H	0.7956	0.8035	0.8239	0.8385

Table 1: Hurst versus number of time scales

Step 1) Let n be the length of the time series $\{y_t\}$ under consideration. For $l = 2, 3, \dots, n/2$, divide the series into n/l subseries of size l . For each subseries, we calculate the sample mean $\{\bar{y}_k(l)\}$ using the following:

$$\bar{y}_k(l) = \frac{1}{l} \sum_{i=(k-1)l+1}^{kl} y_i, \quad k = 1, 2, \dots \quad (14)$$

and compute the overall mean by

$$\bar{y}(l) = \frac{1}{n/l} \sum_{k=1}^{n/l} (\bar{y}_k(l) - \bar{y}(l))^2. \quad (15)$$

Step 2) For each l , calculate the sample variance of the sample means $\{\bar{y}_k(l), k = 1, 2, \dots, n/l\}$, that is,

$$Var(\bar{y}(l)) = \frac{1}{n/l-1} \sum_{k=1}^{n/l} (\bar{y}_k(l) - \bar{y}(l))^2 \quad (16)$$

Step 3) Plot $\log(\bar{y}(l))$ versus $\log(l)$.

Following the above procedure, the estimated Hurst parameter $H = 0.9098$. The computation results are plotted in Figure 3. The upper figure represent $Var(\bar{y}(l))$ versus l (in dash line) and $f(l) = 1/l$ versus l (solid line). The lower figure represent $\log(Var(\bar{y}(l)))$ versus $\log(l)$ (in dash line) and $\log(f(l))$ versus $\log(l)$ (solid line). The figures show that $Var(\bar{y}(l))$ converges to zero with much slower rate than $1/l$, which means the simulated data are long-range dependent.

Next, we give some experimental results to illustrate the relationships between the parameters. First, we come to see how the number of time scales affects the Hurst parameter. To achieve this, we fixed parameters $p = 0.5, q = 0.8, \lambda = 1, \varepsilon = 0.1$ and consider the cases of $m = 1, 2, 3, 4$. The results are shown in Table 1. The results show that for this set parameters $p, q, \lambda, \varepsilon$, the Hurst parameter H increases with the number of time scales. The next experiment is to illustrate the relationship between ε and H . We fix $p = 0.5, q = 0.8, m + 1 = 3, \lambda = 1$, and compare the Hurst values

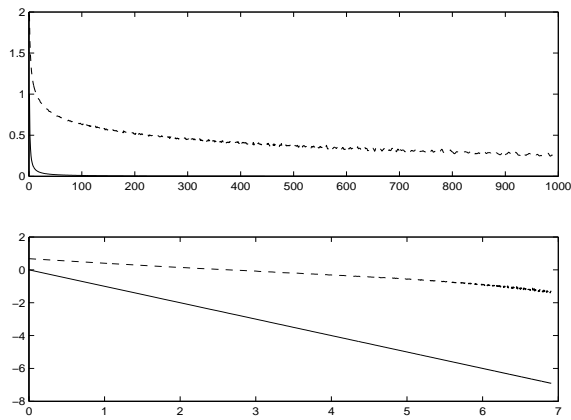


Figure 3: Aggregated Traffic

ε	0.1	0.3	0.5	0.7
estimated H	0.8035	0.5968	0.5152	0.5089

Table 2: Hurst Parameter versus ε

with $\varepsilon = 0.1, 0.3, 0.5$ and 0.7 . The experimental results are given in Table 2

Table 2 shows that for a set of given parameters $p, q, \lambda, \varepsilon$, the Hurst parameter H decreases when ε increases. The third experiment is to show the relationship between λ and H . We let $p = 0.5, q = 0.8, \varepsilon = 0.5, m = 2$, and consider $\lambda = 0.5, 0.7, 0.8, 0.9, 1.2$. The results are shown in Table 3. With given $p = 0.5, q = 0.8, \varepsilon = 0.5, m = 2$, H is not monotone with respect to λ . Instead, it may be a concave function of λ .

3.3 Model Fitting

This subsection addresses how to estimate the system parameters m, p, q, λ and ε . Note that the parameter m is an integer, designating the number of time scales. So, we can first fix m and estimate the other parameters corresponding to each given m . Then, choose the best one from the set of estimated parameters, which has the maximum likelihood value. Let \bar{s} be the maximum number of time scales to be considered, that is, $1 \leq m \leq \bar{s}$. The fitting procedure consists of the following two steps: Step 1. For $1 \leq m \leq \bar{s}$, estimate the parameters $p, q, \lambda, \varepsilon$, and denote their estimated values by $\hat{p}_m, \hat{q}_m, \hat{\lambda}_m, \hat{\varepsilon}_m$, respectively. Step 2. For all $1 \leq m \leq \bar{s}$, compute the likelihood values with parameters $\hat{p}_m, \hat{q}_m, \hat{\lambda}_m, \hat{\varepsilon}_m$, and choose the best one as the final result.

The main task in the above procedure is the first step, which can be divided into three substeps as follows: use the EM algorithm to estimate the stationary distribution π and parameter λ ; Once π is estimated, the second substep is to fit parameters p, q using π ; the third substep is to estimate ε .

λ	0.5	0.7	0.8	0.9	1.2
estimated H	0.4735	0.5524	0.5762	0.4570	0.4309

Table 3: Hurst parameter H versus λ

Let us consider the augmented stochastic process $\{(r_t, x_t), t = 0, 1, \dots\}$. Evidently, $\{(r_t, x_t)\}$ is a Markov process and $\{x_1, x_2, \dots, x_n\}$ are conditionally independent given $\{r_1, r_2, \dots, r_n\}$. In the sequel, we use $f(x_1, \dots, x_n, r_1, \dots, r_n)$, $f_x(x_1, \dots, x_n)$ and $f_r(r_1, \dots, r_n)$ to denote the likelihood function of $\{(x_t, r_t), 1 \leq t \leq n\}$, $\{x_t, 1 \leq t \leq n\}$ and $\{r_t, 1 \leq t \leq n\}$, respectively. If $f(x_t|r_t)$ is the conditional distribution of x_t given r_t , by virtue of (13), we obtain

$$f(x_t|r_t) = \begin{cases} b(x_t|r_t), & \text{if } x_t \geq \sigma(r_t) \text{ and } r_t \neq 0 \\ 1, & \text{if } x_t = 0 \text{ and } r_t = 0 \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where

$$b(x_t|r_t) = \frac{(\sigma(r_t)\lambda)^{x_t - \sigma(r_t)} e^{-\sigma(r_t)\lambda}}{(x_t - \sigma(r_t))!}.$$

If a complete sample path of $\{(r_t, x_t)\}$ is available, we can apply the maximum likelihood approach directly. However, in practice we can observe the number of packets received, but we have no access to the state of $\{r_t\}$, which means that the sampled data are incomplete. As a result, the maximum likelihood approach cannot be applied directly. However, the EM algorithm (see [5] for more details) can be applied to this problem. The EM algorithm is a very useful interactive method for maximizing log-likelihoods which are awkward because there are missing data. Roughly speaking, the EM algorithm calculates the log-likelihood which we would expect to see, given the current update of the maximum likelihood estimate. The next update of the maximum likelihood estimate is obtained by maximizing this expected log-likelihood, which is usually straightforward, as the missing data have been replaced by their expected values.

Next, we give the details of each step.

3.3.1 Step 1. Estimating parameters with given number of time scales

As pointed out before, this step consists of three sub-steps.

Substep 1.1. Estimate π and λ . Note that in the above model, only the stationary distribution π contains parameters p, q , and only $f(x_t|r_t)$ contains parameter λ . So, for simplicity, when applying the EM algorithm, we assume the system parameters

$$\theta = (\pi(0), \dots, \pi(M), \lambda).$$

If we assume that the states $\{r_1, \dots, r_n\}$ are also available, then the likelihood function of the complete-data $\{(r_t, x_t), 1 \leq t \leq n\}$ is given by

$$\begin{aligned} L_n^c(\theta) &= f(x_1, x_2, \dots, x_n, r_1, r_2, \dots, r_n) \\ &= \prod_{t=1}^n \pi(r_t) f(x_t|r_t) \end{aligned} \quad (18)$$

The EM algorithm consists of two steps: E-step and M-step, which are described in the following.

The E-step finds the expected value of the complete-data log-likelihood $\log(L_n^c(\theta))$ with respect to the unknown data $\{r_t, 1 \leq t \leq n\}$ given the observed data $X = \{x_1, \dots, x_n\}$ and the current parameter estimations. That is, we define

$$\Phi(\theta, \theta^k) \triangleq \mathbb{E} \left[\log(L_n^c(\theta)) | X, \theta^k \right] \quad (19)$$

where $\theta^{(k)} = (\pi^{(k)}, \lambda^{(k)})$ denotes the estimated parameters at the k^{th} iteration. The second step (the M-step) of the

EM algorithm is to maximize the expectation computed in the first step. That is, to solve the following optimization problem

$$\theta^{k+1} = \operatorname{argmax}_{\theta} \Phi(\theta, \theta^k) \quad (20)$$

These two steps are repeated as necessary. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

Direct computation gives

$$\begin{aligned} \Phi(\theta, \theta^k) &= \sum_{t=1}^n \mathbb{E} \left[\log(\pi(r_t)) + \log(f(r_t|x_t)) | x_t, \theta^k \right] \\ &= \sum_{i=0}^M \sum_{t=1}^n [\log(\pi(i)) + \log(f(i|x_t))] f(i|x_t, \theta^k) \end{aligned} \quad (21)$$

where

$$f(i|x_t, \theta^k) = \frac{\pi^{(k)}(i) f(x_t|i, \theta^{(k)})}{\sum_{l=0}^M \pi^{(k)}(l) f(x_t|l, \theta^{(k)})} \quad (22)$$

with

$$f(x_t|i, \theta^{(k)}) = \begin{cases} \frac{(\sigma(i)\lambda^{(k)})^{x_t - \sigma(i)} e^{-\sigma(i)\lambda^{(k)}}}{(x_t - \sigma(i))!}, & \text{if } x_t \geq \sigma(i) \text{ and } i \neq 0 \\ 1, & \text{if } x_t = 0 \text{ and } i = 0 \\ 0, & \text{otherwise} \end{cases}$$

With the expectation calculated, we come to update the parameters, that is, to solve the optimization problem (20). To find the expression for $\pi^{(k+1)}(i)$, we introduce the Lagrange multiplier γ with constraint $\sum_{i=0}^M \pi(i) = 1$, and solve the following equation:

$$\frac{\partial}{\partial \pi(i)} \left[\sum_{i=0}^M \sum_{t=1}^n \log(\pi(i)) f(i|x_t, \theta^k) + \gamma \left(\sum_{i=0}^M \pi(i) - 1 \right) \right] = 0 \quad (23)$$

yielding

$$\sum_{t=1}^n \frac{1}{\pi(i)} f(i|x_t, \theta^k) + \gamma = 0$$

or

$$\sum_{t=1}^n f(i|x_t, \theta^k) + \gamma \pi(i) = 0, 0 \leq i \leq M \quad (24)$$

Summing both sides of the above equation over i yields $\gamma = -n$, which combined with (24) yields the update estimate of $\pi^{(k+1)}(i)$:

$$\pi^{(k+1)}(i) = \frac{1}{n} \sum_{t=1}^n f(i|x_t, \theta^k) \quad (25)$$

To find $\lambda^{(k+1)}$, let $\frac{\partial \Phi(\theta, \theta^k)}{\partial \lambda} = 0$, that is,

$$\begin{aligned} \frac{\partial}{\partial \lambda} \left[\sum_{i=0}^M \sum_{t=1}^n [(x_t - \sigma(i))(\log(\sigma(i)) + \log(\lambda))] \right. \\ \left. - \lambda \sigma(i) \right] f(i|x_t, \theta^k) = 0 \end{aligned} \quad (26)$$

from which it follows that

$$\lambda^{(k+1)} = \frac{\sum_{i=0}^M \sum_{t=1}^n [x_t - \sigma(i)] f(i|x_t, \theta^k)}{\sum_{i=0}^M \sum_{t=1}^n \sigma(i) f(i|x_t, \theta^k)} \quad (27)$$

Substep 1.2. Estimate p and q . After the EM algorithm converges, we obtain the estimated values π and λ , denoted by $\hat{\pi}$ and $\hat{\lambda}$, respectively. The next step is to fit the system parameters p, q according to $\hat{\pi}$ by solving the following optimization problem

$$\min_{p, q} \sum_{i=0}^M (\hat{\pi}(i) - \pi(i))^2. \quad (28)$$

Substep 1.3 Estimate ε . With $\hat{p}, \hat{q}, \hat{\lambda}$ available, we proceed to estimating ε . Note that L_n^c does not contain parameter ε . To estimate ε we need more information. On the other hand, we find that only when r_t is in state 0, the arrival process is in 'OFF' state and there is no packet arrival. Let N denote the number of '0'-runs, and let $\xi_t, 1 \leq t \leq N$, be the length of the t -th '0'-run. Using the Markov property of $\{r_t\}$, $\{\xi_t, 1 \leq t \leq N\}$ are independent and identically distributed random variables having the distribution

$$P(\xi_k = i) = \sigma^{i-1}(1 - \sigma), \quad i = 1, 2, \dots$$

where

$$\sigma = \prod_{i=0}^m (1 - p\varepsilon^i)$$

Let $L_N^2(\sigma)$ be the likelihood function of $\{\xi_t\}$ defined by

$$L_N^2(\sigma) = \prod_{k=1}^N [\sigma^{\xi_k - 1} (1 - \sigma)] \quad (29)$$

which contains the parameters p and ε . Note that

$$\log(L_N^2) = \sum_{k=1}^N (\xi_k - 1) \log(\sigma) + N \log(1 - \sigma)$$

and

$$\frac{\partial \log(L_N^2)}{\partial \sigma} = \frac{1}{\sigma} \sum_{k=1}^N (\xi_k - 1) - \frac{N}{(1 - \sigma)}$$

To maximize $\log(L_N^2(\sigma))$, solving $\frac{\partial}{\partial \sigma} [\log(L_N^2(\sigma))] = 0$ yields

$$\hat{\sigma} = 1 - \frac{N}{\sum_{k=1}^N \xi_k} \quad (30)$$

With \hat{p} and $\hat{\sigma}$ obtained, solving

$$\hat{\sigma} = \prod_{i=0}^m (1 - \hat{p}\varepsilon^i) \quad (31)$$

yields time scale parameter $\hat{\varepsilon}$.

3.3.2 Step 2. Computing the likelihood values

With parameters $(\hat{p}_m, \hat{q}_m, \hat{\lambda}_m, \hat{\varepsilon}_m)$, $1 \leq m \leq \bar{s}$, available, we proceed to the Step 2—calculate their likelihood values and choose the best one.

Since there is no observation to (r_1, r_2, \dots, r_n) available, we cannot use L_n^c . Instead, we consider the likelihood function of $\{x_1, x_2, \dots, x_n\}$, $L_n(p, q, \lambda) = f_x(x_1, x_2, \dots, x_n)$. Upon direct computation we, we obtain

$$\begin{aligned} L_n &= \sum_{(r_1, \dots, r_n)} f(x_1, \dots, x_n | r_1, \dots, r_n) f_r(r_1, \dots, r_n) \\ &= \left[\sum_{r_1} f_r(r_1) f(x_1 | r_1) \right] \sum_{(r_2, \dots, r_n)} \prod_{k=2}^n f(x_k | r_k) p_{r_{k-1} r_k} \end{aligned} \quad (32)$$

where $\sum_{(r_2, \dots, x_n)}$ means summation for all the possible values of (r_2, \dots, x_n) . Since the underlying Markov process is assumed to be stationary and its states are not observable, that is, the observed sample path contains no observation of $\{r_t\}$, we use the stationary distribution $\pi = (\pi(0), \dots, p(M))$ to replace its state transition probabilities. In this case, $L_n(p, q, \lambda)$ becomes

$$\begin{aligned} L_n(p, q, \lambda) &= \sum_{(r_1, r_2, \dots, x_n)} \prod_{k=1}^n f(x_k | r_k) \pi(r_k) \\ &= \prod_{k=1}^n \left(\sum_{j=0}^M \pi(j) f(x_k | j) \right) \end{aligned} \quad (33)$$

Therefore, we have

$$\begin{aligned} \log(L_n(p, q, \lambda)) &= \sum_{k=1}^n \log \left(\sum_{j=0}^M \pi(j) f(x_k | j) \right) \\ &= \sum_{k=1, x_k > 0}^n \log \left(\sum_{j=0, x_t \geq \sigma(j)}^M \pi(j) b(x_k | j) \right) \\ &\quad + N_0 \log(\pi(0)) \end{aligned} \quad (34)$$

where N_0 is the number of zeros in (x_1, \dots, x_n) .

Now, let us summarize the above procedure as an algorithm.

ALGORITHM 3.1. (Parameter Fitting Algorithm)

Step 1) For all $1 \leq m \leq \bar{s}$, estimate $\hat{p}_m, \hat{q}_m, \hat{\lambda}_m, \hat{\varepsilon}_m$.

- 1.1) Use the EM-algorithm to estimate $\hat{\pi}_m, \hat{\lambda}_m$
- 1.2) Fit \hat{p}_m, \hat{q}_m by solving (28).
- 1.3) With \hat{p} obtained, solve equation (31) to get $\hat{\varepsilon}_m$.

Step 2) Compute the likelihood values

- 2.1 For all $1 \leq m \leq \bar{s}$, compute the likelihood

$$L_n(\hat{p}_m, \hat{q}_m, \hat{\lambda}_m, \hat{\varepsilon}_m).$$

- 2.2 Find the optimal value

$$\begin{aligned} (\hat{p}, \hat{q}, \hat{\lambda}, \hat{\varepsilon}) &= \operatorname{argmax} \{ L_n(\hat{p}_m, \hat{q}_m, \hat{\lambda}_m, \hat{\varepsilon}_m), \\ &\quad 1 \leq m \leq \bar{s} \} \end{aligned}$$

To see how the above procedure works, let us work out an example.

EXAMPLE 3.2. We use the same routine that generates the data of Example 3.1 to generate another set of data with the parameters $m = 3, p = 0.5, q = 0.8, \lambda = 1, \varepsilon = 0.25$. Then, we estimate the system parameters using the simulated data. For this purpose, we choose

$$\pi^{(0)} = \left(\frac{1}{M} \quad \frac{1}{M} \quad \dots \quad \frac{1}{M} \right)$$

and $\lambda^{(0)} = 0.9$ and use the stop rule

$$\max(\max(\pi^{(k+1)} - \pi^{(k)}), \lambda^{(k+1)} - \lambda^{(k)}) < 0.003.$$

When $m = 3$, the substep 1.1 yields

$$\begin{aligned} \hat{\pi}_3 &= [0.1529, 0.0881706, 0.0881706, 0.0523626, \\ &\quad 0.0881706, 0.0523626, 0.0523626, 0.0400955, \\ &\quad 0.0881706, 0.0523626, 0.0523626, 0.0400955, \\ &\quad 0.0523626, 0.0400955, 0.0400955, 0.0198597] \end{aligned}$$

m	\hat{p}_m	\hat{q}_m	$\hat{\lambda}_m$	$\hat{\varepsilon}_m$
1	0.5920	0.6820	1.62709	x
2	0.5104	0.592	1.1837	0.218
3	0.484	0.790	0.978504	0.28
4	0.22	0.474	0.8828	0.829

Table 4: Estimated Parameters

m	2	3	4
$\log(L(\hat{p}_m, \hat{q}_m, \hat{\lambda}_m))$	-22885.6	-22692.1	-22974.8

Table 5: log-likelihood values

and $\hat{\lambda} = 0.978504$. Direct computation gives $\max(\hat{\pi} - \pi) = 0.0021$. Solving (28) gives $\hat{p} = 0.4840, \hat{q} = 0.79$. Moreover, we find in the simulated sample path there are $N = 880$ '0'-runs with mean $\bar{\xi} = 1.7386$. So, we have $\hat{\sigma} = 0.4248$. Solving (31) yields $\hat{\varepsilon} = 0.28$.

Repeat the above procedure with $m = 1, 2, 4$. When $m = 1$, equation (31) has no solution. The obtained results are given in Table 4.

The likelihood values of the above parameters are computed, which are given in Table 5. Based on the results given by Table 5 and Table 4, we conclude that the estimated system parameters are $\hat{m} = 3, \hat{p} = 0.484, \hat{q} = 0.790, \hat{\lambda} = 0.978504, \hat{\varepsilon} = 0.28$.

4. CONTINUOUS-TIME MODEL

4.1 Continuous-time model description

This section considers a continuous time version of the model studied in the previous section. As before, consider a system consisting of $m+1$ time scales, which are denoted by $t, \frac{t}{\varepsilon}, \frac{t}{\varepsilon^2}, \dots, \frac{t}{\varepsilon^m}$, where $\varepsilon > 0$ is a positive constant denoting the time scale. At the i th time scale t/ε^i , the packet arrival process $x_i(t)$ is a Markov modulated Poisson process. The underlying switching Markov process $r_i(t)$ is a two state Markov process with state space $\{0, 1\}$ and generator matrix

$$Q_i = \frac{1}{\varepsilon^i} Q$$

where $Q = \begin{pmatrix} -p & p \\ q & -q \end{pmatrix}$ with $p > 0, q > 0$ being constants.

When $r_i(t) = 1$, the packets arrive with rate λ ; otherwise, no packet arrives.

Define an augmented Markov process

$$r_t = (r_0(t), r_1(t), \dots, r_m(t)).$$

Let $\Lambda = (\lambda_{ij}), i, j \in \{0, 1, \dots, M\}$, be the infinitesimal generator matrix of r_t , and denote $\lambda(i) = -\lambda_{ii}, i = 0, 1, \dots, M$. Then, by appropriately arranging the states, the infinitesimal generator matrix of r_t can be represented by

$$\Lambda = \Lambda_0 + \frac{1}{\varepsilon} \Lambda_1 + \dots + \frac{1}{\varepsilon^m} \Lambda_m \quad (35)$$

where $\Lambda_j \in \mathbb{R}^{2^{(m+1)} \times 2^{(m+1)}}$, $0 \leq j \leq m$, are $m+1$ infinitesimal generator matrices. Let I denote the two-dimensional identity, \otimes denote the Kronecker product operator, and $I^{\otimes m}$ be the Kronecker product of m I s. With these notations, Λ_i

can be given by

$$\begin{aligned}\Lambda_0 &= Q \otimes I^{\otimes m} \\ \Lambda_1 &= I \otimes Q \otimes I^{\otimes(m-1)} \\ &\vdots \\ \Lambda_i &= I^{\otimes i} \otimes Q \otimes I^{\otimes(m-i)} \\ &\vdots \\ \Lambda_m &= I^{\otimes m} \otimes Q\end{aligned}$$

We assume the system is represented by

$$x_t = \begin{cases} \sigma(r_t) + \eta(r_t), & \text{if } r_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (36)$$

where $\{r_t\}$ is Markov process with state space $\{0, 1, \dots, M\}$ and for given state i , $\eta(i)$ has Poisson distribution. When $r_t = (i_0 i_1 \dots i_m) \in \{0, 1\}^{m+1}$, the arrival rate of $\eta(r_t)$ is $\sigma(r_t)\lambda$.

For example, in the case of $m = 1$, that is, two time scales considered,

$$\begin{aligned}\Lambda &= \begin{pmatrix} -p - \frac{1}{\varepsilon}p & \frac{1}{\varepsilon}p & p & 0 \\ \frac{1}{\varepsilon}q & -p - \frac{1}{\varepsilon}q & 0 & p \\ q & & -q - \frac{1}{\varepsilon}p & -\frac{1}{\varepsilon}p \\ 0 & q & \frac{1}{\varepsilon}q & -q - \frac{1}{\varepsilon}q \end{pmatrix} \\ &= Q \otimes I + \frac{1}{\varepsilon}I \otimes Q\end{aligned}$$

4.2 Parameter Fitting

Let $\{(J_t, T_t), t = 1, 2, \dots\}$ denote a sample path of $\{r_t\}$, where J_t denotes the state visited after the t -th state transition, and T_t denotes the duration of staying on state J_t . By the construction of the model, $\{x_1, \dots, x_n\}$ are conditionally independent given $\{(J_1, T_1), \dots, (J_n, T_n)\}$. In practice, the observed data consist of a sequence of interarrival times T_t and numbers of packets or bits observed, but the system mode J_t is not observable. So, let us assume the underlying Markov process is stationary. Let $L_n = f((T_1, x_1), \dots, (T_n, x_n))$ denote the likelihood function of the observed sample path $\{(T_1, x_1), \dots, (T_n, x_n)\}$. Since $\{(T_1, x_1), \dots, (T_n, x_n)\}$ are conditionally independent given $\{J_1, \dots, J_n\}$, we have

$$\begin{aligned}L_n &= \sum_{(J_1, J_2, \dots, J_n)} f((T_1, x_1), \dots, (T_n, x_n) | (J_1, \dots, J_n)) \\ &\quad \times f(J_1, \dots, J_n) \\ &= \sum_{(J_1, J_2, \dots, J_n)} f(J_1, \dots, J_n) \prod_{k=1}^n f((T_k, x_k) | J_k) \\ &= \sum_{(J_1, J_2, \dots, J_n)} f(J_1, \dots, J_n) \\ &\quad \times \prod_{k=1}^n f((T_k | J_k) f(x_k | J_k, T_k))\end{aligned} \quad (37)$$

where

$$f(x_k | J_k, T_k) = \begin{cases} \frac{(\sigma(J_k)\lambda T_k)^{x_k - \sigma(J_k)} e^{-\sigma(J_k)\lambda T_k}}{(x_k - \sigma(J_k))!}, & \text{if } x_k \geq \sigma(J_k), J_k \neq 0 \\ 1, & \text{if } x_k = 0 \text{ and } J_k = 0 \\ 0, & \text{otherwise} \end{cases}$$

Since no observation to $\{J_k\}$ is available and the Markov chain $\{J_t\}$ is assumed to be stationary, we use the stationary distribution $\prod_{k=1}^n \pi(J_k)$ to take the place of $f(J_1, \dots, J_n)$. In this case, the likelihood becomes

$$\begin{aligned}L_n &= \sum_{(J_1, J_2, \dots, J_n)} \prod_{k=1, x_k \geq \sigma(J_k)}^n \left[\pi(J_k) \lambda(J_k) e^{-\lambda(J_k) T_k} \right. \\ &\quad \left. \times \frac{[\sigma(J_k)\lambda T_k]^{x_k - \sigma(J_k)} e^{-[\sigma(J_k)\lambda T_k]}}{(x_k - \sigma(J_k))!} \right] \\ &= \prod_{k=1}^n \left[\sum_{i=0, \sigma(i) \leq x_k}^M \pi_i \lambda_i e^{-\lambda_i T_k} \frac{[\sigma(i)\lambda T_k]^{x_k - \sigma(i)} e^{-\sigma(i)\lambda T_k}}{(x_k - \sigma(i))!} \right]\end{aligned}$$

Let $\tilde{\pi}$ be the stationary distribution of r_t , that is, $\tilde{\pi} = \frac{1}{(p+q)^{m+1}} (q \ p)^{\otimes(m+1)}$. Note that $\{J_t\}$ is the jump Markov chain of $\{r_t\}$. So, we have

$$\pi = C \pi \text{diag}\{\lambda(0), \dots, \lambda(M)\} \quad (38)$$

where C is a normalizing constant. The above likelihood L_n becomes

$$L_n = \prod_{k=1}^n \left[\sum_{i=0, \sigma(i) \leq x_k}^M \tilde{\pi}_i e^{-\lambda(i) T_k} \frac{[\sigma(i)\lambda T_k]^{x_k - \sigma(i)} e^{-\sigma(i)\lambda T_k}}{(x_k - \sigma(i))!} \right]$$

The optimization problem is to find p, q, λ and ε to maximize L_n , or maximize $\log(L_n)$

$$\begin{aligned}\log(L_n) &= \sum_{k=1}^n \log \left(\sum_{i=0, \sigma(i) \leq x_k}^M \tilde{\pi}_i e^{-\lambda_i T_k} \right. \\ &\quad \left. \times \frac{[\sigma(i)\lambda]^{x_k - \sigma(i)} e^{-\sigma(i)\lambda T_k}}{(x_k - \sigma(i))!} \right)\end{aligned} \quad (39)$$

The same procedure of fitting the discrete time model can be developed for the continuous time model.

5. CONCLUDING REMARKS AND FUTURE WORK

This paper proposed a multiple time scale Markov process modulated Poisson process to model the Internet traffic. In order to characterize the self-similarity of the traffic, in the proposed model the underlying Markov processes have many time scales, and the subsystems of different time scales have the same statistic characteristics. For the discrete time model, an extended EM algorithm is proposed to fit the system parameters.

To extend the modulating Markov chain to high order Markov chain is an interesting research direction. A discrete time random process $\{r_t, t = 1, 2, \dots\}$, with $r_t \in S = \{0, 1, 2, \dots\}$, is said to be a Markov chain of order w if its probability distribution satisfying

$$P(r_{t+1} | r_t, \dots, r_1) = P(r_{t+1} | r_t, \dots, r_{t-w})$$

that is, the next step state transition depends on the past w steps. Obviously, when $w > 1$, w -order Markov chain is a natural generation to first order Markov chain, and obviously, it has longer dependence than the first order Markov chain. So, replacing the underlying first order Markov chain with a high order Markov chain in the model proposed in the paper will produce a useful model.

6. REFERENCES

- [1] A. T. Andersen and B. F. Nielsen, A Markovian Approach for Modeling Packet traffic with long-range dependence, *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 5, 1998.
- [2] R. Bellman, *Introduction to matrix analysis*. New York: McGraw-Hill, 1960, chapter 20.
- [3] J. Beran, Statistics for long-memory processes. Monographs on statistics and Applied Probability, Chapman and Hall, New York, 1994.
- [4] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, On the self-similar nature of Ethernet traffic. *Proc. ACM SIGCOMM'93*, pp. 183-193, 1993.
- [5] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, New York: Wiley, c1997.
- [6] K. Park and W. Willinger, Self-similar network traffic and performance evaluation, John Wiley and Sons, Inc. New York, 2001.
- [7] R. Ritke, X. Hong and M. Gerla, Contradictory relationship between Hurst parameter and queuing performance(extended version), *Telecommunication Systems*, Vol. 16, 159-175, 2001.
- [8] S. Robert and J. Y. L. Boudec, New models for pseudo self-similar traffic, *Performance Evaluation*, Vol. 30, 57-68, 1997.
- [9] S. L. Scott, P. Smyth, The Markov modulated Poisson process and Markov Poisson cascade with applications to Web traffic modeling, *Bayesian Statistics*, Vol. 7, 2003.
- [10] S. H. Shahram and L. N. Tho, MMPP models for multimedia traffic, *Telecommunication Systems*, 15, pp. 283-293, 2000.
- [11] W. Willinger, M. Taqqu, and A. Erramilli, A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. In F. P. Kelly, S. Zachary, and I. Ziedins, eds. *Stochastic Networks: Theory and Applications*, pp. 339-366, Clarendon Press, Oxford, UK, 1996.
- [12] T. Yoshihara, S. Kasahara and Y. Takahashi, Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process, *Telecommunication Systems*, Vol. 17, pp. 185-211, 2001.