An HMM Approach to Internet Traffic Modeling

Alberto Dainotti*, Antonio Pescapè*, Pierluigi Salvo Rossi*, Giulio Iannello†, Francesco Palmieri‡, Giorgio Ventre*
* Univ. di Napoli "Federico II" (Italy), † Univ. Campus Bio-Medico di Roma (Italy), ‡ Seconda Univ. di Napoli (Italy)
* {alberto,pescape,salvoros,giorgio}@unina.it, † g.iannello@unicampus.it, ‡ francesco.palmieri@unina2.it

***Abstract*— Traffic modeling is a fertile research area. This paper proposes a packet-level traffic model of traffic sources based on Hidden Markov Model. It has been developed by using real network traffic and estimating in a combined fashion Packet Size and Inter Packet Time. The effectiveness of the proposed model is evaluated by studying several traffic types with strong differences in terms of both applications/users and protocol behavior. Indeed, we applied our model to real traffic traces of Age of Mythology (a Multi Player Network Game), SMTP, and HTTP. An analytical basis and the mathematical details regarding the model are given. Results show how the proposed model captures first-order statistics, as well as temporal dynamics via auto- and cross-correlation. Also, the capability to accurately replicate the considered traffic sources is shown. Finally, preliminary results for model-based traffic prediction reveal encouraging.**

I. Introduction

Internet traffic modeling is an important and essential task to understand and solve performance-related issues of current and future networks. Many efforts have been focused on modeling of source traffic related to specific application-level protocols, also with the purpose to conduct realistic network traffic simulations and emulations (i.e. generating synthetic traffic in real networks). Although it has been often overlooked by the networking community, packet-level analysis offers very interesting insights [1] [2] [3]. Packet-level traffic models express traffic flows in terms of Inter Packet Time (IPT) and Packet Size (PS), basing the analysis on few simple variables, but related to the lowest/deepest point of view. Network devices (Routers, Switches, Access Points) often operate on a packet-by-packet basis (i.e. buffer management), and network problems (Loss, Delay, Jitter) happen at packet level. Other advantages of studying traffic by observing IPT and PS are the avoidance of any assumption regarding the application layer protocol characteristics, and the possibility to study, in the same manner, different kind of sources and even mixes of them.

The modeling approach here proposed relies on a Hidden Markov Model (HMM) on the basis of a more general Bayesian approach that has revealed being well-performing in packet-channels modeling [4]. The idea is that similar Bayesian models may be used for effective modeling of packet-level environments, thus obtaining a powerful and homogeneous analytical framework for heterogeneous scenarios (both in terms of traffic sources and end-to-end network paths). Indeed, in this work we investigate on HMMs capabilities (via learning, modeling, and prediction) to construct realistic packet-level models from empirical traffic traces considering the marginal distributions and the auto and mutual covariances of IPT and PS. We apply our approach to several traffic traces coming from various application-layer protocols and representing very different Internet applications. More precisely we apply our model to (i) traffic generated by *Age of Mythology* (AoM) [5], which is a Multi Player Network Game; (ii) SMTP traffic; (iii) HTTP traffic.

The rest of the paper is organized as follows. In Section II a briefly description of the motivations is given. Section III provides details on the proposed analytical model, providing insights about model statistics and the learning stage. Section IV describes the measurement and modeling approach. In Section V we show results of AoM, SMTP, and HTTP traffic modeling. Section VI presents preliminary results on the prediction capability of the proposed model. Section VII ends the paper with conclusion remarks.

II. Motivation

This work proposes a packet-level model, based on HMM, of traffic sources. To the best of our knowledge, few works using HMMs at packet level are present in literature. We found approaches to Internet traffic modeling able to capture temporal structures based on MMPP (Markov Modulated Poisson Process) [6] and BMAP (Batch Markovian Arrival Process) [7] [2]. Recently an interest in HMM-based models has grown [8], though their interest is focused on higher-level profile modeling. HMM packet-level modeling has been partially explored in [1], where a network traffic classification is proposed, and in [9], where IPT and PS of both aggregated and WWW traffic are disjointly analyzed. Therefore, an HMM packet-level model provides the following benefits when compared with higher level approaches: (i) simple/concise and at the deepest point of view; (ii) switching devices often operate on a packet-by-packet basis; (iii) most network performance problems (e.g. Loss, Delay, Jitter) happen at packet level; (iv) it is independent of protocols evolution and it is applicable to different applications/protocols; (v) it is usable in traffic generators and simulators; (vi) traffic at packet level remains observable after encryption made by, for example, end-to-end cryptographic protocols such as SSL or IPSec; (vii) packet-level traffic models make robust approaches to traffic profiling for anomaly detection. Also, differently from [1], this paper points the attention on accurate source traffic modeling for replicating real traffic. To this purpose, in addition to the first

[0]This work has been partially supported by PRIN 2004 Quasar Project, by CONTENT NoE, Onelab and NETQOS EU projects.

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

order statistics, correlations are taken into account to infer traffic dynamics and the capability to replicate and predict the source behavior is provided.

To highlight the significance of the proposed approach we underline that, to the best of our knowledge, it extends the results present in literature in that: **(I)** our first attempt to apply HMM to packet-level modeling represents an early success in the field of traffic modeling; **(II)** it allows IPT/PS joint description; **(III)** it allows prediction; **(IV)** it is derived by real traffic; **(V)** it explicitly takes into account the traffic dynamics over the network; **(VI)** it has been tested on three different traffic types (quite different from each other in terms of both used protocols and users/applications behavior), deriving analogies and differences on the equivalent traffic models; **(VII)** results obtained with the analyzed traffic types make the proposed model generalizable; **(VIII)** as regards games traffic, there exist a number of good traffic characterizations [10] [11] but there are not much models.

III. The Analytical Model

We propose a statistical model for packet-level network traffic. We consider an HMM in which the state variable is discrete[1], $x_n \in \{s_1, \ldots, s_N\}$, and the observable variable is a continuous bi-dimensional vector, $\mathbf{y}_n = [d_n, b_n]^T$. The first and second components of $\mathbf{y}_n$ represent $10\log_{10}(\mathrm{IPT}/1\mu\mathrm{s})$ and the PS for the n-th packet, respectively in dBμ and in bytes. We measure IPT with a resolution of $1\mu s$ (as explained in Section IV) and apply a logarithmic transformation because they range over several orders of magnitude. The state variable has been introduced to account for memory and correlation phenomena between IPT and PS. We assumed that IPT and PS are statistically independent given the state. $\Lambda = \{\mathbf{A}, \mathbf{g}^{(t)}, \mathbf{w}^{(t)}, \mathbf{g}^{(p)}, \mathbf{w}^{(p)}\}$ is the set of parameters characterizing the model, denoting the state transition matrix, the conditional IPT and PS distribution vectors respectively, i.e.

- $A_{ij} = Pr(x_{n+1} = s_j | x_n = s_i)$;
- $d_n | x_n = s_i \sim Gamma(g_i^{(t)}, w_i^{(t)})$;
- $b_n | x_n = s_i \sim Gamma(g_i^{(p)}, w_i^{(p)})$;

then the conditional pdf's for IPT and PS are[2]:

$$f_i^{(t)}(d) = \frac{(d/w_i^{(t)})^{g_i^{(t)}-1} e^{-(d/w_i^{(t)})}}{w_i^{(t)}\Gamma(g_i^{(t)})} (d>0) ,$$

$$f_i^{(p)}(b) = \frac{(b/w_i^{(p)})^{g_i^{(p)}-1} e^{-(b/w_i^{(p)})}}{w_i^{(p)}\Gamma(g_i^{(p)})} (b>0) .$$

Summarizing we have a model where x_n is a discrete random variable whose dynamic behavior is governed by the transition matrix $\mathbf{A}$, with a Markovian assumption for the evolution, and $\mathbf{y}_n$ is a bi-dimensional continuous random variable describing IPT and PS as mixtures of conditionally independent (given the state) Gamma distributions.

[1] *Notation* - Upper and lower bold case letters denote respectively matrices and column vectors, $[.]^T$ and $\mathcal{E}\{.\}$ denote transpose and expectation.

[2] The choice of Gamma distributions for IPT and PS is because a mixture of normal distributions can easily approximate a general distribution, Gamma is practically very similar to a normal distribution and has the desirable characteristic to be null for negative values (being negative IPT and PS meaningless).

A. Model Statistics

Denoting $\mathbf{q} = [q_1, \ldots, q_N]^T$ the steady-state probability distribution, i.e. $q_i = \lim_{n\to\infty} \Pr(x_n = s_i)$, and

$$\begin{cases} \mu_i^{(t)} = g_i^{(t)} w_i^{(t)} , & \sigma_i^{(t)} = \sqrt{g_i^{(t)}} w_i^{(t)} \\ \mu_i^{(p)} = g_i^{(p)} w_i^{(p)} , & \sigma_i^{(p)} = \sqrt{g_i^{(p)}} w_i^{(p)} \end{cases} , \quad (1)$$

the IPT and PS conditional means and standard deviations, respectively, then the average IPT and PS and standard deviations of the model are:

$$\begin{cases} \mu^{(t)} = \sum_{i=1}^N q_i \mu_i^{(t)}, \sigma^{(t)} = \sqrt{\sum_{i=1}^N q_i \mu_i^{(t)} (1+g_i^{(t)}) w_i^{(t)} - (\mu_i^{(t)})^2} \\ \mu^{(p)} = \sum_{i=1}^N q_i \mu_i^{(p)}, \sigma^{(p)} = \sqrt{\sum_{i=1}^N q_i \mu_i^{(p)} (1+g_i^{(p)}) w_i^{(p)} - (\mu_i^{(p)})^2} \end{cases}$$

IPT and PS pdf's are the following:

$$f_{\mathrm{IPT}}(d) = \sum_{i=1}^N q_i f_i^{(t)}(d) , \; f_{\mathrm{PS}}(b) = \sum_{i=1}^N q_i f_i^{(p)}(b) .$$

The average duration and the conditional (given that state) duration in the state s_i are, respectively:

$$\varphi_i = \frac{q_i}{1 - A_{i,i}} , \quad \phi_i = \frac{1}{1 - A_{i,i}} . \quad (2)$$

IPT and PS auto- and cross-correlations of the model are[3]:

$$\begin{aligned} R_{tt}(m) &= \mathcal{E}\{d_n d_{n+m}\} \\ &= \begin{cases} \mathbf{q}^T \mathbf{E}_{II}^{(t)} \mathbf{1} & m = 0 \\ \mathbf{q}^T \mathbf{E}^{(t)} \mathbf{A}^{|m|-1} \mathbf{E}^{(t)} \mathbf{1} & m \neq 0 \end{cases} , \\ R_{tp}(m) &= \mathcal{E}\{d_n b_{n+m}\} \\ &= \begin{cases} \mathbf{q}^T \mathbf{E}_{II}^{(tp)} \mathbf{1} & m = 0 \\ \mathbf{q}^T \mathbf{E}^{(t)} \mathbf{A}^{|m|-1} \mathbf{E}^{(p)} \mathbf{1} & m \neq 0 \end{cases} , \\ R_{pp}(m) &= \mathcal{E}\{b_n b_{n+m}\} \\ &= \begin{cases} \mathbf{q}^T \mathbf{E}_{II}^{(p)} \mathbf{1} & m = 0 \\ \mathbf{q}^T \mathbf{E}^{(p)} \mathbf{A}^{|m|-1} \mathbf{E}^{(p)} \mathbf{1} & m \neq 0 \end{cases} , \\ R_{pt}(m) &= \mathcal{E}\{b_n d_{n+m}\} \\ &= \begin{cases} \mathbf{q}^T \mathbf{E}_{II}^{(pt)} \mathbf{1} & m = 0 \\ \mathbf{q}^T \mathbf{E}^{(p)} \mathbf{A}^{|m|-1} \mathbf{E}^{(t)} \mathbf{1} & m \neq 0 \end{cases} . \end{aligned}$$

To show traffic dynamics without the biasing effects of average IPT and PS, in Section V covariances are taken into account instead of correlations.

B. Learning the Model Parameters

The Expectation-Maximization algorithm is an optimization procedure that allows learning of a new set of parameters for a stochastic model according to improvements of the likelihood of a given sequence of observable variables. For structures like HMM's this optimization technique reduces to the Baum-Welch algorithm [12] studied for discrete and continuous observable variables with a broad class of allowed conditional pdf's. More specifically, given a set of observable sequences $\mathbf{Y} = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)})$ referred to as the *training set*, we want to find the set of parameters such that the likelihood $\mathcal{L}(\mathbf{Y}; \Lambda) = \Pr(\mathbf{Y}|\Lambda)$ of the training set is maximum. The Baum-Welch algorithm is an iterative procedure looking for

[3] $E_{i,j}^{(t)} = A_{i,i} g_i^{(t)} w_i^{(t)} \delta_{i,j}$; $E_{i,j}^{(p)} = A_{i,i} g_i^{(p)} w_i^{(p)} \delta_{i,j}$; $E_{II\ i,j}^{(t)} = A_{i,i}(1+g_i^{(t)}) g_i^{(t)} (w_i^{(t)})^2 \delta_{i,j}$; $E_{II\ i,j}^{(p)} = A_{i,i}(1+g_i^{(p)}) g_i^{(p)} (w_i^{(p)})^2 \delta_{i,j}$; $E_{II\ i,j}^{(tp)} = A_{i,i} g_i^{(t)} w_i^{(t)} g_i^{(p)} w_i^{(p)} \delta_{i,j}$; $E_{II\ i,j}^{(pt)} = E_{II\ i,j}^{(tp)}$; $\mathbf{1}$ is a vector whose elements are 1, and $\delta_{i,j}$ is the delta of Kronecker.

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

TABLE I

TRAFFIC TRACES DETAILS.

Traffic	Link	Protocol	Port	Date	Size	Pkts	Sessions
AoM	LAN	UDP	2300	8/2003	12 MB	180K	6
SMTP	WAN	TCP	25	9/2005	3 GB	43M	56K
HTTP	WAN	TCP	80	7/2004	60 GB	830M	1M

TABLE II

AVERAGE STATISTICS.

		av. IPT (dBμ)	av. PS (bytes)	IPT s.d. (dBμ)	PS s.d. (bytes)
AoM	data	47.116	12.396	7.60	3.97
	starting	41.765	68.500	10.62	29.76
	trained	47.115	12.397	7.61	3.98
SMTP	data	40.091	710.5	18.55	619, 3
	starting	48.267	730, 5	19.65	347, 3
	trained	42.127	709, 6	18.14	640, 4
HTTP	data	51.43	542.3	16.16	324, 2
	starting	48.66	730.5	19.46	347.3
	trained	53.43	540.6	17.35	348.2

a local maximum of the likelihood function which typically depends on the starting point Λ. When necessary, multiple trainings with different initial conditions provide the global solution. The Baum-Welch for the proposed source-traffic model is based on the following equations:

$$\hat{A}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k-1} \alpha_n^{(k)}(i) A_{ij} f_j(\mathbf{y}_{n+1}^{(k)}) \beta_{n+1}^{(k)}(j)}{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k-1} \alpha_n^{(k)}(i) \beta_n^{(k)}(i)},$$

$$\hat{g}_i^{(t)} \hat{w}_i^{(t)} = \frac{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k} \alpha_n^{(k)}(i) \beta_n^{(k)}(i) d_n^{(k)}}{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k-1} \alpha_n^{(k)}(i) \beta_n^{(k)}(i)},$$

$$\hat{g}_i^{(p)} \hat{w}_i^{(p)} = \frac{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k} \alpha_n^{(k)}(i) \beta_n^{(k)}(i) b_n^{(k)}}{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k-1} \alpha_n^{(k)}(i) \beta_n^{(k)}(i)},$$

$$\hat{g}_i^{(t)} (\hat{w}_i^{(t)})^2 = \frac{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k} \alpha_n^{(k)}(i) \beta_n^{(k)}(i) (d_n^{(k)} - \mu_i^{(t)})^2}{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k-1} \alpha_n^{(k)}(i) \beta_n^{(k)}(i)},$$

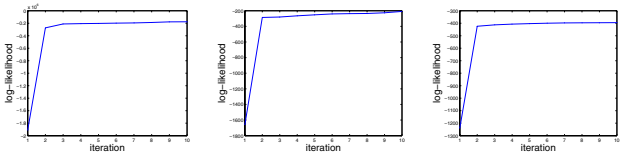
$$\hat{g}_i^{(p)} (\hat{w}_i^{(p)})^2 = \frac{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k} \alpha_n^{(k)}(i) \beta_n^{(k)}(i) (b_n^{(k)} - \mu_i^{(p)})^2}{\sum_{k=1}^{K} \frac{1}{P^{(k)}} \sum_{n=1}^{L_k-1} \alpha_n^{(k)}(i) \beta_n^{(k)}(i)},$$

where $f_i(\mathbf{y}_n^{(k)}) = f_i^{(t)}(d_n^{(k)}) f_i^{(p)}(b_n^{(k)})$, the likelihood is $P^{(k)} = \mathcal{L}(\mathbf{y}^{(k)}; \Lambda) = \sum_{i=1}^{N} \alpha_n^{(k)}(i) \beta_n^{(k)}(i)$, and the Forward and Backward variables are computed according to $\alpha_n^{(k)}(j) = \sum_{i=0}^{N-1} \alpha_{n-1}^{(k)}(i) A_{ij} f_j(\mathbf{y}_n^{(k)})$, $\beta_n^{(k)}(i) = \sum_{j=0}^{N-1} A_{ij} f_j(\mathbf{y}_{n+1}^{(k)}) \beta_{n+1}^{(k)}(j)$.

IV. MEASUREMENT APPROACH AND TRAFFIC TRACES

As regards games traffic, we studied traffic generated by Age of Mythology, a Microsoft Real Time Strategy Multiplayer Game. The traffic traces have been provided, in Tcpdump format, by the Worcester Polytechnic Institute (WPI), MA (USA) [13] and they consist of packet sequences of complete gaming sessions, between two players, captured in a LAN environment. For SMTP and HTTP instead, we captured traffic by passively monitoring the WAN access link at *University of Napoli "Federico II"* network during the period *January 2004 - December 2005*. For this purpose we developed Plab, a software platform to capture and analyze network traffic both online and offline, which was also used to process AoM traces. Both Plab and our traffic traces used in this paper are freely and publicly available at [14].

In Table I details about the traffic traces that we analyzed are given. With the term "session", in the case of AoM, we mean all traffic exchanged from the beginning to the end



(a) AoM (b) SMTP (c) HTTP

Fig. 1. Log-likelihood vs. iteration.

of a match, whereas in the case of SMTP/HTTP we mean all the traffic exchanged between two hosts related to ports TCP 80 (for HTTP) and 25 (for SMTP) with a timeout of 15 minutes. Only six gaming sessions were studied, because packet-level traffic of RTS games has been demonstrated being very predictable and strongly dependent from the specific game application whereas it is poorly dependent from user behavior [15]. As regards SMTP and HTTP traces, instead, we observed a much larger set of sessions. This is because of the more complex nature of such traffic (see [16] for a traffic characterization) and also because we could gather our own traces. To preserve privacy, for each packet we kept only the IP, UDP, and TCP headers and we scrambled IP addresses using the wide-tcpdpriv tool from the MAWI-WIDE project [17]. As regards the time resolution of the measurements, the packet timestamp resolution provided by the Libpcap library (which is used both by Tcpdump and Plab), and by the kernel drivers that it links to, is of $1\mu s$. An important aspect of our methodology is that in the evaluation of IPT and PS distributions we did not take into account packets with empty payload. Since we wanted to characterize the traffic generated by the applications, independent as much as possible of TCP itself, we decided to drop all TCP-specific traffic, like connection establishment packets (SYN-ACK-SYNACK) and pure acknowledgment packets [18]. For the same reason, in the estimation of the packet size, we measured the byte length of the TCP payload or, in the case of AoM, we considered the UDP payload. These choices make our results usable for simulation purposes as an input for TCP state machines and UDP/IP stacks, like in D-ITG [14] and TCPlib [3].

We decided to apply the proposed model to the above-mentioned traffic types for several reasons. First, the selected traffic types (based on both UDP and TCP) are quite different from each other, due to users and applications behavior. Second, games traffic represents a new and interesting traffic class. Also, network game traffic generates a significant share of todays Internet traffic. It is reported that $3 - 4\%$ of all packets in a backbone could be associated with only 6 popular games [19]. Third, HTTP and SMTP represent two applications largely used over Internet (the most used by common users).

V. EXPERIMENTAL RESULTS

This Section presents preliminary results of our model when it is applied to AoM, SMTP, and HTTP traffic. As regards SMTP and HTTP, we will refer to the definition of session given in Section IV. For each session between two hosts,

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

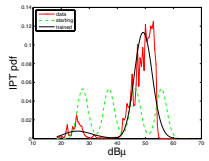
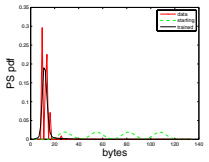


Fig. 2. AoM: IPT&PS pdf.



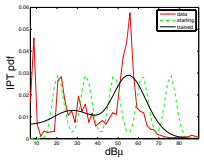
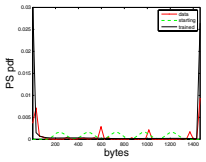


Fig. 3. SMTP: IPT&PS pdf.



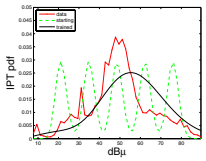
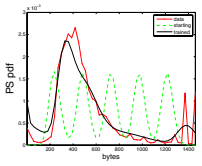


Fig. 4. HTTP: IPT&PS pdf.



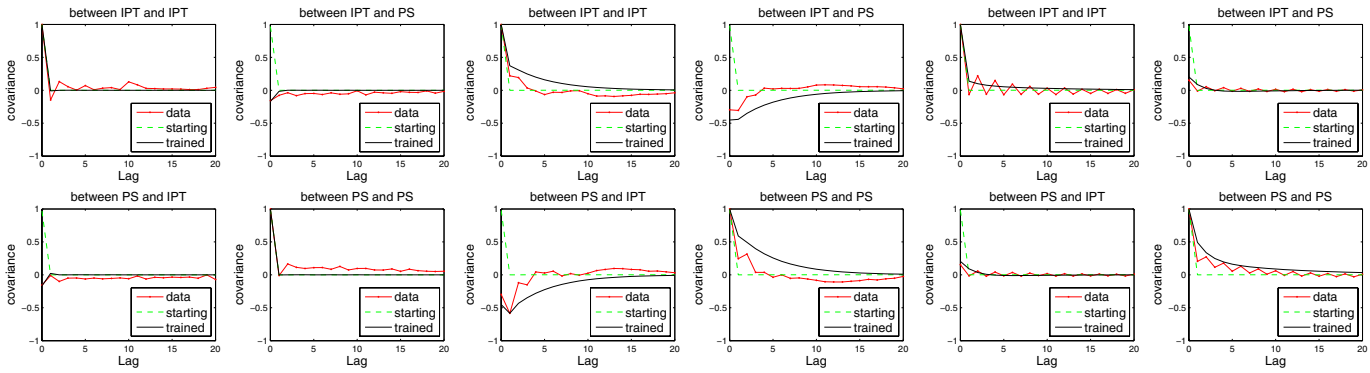


Fig. 5. AoM: IPT&PS auto- and cross-covariance. Fig. 6. SMTP: IPT&PS auto- and cross-covariance. Fig. 7. HTTP: IPT&PS auto- and cross-covariance.

two different flows of data can be identified, which we called *upstream* and *downstream*. Upstream packets have TCP port 25 and 80 as destination port, whereas downstream packets have source TCP port 25 and 80 (in the case of SMTP and HTTP respectively).

In this paper we concentrate on the traffic sources represented by HTTP and SMTP clients, we therefore model only upstream traffic. We adopt the same approach for AoM, modeling the traffic flowing in the outbound direction when seen from the point of view of a specific peer (i.e. leaving the workstation of a gaming user). Anyway, being the observed traces related to matches with two players, the traffic flowing in the other direction is almost symmetrical. As regards downstream traffic, it is worth mentioning that in the case of SMTP the vast majority of downstream flows - for each session - are made of only few packets (about 5) of small size. Thus they represent a very small portion of SMTP traffic. This can explained by SMTP protocol specifications: the peer acting as a server usually answers to requests and data transfers from the client with small messages that must have a numeric ID prepended. As for HTTP instead, strong volumes of traffic are generated in both directions, this is due to the intrinsic nature of the Web traffic.

We used the model with $N = 4$ states for AoM traffic and $N = 5$ states for SMTP and HTTP traffic, due to their more complex structure. All three cases reached convergence in terms of likelihood after a few iterations, as shown in Fig. 1. In the following we consider results obtained with 10 iterations. They exhibit good precision in matching mean and standard deviation for the marginal distributions, as shown in Table II.

A. *AoM traffic*

Fig. 2 shows the pdf for the training set and for the starting and training models. Packets' payloads are usually smaller than 20 bytes and are concentrated around few close values. On the contrary, it is interesting to note that the IPT distribution shows a bi-modal behavior, with the main modes separated by more than two orders of magnitude. We found similar behaviors in other real-time strategy games, where stations typically send periodic update packets plus additional update packets when a user action must be immediately transmitted. $N = 4$ states showed to be sufficient to capture the behavior of the real data. Fig. 5 shows the results obtained for auto- and cross-covariance. We found that all four covariances rapidly decay, an aspect that is well captured by the trained model. Also note (see cross-covariance at Lag 0) the presence of a small dependence between IPT and PS of the single packet, well captured by the trained model. Finally, in Fig. 8, which shows the training set and a synthetically-generated set of IPT-PS pairs, it can be seen that the model is able to accurately reproduce the AoM traffic pattern.

B. *SMTP Traffic*

We present results from the sessions with less than 100 packets, which we defined as *short-lived*, and which account for $\sim 97\%$ of the SMTP sessions. This is because we found that there are other sessions which exhibit extremely different statistical properties. This is confirmed by a K-means clustering with a few features per session, e.g. number of packets, bytes, IPT and PS mean and variance. Note that considering only this class does not affect our approach, as we do not want to provide a comprehensive model for SMTP traffic. At this stage we want to show the applicability of the proposed approach also to this kind of traffic (quite different in terms of users and protocol behavior from the AoM traffic). Note that SMTP traffic exhibits a very complex structure both in terms of marginal distributions and correlations. Fig. 3 shows how the trained model is able to follow the main envelope

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

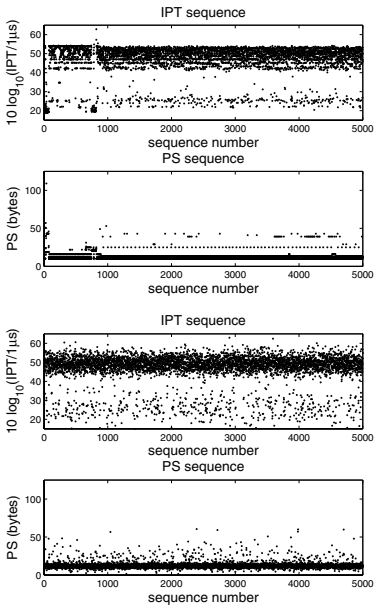


Fig. 8. AoM: IPT&PS training (up) and synthetic (down) traces.

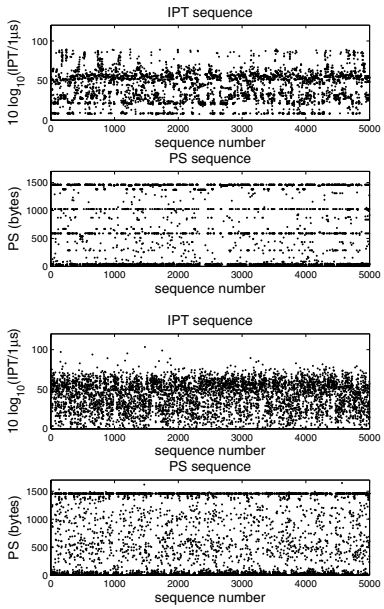


Fig. 9. SMTP: IPT&PS training (up) and synthetic (down) traces.

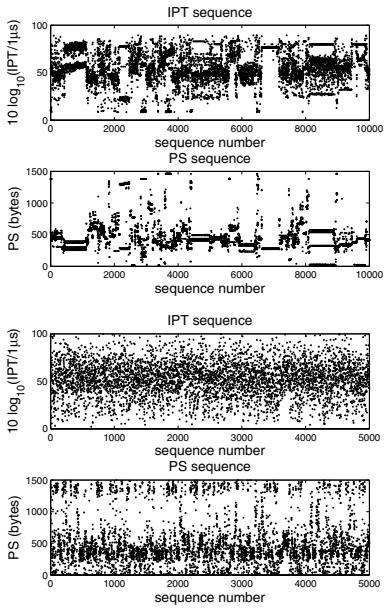


Fig. 10. HTTP: IPT&PS training (up) and synthetic (down) traces.

of the marginal distributions, capturing at the same time the temporal dynamic of the traffic, as shown in Fig. 6. This can also be seen looking at the real and synthetic (generated from the trained model) traffic traces, shown in Fig. 9.

C. *HTTP Traffic*

Fig. 4 shows the fitting of marginal distributions. Note that IPT are spread over 8 orders of magnitude, but the majority of them is concentrated approximately between $10ms$ and $1s$. These values are compatible with RTTs found in Wide Area Networks. Indeed HTTP clients often perform a lot of subsequent requests to the same server. In the case of Web, for example, the first request of an HTML document is typically followed by more requests for the embedded objects. If such objects are small enough to be sent within one or few packets (as often is the case [20]), the requests are sent with intervals close to the RTT from the client to the server. The correlation structures of HTTP, shown in Fig. 7, are very interesting. They present correlation at several lags with an oscillating behavior. The envelope decays faster for cross- than auto-covariances, and it is accurately captured by the trained model. It is worth saying that the model trained with single sessions (here not reported due to lack of space) captured the oscillating behavior too, while for the whole traffic, where different kinds of sessions are considered, this is quite hard and was not possible. Also for HTTP, in Fig. 10 we show the capability of the model to jointly reproduce time series of both IPT and PS by synthetic generation of traffic patterns. Finally, as for IPT, the HMM model captures the characteristics of real data but seems to slightly under-estimate small values in favor of larger ones.

D. *Discussion*

Indeed they differ for the behavior of the marginal distributions of IPT and PS but also for their correlation structure. As for the last point, it is worth noting that we found larger autocorrelations with a slower decay for SMTP and HTTP traffic when compared to AoM. Such behavior can be partially explained by the influence of TCP end-to-end flow control, which introduces dependencies between IPT. Indeed, while SMTP and HTTP run over TCP, AoM traffic is carried by UDP packets. Furthermore, rigid application-level protocol rules of SMTP and HTTP induce more structure into their traffic patterns. On the other side, as regards AoM, the interaction of the gaming user introduce more randomness into the traffic.

For all the three classes of traffic the HMMs have shown the capability to jointly model IPT and PS. The proposed models succeeded in effectively capturing first-order statistics and temporal dynamics of real network traffic. Training models for AoM, SMTP, and HTTP required few iterations, and though SMTP and HTTP traffic present a much more complex structure, they only required one more state (with respect to AoM) for effective modeling. Then, the flexibility of an HMM approach, even when applied to a low-level traffic modeling, appears quite encouraging. Concluding, it is worth highlighting that the more exciting result of the proposed model is, in our opinion, the capability to fit at the same time both IPT and PS statistics and dynamics, even if not obtaining extreme accuracy, of three heterogeneous traffic sources with a relative small set of parameters.

VI. Prediction

The trained models have been used for prediction purposes of a sample trace. The objective is to show the capability of the model to furnish the expected short-term future behavior of

1-4244-0357-X/06/$20.00 ©2006 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.

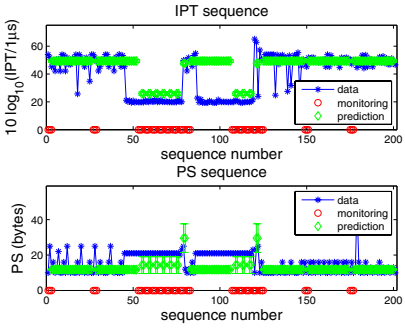


Fig. 11. AoM: monitoring and prediction.

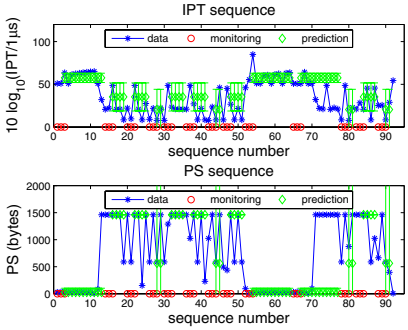


Fig. 12. SMTP: monitoring and prediction.

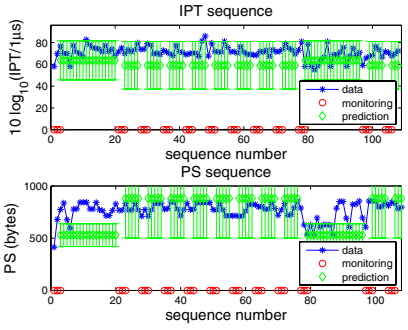


Fig. 13. HTTP: monitoring and prediction.

the traffic with sufficient accuracy. Such a characteristic results particularly appealing when thought as part of a more complex network-sensing and adaptive-management system. In order to give an idea of the potentiality of the proposed approach, we performed (off-line) the following basic steps on the traces described previously: *Monitoring* - W samples (in terms of IPT-PS pairs) are observed iteratively to obtain an estimate of the current state via the Viterbi algorithm [12]; *Prediction* - on the basis of the current state estimate and of the trained model parameters, the traffic is assumed to behave in a known fashion. Figs. 11-13 show results in the case of AoM, SMTP, and HTTP traffic. The blue lines with asterisks represent the real data. We considered a W=3-sample observation to obtain current-state estimate. Also, we assume that the traffic holds on conditional mean values (Eq. (1)) for a number of samples proportional to the average duration (Eq. (2)) of the state. Blue asterisks, red circles, and green diamonds represent real data, monitored samples, and predicted samples, respectively[4]. Comparing the frequent superposition between blue asterisks and green diamonds, it can be noticed how the model captures and predicts the traffic dynamics for all the three different considered traffic. Such a result is quite surprising if we look at the source behavior when being in the states with large average duration, i.e. states whose conditional mean values are 49 dBμ and 12 bytes in case of AoM, 57 dBμ and 25 bytes in case of SMTP, and 64 dBμ and 530 bytes in case of HTTP. Note again that, due to the joint modeling we proposed, estimation of the state variable allows to infer knowledge about both IPT and PS expected behavior simultaneously.

VII. Conclusion

In this paper we proposed a model of traffic sources at packet level. It has been shown how the proposed HMM (based on a joint representation of IPT and PS) approach is able to capture both first-order statistics and temporal structures of the traffic generated by a number of heterogeneous sources. The capability to accurately replicate and predict traffic makes the proposed approach quite promising. We are currently validating the proposed approach with many other traffic sources (e.g. Instant Messaging, FTP, POP3, IMAP, SSH, Telnet, p2p, Worms ...).

[4]For better precision we also reported a confidence interval proportional to the conditional standard deviation (Eq. (1)).

References

[1] C. Wright, F. Monrose, G. Masson, "HMM Profiles for Network Traffic Classification", *VizSEC/DMSEC*, pp. 9-15, Oct. 2004.

[2] P. Salvador, A. Pacheco, R. Valadas, "Modeling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs", *Computer Networks*, Vol. 44, pp. 335-352, Oct. 2004.

[3] P. Danzig, S. Jamin, R. Caceres, D. Mitzel, D. Estrin, "An Empirical Workload Model for Driving Wide-Area TCP/IP Network Simulations", *Internetworking: Research and Experience*, Vol. 3(1), pp. 1-26, Mar. 1992.

[4] G. Iannello, F. Palmieri, A. Pescapè, P. Salvo Rossi, "End-to-End Packet-Channel Bayesian Model applied to Heterogeneous Wireless Networks", *IEEE GLOBECOM*, pp. 484-489, Nov. 2005.

[5] `http://www.microsoft.com/games/ageofmythology/`

[6] L. Muscariello, M. Mellia, M. Meo, M.A. Marsan, R. Lo Cigno, "Markov models of internet traffic and a new hierarchical MMPP model", *Computer Communications Journal*, Vol. 28(16), pp. 1835-1851, Oct. 2005.

[7] A. Klemm, C. Lindemann, M. Lohmann, "Modeling IP traffic using the batch Markovian arrival process", *Performance Evaluation* Vol. 54(2), pp. 149-173, Oct. 2003.

[8] Y. Hafri, C. Djeraba, P. Stanchev, B. Bachimont, "A Markovian Approach for Web User Profiling and Clustering", *PAKDD*, pp. 191-202, Apr. 2003.

[9] E. Costamagna, L. Favalli, F. Tarantola, "Modeling and Analysis of Aggregate and Single Stream Internet Traffic", *IEEE GLOBECOM*, pp. 3830-3834, Dec. 2003.

[10] T. Lang, G.J. Armitage, P. Branch, H. Choo. "A Synthetic Traffic Model for Half-Life", *ATNAC*, Dec. 2003

[11] W. Feng, F. Chang, W. Feng, J. Walpole, "A Traffic Characterization of Popular On-line Games", *IEEE/ACM Trans. on Networking*, Vol. 13(3), pp. 488-500, Jun. 2005.

[12] L.R. Rabiner, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Procs. of the IEEE*, Vol. 77(2), pp. 257-285, Feb. 1989.

[13] `http://nile.wpi.edu/downloads`

[14] `http://www.grid.unina.it/Traffic/`

[15] M. Claypool, "The Effect of Latency on User Performance in Real-Time Strategy Games", *Computer Networks*, Vol. 49(1), pp. 52-70, Sept. 2005

[16] A. Dainotti, A. Pescapé, G. Ventre, "A Packet-level Characterization of Network Traffic", CAMAD, pp. 38-45, June 2006.

[17] `http://www.wide.ad.jp/wg/mawi/`

[18] R. Caceres, P. Danzig, S. Jamin, D. Mitzel, "Characteristics of Wide-Area TCP/IP Conversations", *ACM SIGCOMM Computer Communication Review*, Vol. 21(4), pp. 101-112, Sept. 1991.

[19] S. McCreary, K. Claffy, "Trends in wide area IP traffic patterns - A view from Ames Internet Exchange", *ITC Specialist Seminar of Measurement and Modeling of IP Traffic*, pp. 1-11, Sept. 2000.

[20] F.D. Smith, F.H. Campos, K. Jeffay, D. Ott, "What TCP/IP Procotol Headers Can Tell Us About the Web". *ACM SIGMETRICS*, pp. 245-256, Jun. 2001.

1-4244-0357-X/06/$20.00 ©2006 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2006 proceedings.