Studies for Modeling Basic Aspects of Source Traffic

João V. P. Gomes[1,2,a], Pedro R. M. Inácio[1,2,b], Branka Lakic[3],
Mário M. Freire[2], Henrique Silva[4], Paulo P. Monteiro[1,3]
[1]Nokia Siemens Networks Portugal S.A., Rua Irmãos Siemens, nº 1, 2720-093 Amadora, Portugal
[2]IT-Networks and Multimedia Group, Department of Computer Science, University of Beira Interior,
Rua Marquês de Ávila e Bolama, 6201-001 Covilhã, Portugal
[3]Institute of Telecommunications, University of Aveiro, 3810-193 Aveiro, Portugal
[4]Department of Electrical and Computers Engineering, University of Coimbra, Pólo II, 3030-290 Coimbra, Portugal
e-mail (corresponding authors): [a]jgomes@penhas.di.ubi.pt, [b]pedro.inacio@nsn.com

***Abstract* — This paper summarizes a study conducted over traffic generated and received at a network terminal machine. The results it includes may be used to simulate the processes of the interarrival times, of the packet sizes and of the bit count per time unit, at the source level and for several types of telematic applications (namely Voice over IP and File Sharing). Analysis shows that the Weibull and Normal distributions may be fitted to the greater part of the empirical data and that most of the traces exhibit positive correlation.**

I. INTRODUCTION

Traffic modeling and simulation plays an important role in the specific area of Traffic Monitoring and Analysis (TMA) and in the more general area of Telecommunications, for it provides practitioners and system testers with efficient tools to evaluate the performance of networks and of their elements. Due to the increasingly complex behavior of telematic applications and automatic switching devices, perfect synthesis of network traffic is rather difficult, and approximated models constitute often an attractive and viable choice.

This paper summarizes a study conducted over traffic generated and received at a network terminal machine. Its main purpose is to provide practitioners with an idea of the main statistical properties of the three following basic aspects of such traffic: bit count per time unit, interarrival times and packet sizes. The values included below concretize several theoretical models that may be used to approximate the aforementioned traffic aspects, since their application in computer based simulations was on the basis of this study.

This paper is structured as follows. Section II contains a brief mention to traffic modeling and analysis in the past. Section III discusses the analysis conducted for the traffic traces collected at a personal computer, and summarizes the most important results of the study. The main conclusions may be found in Section IV.

The authors would like to acknowledge financial support from Fundação para a Ciência e Tecnologia (grant contracts SFRH/BDE/15592/2006 and SFRH/BDE/15643/2006), from Nokia Siemens Networks Portugal S.A. and from project PTDC/EIA/73072/2006 TRAMANET: Traffic and Trust Management in Peer-to-Peer Networks. Branka Lakic is currently with the European Patent Office, in Netherlands. This manuscript summarizes part of the work discussed in a lengthier magazine paper, submitted to the ACM Transactions on Multimedia Computing, Communications and Applications.

II. TRAFFIC MODELING IN THE PAST

A. Network Aggregation Points

The majority of the studies about traffic from computer networks are conducted with the purpose of developing models to simulate the behavior of the data flows in aggregation points. Most of them explore the self-similarity property of the traffic claiming that, in that point of the network, the byte count process is self-similar.

In [7], the authors analyzed traffic from Local Area Networks (LANs) and demonstrated that a superposition of many ON/OFF sources exhibiting the *Noah effect* results in self-similar traffic. In what is comes to traffic from Wide Area Networks (WAN), it is claimed by some studies that it exhibits the properties of asymptotically second order self-similar processes. In [6], the authors studied the File Transfer Protocol (FTP), World Wide Web (WWW) and Telnet, coming to the conclusion that transmission of files over such protocols results also in self-similarity at the WAN level.

B. Voice over Internet Protocol, Video and Data

Voice over Internet Protocol (VoIP) traffic is increasing within computer networks, motivating the development of models for this traffic class that can help to better shape the Quality of Service (QoS) policies. While the generation of Internet Protocol (IP) packets depends of the codec used by the VoIP protocol and cannot, most of the times, be modeled by one generic distribution, the duration of calls and the call interarrival times is commonly described by the exponential distribution [4]. Moreover, although spurts and gaps were initially modeled using the exponential distribution, recent studies show that the lognormal distribution is the best fit [3].

In the case of video traffic, the data transmitted is formed by video, voice and system data. However, the majority of the existent models are based only in the video part of the data. The importance of autocorrelation for video modeling is also claimed by several authors, who developed different approaches to embed the autocorrelation in the proposed models. Ansati et al. [1] extracted the parameters needed to model video traffic using Fractional Auto Regressive

Integrated Moving Average (FARIMA) processes, from different videos sequences.

Bardord and Crovella [2] modeled WWW traffic using ON/OFF processes. The ON periods are the ones where a given source is transmitting, whereas the OFF periods are related with the time between two downloads. The authors used mostly Pareto to describe the ON and the OFF times.

In [5], different distributions were used to model distinct properties of FTP traffic: exponential for the session interarrival times; lognormal for the connection size; both for burst size; and exponential, uniform and lognormal for connection interarrival times.

III. Fitting Distributions and Measuring Autocorrelation

As previously stated, most of the efforts in the area of traffic modeling are placed upon network aggregation points. The reason behind this fact lies in the analysis of the queuing effects in data forwarding nodes, which has received a lot of attention from the telecommunications community in the last few years. From the perspective of the aggregation point, the characteristics of the particular flows are often irrelevant, but from the perspective of QoS mechanisms acting near the edges, such properties become rather important.

The conclusions reported herein do not address all the vicissitudes of the traffic generated at the source level. Actually, they constitute a tentative to provide a simplified and useful view of the behavior of the traffic, which may be adopted by researchers aiming to conduct network traffic simulations with some control over the characteristics of the individual data flows.

A. Data Sets and Traffic Capturing Scenarios

The main driver for the study of source traffic was on the analysis of its effects in small LANs, directly connected to an access network. The traces used in the scope of this work were intentionally created with personal computers, and collected by tapping the connection between the (Ethernet) LAN and the Internet. Some of the traces were collected in a final branch of a large local network over Ethernet. The several types of traffic were named accordingly to the applications used to generate them. These applications are amongst some of the most popular ones at the time this manuscript was written, namely: *Skype*, *eMule* and *MSN*. The Web related traffic was divided into several subclasses, including *Browsing* (termed simply *Web*), *file download from a server* and *Streaming*, since the operational mode of the two services is slightly different, thought may be supported by the same protocol. Because of the same reason, file sharing was also investigated for the *file upload* and *download* operations. The collection is complemented with five recordings of the typical use of the Internet by a residential or corporate user.

Each one of the collected traces was divided into three data sets, hereinafter termed as IN, OUT and MIX. IN refers to the INcomming communications (from the terminal machine perspective), while OUT refers to the OUTgoing portion of the traffic. MIX contains both types.

B. Theoretical Models

The analysis apparatus was set to construct the cumulative distribution functions of the empirical processes of the *interarrival times*, *packet sizes* and *byte count per second*. Several theoretical models were then fit to the data using well documented estimators (when available, maximum likelihood estimators were used). From all the theoretical distributions, we would like to emphasize the following ones: *Pareto distribution*, with the location and shape parameters, denoted afterwards by $P(x_{min},s)$; *Gaussian distribution*, with average and variance parameters, herein termed by $N(a,b)$; and the *Weibull distribution*, with scale and shape parameters, symbolized by $W(a,b)$. Note that the Lognormal, the Exponential, the Rayleigh and the Gamma distributions were also taken into consideration during this work, thought they did not fit satisfactorily any of the data sets (see table I below).

The criterion for selecting the best fit amongst all models was based on the *Kolmogorov-Smirnov goodness of fit test*. i.e. the model presenting the smallest *discrepancy* value D^k was momentarily labeled as the *most suitable one*, where $D^k=max_{x\epsilon\Omega}|F^k(x_i)-F_e(x_i)|$, $F^k(x_i)$ and $F_e(x_i)$ denote the theoretical and the empirical cumulative distribution functions, respectively, and Ω is the set of all empirical occurrences. The final decision about the quality of the fit was made recurring to human discernment, so as to exclude the cases where the model with the smallest discrepancy was noticeably too far from the empirical data. When found pertinent, we chose to include the most decisive probability peaks of a given distribution, instead of the indication of a model and of the estimations of their parameters. To make the best use of the available space of this document, it was decided to sum up the most significant results in the form of a table, included in a subsection below (*D. Summary*). For the sake of coherence, additional comments regarding this subject are postponed to that section.

C. Autocorrelation

Because the autocorrelation structure of the bit count per time unit plays an important role in the study of queuing effects (long-range dependent sequences are positively correlated), it was decided to explore that statistical aspect also. The calculation of this metric was made recurring to the standard autocorrelation formula to all lags smaller than 40 seconds, since some of the studied traces were no longer than 1 minute (e.g. some VoIP calls). To obtain an idea of the amount and type of correlation that needs to be simulated when forging individual flows of traffic, the minimum and the maximum of the aforementioned 40 values were plotted against the designation of the applications that generated the respective traces (sorted in an increasing manner by the minimum autocorrelation value). These charts are included in Fig. 1.

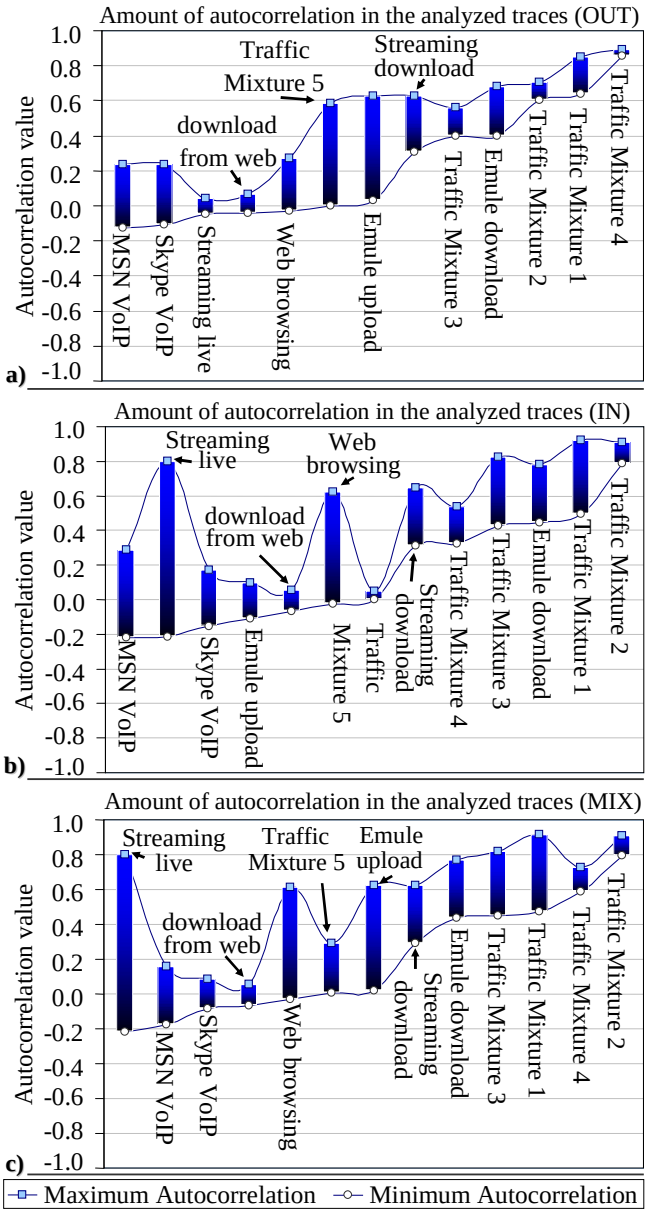

Fig. 1. Variation interval of the first 40 values (i.e. 40s) of the autocorrelation function of the byte count per time unit process, calculated for the a) OUT, b) IN and c) MIX data sets.

As may be concluded from careful observation, none of the traces exhibits anti-persistence. Several types of traffic present different correlation properties depending on the direction of the transmission, being that particularly evident for highly asymmetric applications (e.g. eMule, Browsing). It is interesting to notice that the incoming communications are more auto-correlated then the outgoing ones, mostly due to the asymmetry of the traffic and size of files / packets being exchanged in both directions. Also worth of mentioning is the fact the communications concerning Traffic Mixture 5 seem to be more random that the others mixtures. The reason for such to happen lies in the random behavior of the user of the computer, which was merely using the computer to check mail, surf the web and chat with friends. No large files were transmitted during such communications. The variation interval of the autocorrelation of VoIP is also small and near to zero, motivated by the random influence of human talk.

D. Summary

Table I summarizes the results obtained for each one of the data sets. It contains the values of the parameters for the model that best fits the several aspects under study when such is applicable. In the case where no model could be adapted to the cumulative distribution curves because of two or three especially probable values (or contiguous interval of values), those are explicitly written in the respective cells along with their frequency.

As may be concluded from observation of the table, Weibull was able to model most of the experimental data for several aspects of the traffic, shaping perfectly all but two cases of the interarrivals process (one of those two could not be modeled using any distribution). In the case of the byte count process, Weibull and Normal were the distributions that best modeled the empirical traces. The packet size values were represented, in the great majority of the cases, by a bi- or tri-modal distribution. In spite of the popularity the Pareto distribution has gained in the past, our analysis found it only useful to model the packet sizes of the Skype VoIP traffic and the interarrivals of streaming broadcast.

IV. CONCLUSIONS

This paper presents a study of the statistical properties of traffic collected near terminal nodes of LANs. After the inclusion of a short overview to the subject of traffic analysis, which contextualizes and differentiates the study by the approach we have taken, the experimental apparatus was briefly explained and its results provided in a condensed form. Several well known distributions were adapted to the empirical processes of the byte count per second, interarrival and packet sizes. It was concluded that Weibull distribution proved itself to be suitable for modeling the majority of the processes, especially for the interarrival process. The Gaussian distribution was able to fit the byte count per second of a big part of the traces, though Weibull was also close to the experimental values. It was also concluded that the distribution of the packet sizes is dominated by probability peaks in many cases, being thus better to model such aspect recurring to bi-/tri-modal or empirical distributions. The results concerning the autocorrelation of the byte count per second corroborate the theory of self-similarity at network aggregation points, since persistency is already present in some flows at the source level.

Future research paths concerning the results in this manuscript include their usage in traffic simulation or prediction works. Planning of next generation networks is one of the possible applications of those works.

Table I
Results of the analysis of several traffic properties for the various data sets.

Traffic		Byte Count	Packet Size	Interarrivals
Web	OUT	W(0.24, 27.23)	W(0.37, 50.95)	W(0.38, 0.05)
	IN	W(0.17, 11.36)	60(26%), ≥ 1484(52%)	W(0.38, 0.03)
	MIX	W(0.19, 40.95)	54-62(48%), ≥ 1484(30%)	W(0.25, 5.3E-3)
Skype VoIP	OUT	W(10.16, 5099.29)	N(144.68, 626.10)	W(0.59, 0.06)
	IN	W(11.82, 5063.86)	N(143.94, 554.09)	W(2.99, 0.03)
	MIX	W(21.38, 10047.74)	N(145.18, 802.07)	W(1.29, 0.02)
Streaming download	OUT	N(2025.45, 399485.72)	54(99%)	W(0.34, 6.2E-3)
	IN	N(104784.07, 1006.94E6)	1514(95%)	W(0.22, 3.6E-4)
	MIX	N(106554.18, 1115.16E6)	54(34%), 1514(62%)	W(0.17, 4.9E-5)
Streaming broadcast	OUT	W(1.04, 108.39)	45-84(93%)	no fit
	IN	N(5084.77, 19485069.77)	517-526(88%)	P(8E-6, 0.27)
	MIX	N(5257.23, 19611722.80)	45-84(23%), 517-526(71%)	W(0.66, 0.06)
eMule (file upload)	OUT	W(3.04, 6422.23)	W(0.48, 703.99)	W(0.46, 0.09)
	IN	W(2.48, 577.73)	W(1.27, 47.45)	W(0.85, 0.13)
	MIX	W(3.32, 7002.00)	W(0.41, 208.44)	W(0.43, 0.03)
eMule (file download)	OUT	W(1.18, 1968.07)	W(0.68, 43.11)	W(0.68, 0.04)
	IN	N(14033.01, 97066134.82)	60-72 (40%), ≥ 1494 (23%)	W(0.89, 0.04)
	MIX	N(15958.23, 110.16E6)	W(0.26, 92.68)	W(0.51, 1.2E-2)
File download from web	OUT	N(1550.54, 160774.45)	54-74(99%)	W(0.35, 1.5E-2)
	IN	N(51789.56, 190653714.80)	1514(99%)	W(0.33, 8.5E-3)
	MIX	N(53370.46, 199739532.26)	54-74(44%), 1514(56%)	W(0.19, 4.1E-4)
MSN VoIP	OUT	N(5456.51, 2940917.82)	W(8.77, 128.26)	W(1.67, 0.03)
	IN	N(2693.21, 6356671.55)	W(4.99, 133.54)	W(1.19, 0.03)
	MIX	N(8238.40, 4854388.43)	W(5.04, 141.32)	W(1.22, 0.02)
Mail, MSN and file sharing traffic	OUT	W(0.38, 2546.25)	54-77(77%), ≥ 1402(14%)	W(0.47, 0.03)
	IN	W(0.31, 5230.75)	60-62(10%), 1099(15%), ≥ 1484(61%)	W(0.50, 0.02)
	MIX	W(0.40, 10338.75)	54-77(40%), 1099(9%), ≥ 1402(41%)	W(0.32, 4.1E-3)
File Sharing, download from web and MSN traffic	OUT	N(7050.82, 30199712.90)	54-74(88%), ≥ 1414(10%)	W(0.63, 0.02)
	IN	W(0.87, 44999.36)	60-66(6%), 1514(69%)	W(0.71, 0.02)
	MIX	W(0.93, 57074.99)	54-74(41%), ≥ 1414(45%)	W(0.51, 6.5E-3)
File Sharing, streaming download and MSN traffic	OUT	N(7728.56, 30210101.80)	54-77(73%), ≥ 1506(17%)	W(0.57, 0.03)
	IN	W(1.19, 19625.19)	60-66(19%), ≥ 1506(43%)	W(0.73, 0.03)
	MIX	W(1.26, 29481.66)	54-77(47%), ≥ 1506(30%)	W(0.50, 0.01)
Skype, streaming download and file sharing traffic	OUT	N(8873.73, 36352160.02)	P(42, 0.89)	W(0.77, 0.03)
	IN	N(5546.58, 40574100.44)	P(60, 1.13)	W(1.00, 0.04)
	MIX	N(14419.21, 88421057.85)	P(42, 0.88)	W(0.68, 0.01)
Web, mail and instant messaging traffic	OUT	W(0.27, 30.01)	54-62(61%), ≥ 1482(12%)	W(0.24, 0.02)
	IN	W(0.17, 2.28)	60-93(28%), ≥ 1506(49%)	W(0.22, 0.01)
	MIX	W(0.33, 285.35)	54-93(62%), ≥ 1482(18%)	W(0.40, 0.04)

REFERENCES

[1] Nirwan Ansari, Hai Liu, Yun Q. Shi, and Hong Zhao. On Modeling MPEG Video Traffics. *IEEE Transactions on Broadcasting*, 48(4):337–347, December 2002.

[2] Paul Barford and Mark Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. *ACM SIGMETRICS Performance Evaluation Review*, 26(1):151–160, June 1998.

[3] E. Casilari, H. Montes, and F. Sandoval. Modelling of Voice Traffic Over IP Networks. In *Proceedings of The Third International Symposium On Communication Systems, Networks and Digital Signal Processing*, pages 411–414, Stafford, UK, June 2002.

[4] Roger L. Freeman. *Telecommunication System Engineering*. Wiley-IEEE Press, 4th edition, June 2004.

[5] J. Ishac. FTP Traffic Generator. Technical report, Case Western Reserve University, January 2001.

[6] W. Willinger, Vern Paxson, and M. S. Taqqu. Self-similarity and Heavy Tails: Structural Modeling of Network Traffic. In Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, editors, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, chapter Applications, pages 27–53. Birkhäuser, Boston, September 1998.

[7] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson. Self-similarity Through High-variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, February 1997.