This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.

Reducing Network Traffic Data Sets

Alessio Botta, Alberto Dainotti, Antonio Pescapè, and Giorgio Ventre
University of Napoli "Federico II", E-mail: {a.botta,alberto,pescape,giorgio}@unina.it

***Abstract*— In the study of network traffic, the collection and the processing of measurement data sets play a fundamental role. Due to the large size of typical traffic traces, their analysis is often heavy in terms of computational time and resources. In addition, even when the data sets are small, due to the intrinsic redundancy of the data, there is no need to consider the entire data sets in the processing stages. To cope with these issues, we use an *Entropy*-based methodology to reduce network traffic data sets obtained by measurements over real networks. The off-line approach we used is based on the *Marginal Utility* concept, and reveals encouraging results when applied to real data captured over real networks, especially when dealing with large amounts of data. To show the applicability of our approach, we present and discuss results obtained in the analysis and characterization, at packet-level, of traffic traces from two popular network games: *Counter-Strike* and *Age of Mythology*. Thanks to the differences between the two considered on-line games and their traffic traces we can draw pros and cons in realistic scenarios.**

I. Introduction

Traffic analysis is a wide and fertile research area. Understanding the properties of traffic generated by new applications and flowing in current networks allows to improve their performance and to design new architectures efficiently. Performance evaluation of networking systems as well as traffic characterization, modeling, simulation and emulation are all activities that need to rely on realistic data. However, the collection of data traces from real networks often requires managing large quantities of data. In one hour, the collection of 60 byte packet headers on an OC-48 link can generate 600 GB of data [4]. As the amount of data increases, the time required for their analysis raises, and when working with large data sets several problems can occur (e.g. the software environment for data processing and statistical analysis runs *out of memory*). Another typical example in which data reduction is necessary, is in the context of distributed and cooperative systems for traffic analysis and detection, when a central node must collect several traffic data sets from other nodes and needs to process them. Each node must stream a reduced quantity of its collected data while still preserving the information content needed by the central node to perform its specific analysis and to correlate the data originating from different sources [5].

Therefore, our reduction approach starts from the following idea: collecting a large amount of data (i.e. information) does not necessarily imply that successive analysis stages must use all this data. Indeed, the considered data set may comprise information redundant with respect to the analysis to be performed. Methods can be therefore derived to reduce the original data set with an acceptable loss of the properties of interest. To reach this target we need both methods that work *effectively* and techniques to measure the *effectiveness* of the introduced method.

Exploiting the intrinsic redundancy of network traffic data, in this paper we present a methodology, based on the idea introduced in [30], to reduce network traffic data sets. We show its application in the context of *packet-level* analysis of application traffic and we assess its *effectiveness* discussing a number of statistical properties of both entire and reduced data sets. With the term *packet-level* analysis we mean the study of the statistical properties of inter-packet times and packet sizes generated by a specific Internet application [19] [18] [20].

The adopted approach is applied to the traces of applications drawing increasing interest in the research community: *on-line games*. More precisely, we apply our methodology to the traffic generated by *Counter-Strike* [3] and *Age of Mythology* [13]. Network games, are indeed an interesting class of emerging applications which have recently gained considerable attention because of their different - from traditional Internet applications - QoS requirements and traffic properties, connected to their growing popularity and spread in current networks [22] [23] [21] [20] [18].

In [30] we presented the analytical basis and a proof of concept of an *Entropy*-based reduction methodology. In this work, to assess the efficacy of such approach, we perform a deep analysis, extending previous results. We consider the traffic generated by two *network games* (*Counter-Strike* and *Age of Mythology*) presenting very different properties. The performed analysis aims at evaluating the effects of the reduction on real network traffic data, in terms of marginal distributions, tail behavior, auto-correlation functions, and time-scale properties. Finally, we make the statistical tools used in this work available at [25].

The paper is organized as follows. Section II presents related work. In Section III we give a brief overview of the methodology we used, and we show how this was applied to *Counter-Strike* and *Age of Mythology* traffic. The experimental results are presented in Section IV. Finally, we discuss the findings in Section V drawing also some final conclusions.

II. Related work

The problem of data reduction for collection and analysis is common to several research areas (e.g. genetic [27] and economy [28]). In the field of network traffic, several aspects of data reduction can be considered.

[0] This work has been partially supported by the MIUR in the context of PRIN 2006 RECIPE Project, by CONTENT EU NoE, OneLab and NETQOS EU projects. The authors would like to thank Pierluigi Salvo Rossi for his valuable hints, time, and suggestions.

A deeply investigated approach is based on *sampling*. In this area, some techniques have been developed to reduce the amount of collected data while still being able to extract faithful/reliable information related to the overall traffic traversing a network link. Sampling is usually used for monitoring purposes and to study flow properties, traffic matrices, etc. and many works have been devoted to understand how sampling affects several traffic properties [9] [10] [15].

Some recent approaches aim to use the Principal Component Analysis (PCA) [26] to reduce the dimension of the considered parameter space. This, in turn, results in a reduction of the quantity of information necessary to represent the data.

Other approaches are based on compression techniques. In [17] the authors study the compression of network measurement data of different granularity (SNMP, NetFlow, packet headers), also considering joint coding of data originating from different monitoring points

Finally, some research has been devoted to the analysis of the measurement time period. In [16] the problem of how long the measurement time for collecting traffic traces should be, for classification purposes, has been examined.

III. Reducing Network Traffic Data Sets

The methodology we use in this work is based on the concepts of *Entropy*, *Kullback-Leibler distance*, and *Marginal Utility*. The analytical background has been already introduced in [30]. In such paper it was shown how, starting from these concepts, it is possible to obtain an equation useful to test the goodness of the reduction. Here we recall such formula that represents the concept of off-line *Marginal Utility*. For more information please refer to [30]

The *Kullback-Leibler distance* is used to test whether the addition of a block of traffic samples to an existing set improves our knowledge regarding its marginal distribution. Such metric measures the information gain of the new block with respect to the existing set of traffic samples.

Let M be the size of the data set to be reduced, we divide it into z non-overlapping blocks. Each of them will have a size of $N = \lfloor M/z \rfloor$. We then compute the following expression for $m = 1, \ldots, z$

$$U^z(S^m) = \sum_{x_i \in A} P(x_i^z) \cdot Y_i^m \tag{1}$$

where:

$$Y_i^m = \begin{cases} -\log(P(x_i^z)), & \text{if } P(x_i^m) = 0 \\ \log(\dfrac{P(x_i^z)}{P(x_i^m)}), & \text{otherwise} \end{cases} \tag{2}$$

The reduced data set will be composed of the first j blocks with j being the index for which $U^z(S^j)$ becomes arbitrary smaller than the *Entropy* of the entire data set [8]. If $(j < z)$ we have obtained a reduction of the dataset loosing a controlled quantity of information content.

Here we show how this general methodology can be applied to real network measurements used for traffic characterization and modeling. This work falls in a more general research framework for packet-level analysis of network traffic [19] [18] [20]. For *packet-level* analysis, we mean analyzing network traffic in terms of *Inter-Packet Times* and *Packet Size*, instead of, e.g., flows, connections, sessions, aggregate traffic, etc.

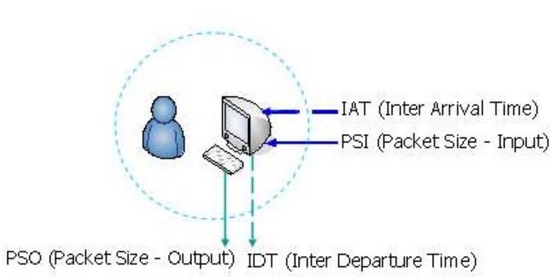

IAT (Inter Arrival Time)
PSI (Packet Size - Input)
PSO (Packet Size - Output)
IDT (Inter Departure Time)

Fig. 1. Packet-level approach: considered variables

With the aim to study network games because of their timeliness, here we consider traffic traces of two popular network games, *Counter-Strike* (CS) and *Age of Mythology* (AoM). As we can see in Table I, the AoM trace is very small, in terms of both bytes and packets, compared with CS. The reason for choosing such a small trace is twofold: first, to verify the applicability of the approach to very small traces, and, second, exploiting the different characteristics of the two games, to investigate the generalizability of the proposed approach.

As for the datasets to be reduced, in this work we consider *Inter-Arrival Times* (IAT) between incoming packets and their corresponding *Packet Size In* (PSI) as well as *Inter-Departure Times* (IDT) between outgoing packets and their corresponding *Packet Size Out* (PSO) (see Fig. 1). The PSI and the PSO (measured in bytes) represent the length of UDP payload, while the IAT and the IDT refer to inter-packet times (measured in seconds) between two consecutive IP packets. The point of view is from the server in the case of *Counter-Strike*, and from a game peer in the case of *Age of Mythology*. The number of disjoint subsets we consider, for each variable, is equal to 100. Also, we consider the reduced dataset to be representative of the entire one when the *Marginal Utility* becomes about $200 - 300$ times smaller than the *Entropy* of the entire dataset.

We make the Matlab scripts used for data reduction and for the statistical analysis freely available at [25].

To analyze the effectiveness of the adopted approach, we compare different statistical properties of the entire and reduced data sets. In particular, to observe the effect of the reduction on the marginal distribution, we plot the quantile of the entire set against the quantile of the reduced one (QQ-plot) and the Probability Density Function (PDF) of the entire and reduced set in the same figure. To highlight the tail behavior of the distributions, we plot the Complementary Cumulative Distribution Function (CCDF). This aspect is important because it has been demonstrated that the presence of particular tail behaviors (e.g. heavy-tails) in statistical distributions related to network traffic is responsible, in some cases, for significant

TABLE I

TRAFFIC TRACES DETAILS.

Application	Scenario	Packets	Size	Log Time
Counter-Strike	Server over WAN	20.000.000	1.6 GB	7h:50m
Age of Mythology	Client over LAN	68613	5 MB	0h:56m

phenomena. Besides the marginal distributions, to investigate the impact of the reduction on both the temporal structure of the samples and their mutual dependencies, we also look at the modifications to the scaling properties introduced by the reduction and we analyze the autocorrelation function of the entire and reduced sets.

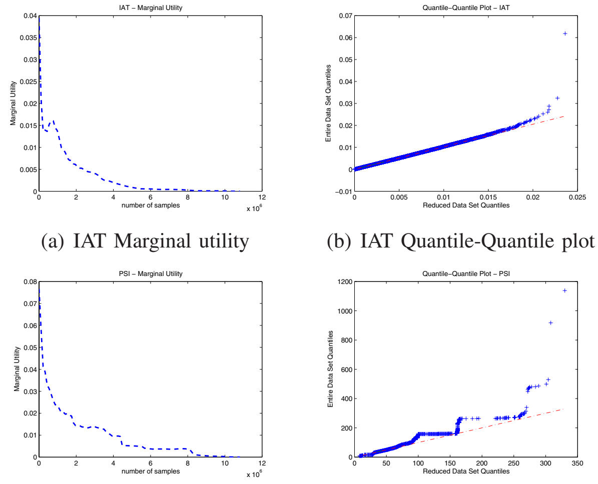
IV. EXPERIMENTAL RESULTS

A. *On the Reduction of Counter-Strike traffic traces*

The traffic trace of *Counter-Strike* server we analyzed comes from one of the most popular on-line gaming communities in the Northwest region of USA, namely *mshmro.com* [11]. *Counter-Strike* is a first-person-shooter (FPS) game, whose wide-spreading goes back as far as year 2000 (with more than 20.000 active servers), when measurements indicated that the application was generating a large percentage of all observed UDP traffic behind DNS and RealAudio traffic [1]. The trace we used (Table I) has been obtained by capturing all the traffic flowing in and out from the gaming server for about 8 hours. The server itself was configured with a maximum capacity of 22 players, which was often reached. The trace collection was limited to 20.000.000 packets (about 8 hours) but the traffic to (10.809.129 $pkts$) and from (9.190.871 $pkts$) the server shows similar behavior even for the rest of the day [23]. This trace has been collected and used in [22] and [23].

1) Entropy-based data sets reduction:

a) IAT: In Fig. 2(a), the *Marginal Utility* of the IAT series, as a function of the number of samples, decays very fast. Ending the reduction with a *Marginal Utility* of 0.021, we obtain a reduced set composed of only 2 experiments.



(a) IAT Marginal utility

(b) IAT Quantile-Quantile plot

(c) PSI Marginal utility

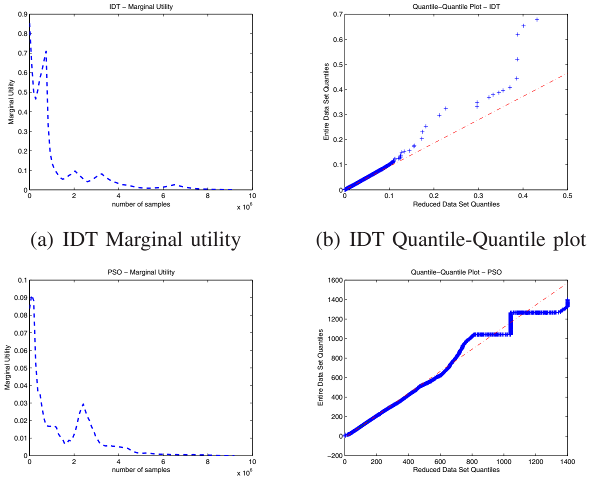
(d) PSI Quantile-Quantile plot

Fig. 2. Reducing IAT and PSI Time Series of *Counter-Strike*[30].

The QQ-plot between the entire and the reduced data sets (reported in Fig. 2(b)) shows that the approximation is quite good for over the 99.9% of the distribution. In Fig. 4, a comparison of the probability density function (PDF) diagrams of the two sets shows the goodness of the approximation. Also, the mean and standard deviation values are preserved (see first row of Table II). Furthermore, in Fig. 5(a), we compare the CCDF, with the y axis plotted in logarithmic scale, of the two distributions. This figure shows a different behavior regarding the largest values. Indeed, while the largest value of the entire data set is $0.82s$, the largest value of the reduced one is $0.0236s$. However, the samples of the entire data set greater than this last value account for only the 0.003% of the distribution, and the difference between the two maximum values is slightly more than one order of magnitude. Therefore, this cannot be considered as a significative change in tail behavior.

b) PSI: Fig. 2(c) shows the *Marginal Utility* as a function of the number of samples. In contrast with the previous case, it falls down slowly. Even so, we obtain a net reduction of 91% (see the second row of Table II). In Fig. 2(d), the QQ-plot (between the entire and the reduced data sets) shows a good approximation until 90 bytes (i.e. the $99.8th$ percentile of the entire data set). The PDF diagrams in Fig. 4 show how close the distributions are. However, there are very few values related to larger packets, which are not present in the reduced data-set. Their amount is so small that cannot be seen in the diagram without zooming several times, and they can be considered as outliers.

Concluding the incoming traffic trace analysis, if we consider the size of the largest of the two reduced data sets (IAT and PSI), we can approximate the entire data set using about 1 million of samples. This means that we have obtained a reduction of about 90%.



(a) IDT Marginal utility

(b) IDT Quantile-Quantile plot

(c) PSO Marginal utility

(d) PSO Quantile-Quantile plot

Fig. 3. Reducing IDT and PSO Time Series of *Counter-Strike*[30].

c) IDT: For this variable, the *Marginal Utility* decays to zero more slowly than the IAT case (Fig. 3(a)). As shown in the third row of Table II, we obtain a reduction of 59%. In Fig. 3(b) and 5(b)) we can see that the approximation is quite good for over the 99.9% of the distribution, and mean and standard deviation are well approximated (see third row of Table II). Figs. 4 and 5(b) show that the two distributions are close in the main part and in the tail too.

d) PSO: We sketch the *Marginal Utility* against the number of samples in Fig. 3(c). The QQ-plot in Fig. 3(d) shows a good approximation up to about 500 bytes, which accounts for 99.2% of the original data set. In the fourth row of Table II a summary of the conducted analysis is reported.

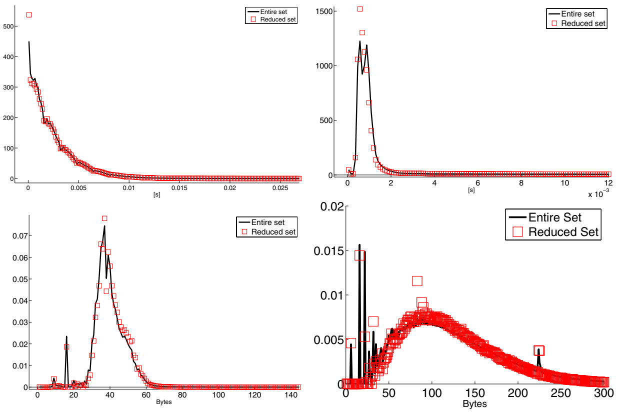
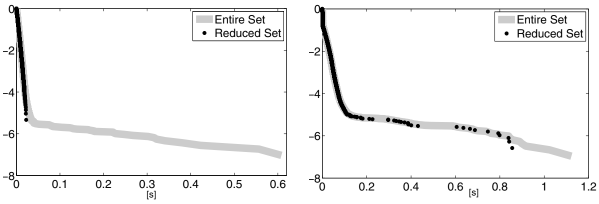


Fig. 4. Counter-Strike PDFs (clockwise: IAT, IDT, PSO, PSI).

Finally, the outgoing traffic is well approximated by an IDT/PSO series of about 4 millions of samples.



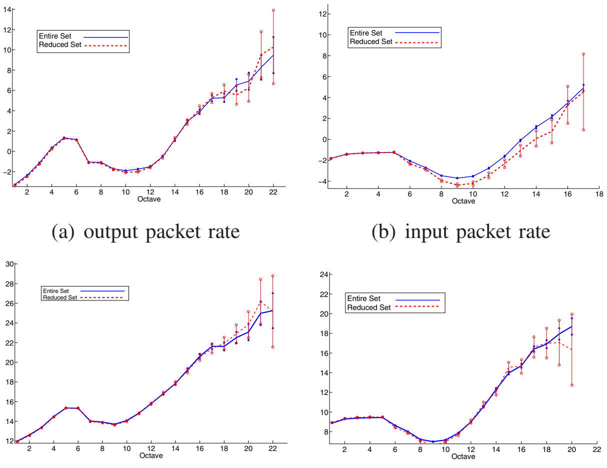
(a) IAT

(b) IDT

Fig. 5. CCDF of Counter-Strike.

2) Wavelet Analysis of CS Reduced Data Sets: The reduction criterion we use here is based on the analysis of the marginal distributions of traffic data samples. But, in the study of network traffic also temporal structures and dependencies (e.g. *long range dependence* and *scaling* behavior) can be of interest. In this section, we briefly show a time-frequency analysis based on the *Wavelet Transform*, revealing similar behaviors between the entire and reduced data sets. We use the *Logscale Diagram* (*LD*), which shows the trend of the energy of the wavelet coefficients at each time scale, allowing to estimate the scaling behavior of the considered process and the corresponding *Hurst* parameter (see [12]).

From the *Counter-Strike* IAT and IDT data sets, we calculated the packet rate time series, with a period of 1 *ms*, of traffic flowing in both directions (to and from the server).



(a) output packet rate

(b) input packet rate

(c) output byte rate

(d) input byte rate

Fig. 6. Logscale Diagram comparison of *CS* reduced and original data sets.

Let S_j, S_j^1 be the logarithms of the energy of the wavelet coefficients at scale j of respectively the entire and reduced data sets. We found $S_j =_{\sigma_j} S_j^1$ (for $j = 1, ..., 17$ in the case of IAT and for $j = 1, ..., 22$ in the case of IDT) where the $=_{\sigma_j}$ operator takes into account their confidence intervals. This can be seen in the LDs in Figs. 6(a) and 6(b), where, at each scale, the confidence intervals of the two diagrams always intersect. It is worth noting that we found the same results for the byte rate time series (Figs. 6(c) and 6(d)), which were obtained by combining information from the IAT and PSI series as well as IDT and PSO series. This comparison is indeed important, since it is highlights properties of real network traffic by combining information on packet arrival times and their size. The analysis in this section shows that, for the considered data sets, the reduction did not heavily affect the traffic temporal structures.

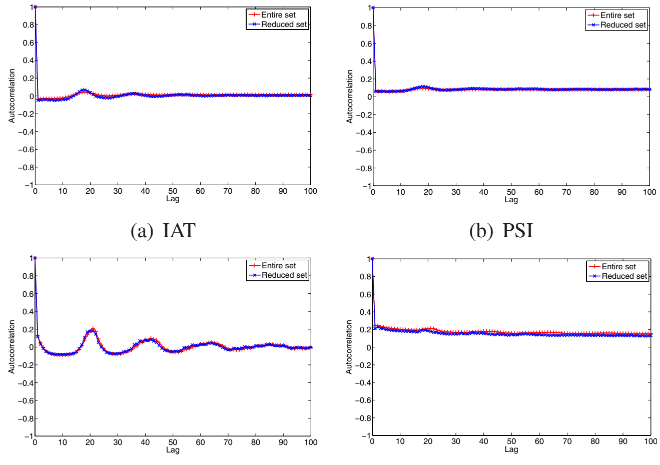
3) Effects of the Reduction on the Autocorrelation: Beside the wavelet spectrum of the packet rate and byte rate series, we study the behavior of the Autocorrelation function for both complete and reduced data sets of IDT and PS. This is done to further assess the impact of the adopted approach on the samples temporal behavior and their mutual dependencies.

In Fig. 7 the autocorrelation plots, until lag 100, are reported for all the data traces. As shown, for all the considered variables, the autocorrelation values of the reduced sets are very close to those of the original sets. In particular, the Root Mean Square (RMS) value of the error introduced by the reduction ranges from 0.0128 (for the IAT series) to 0.0232 (for the PSO series). To better observe the effect of the reduction, a zoomed view of the IAT autocorrelation is reported in Fig. 8. IAT is the variable that presents more correlation among the samples, also, its autocorrelation plot reveals an oscillating trend. The view of Fig. 8 allows to verify that the trend of the reduced-set autocorrelation is very similar to that of the original set. This witnesses that the temporal structure of the samples is preserved even in the presence of a such particular behavior.

TABLE II

Counter-Strike DATA SET REDUCTION[30].

	Size [sample]	**Mean**	**StDev**	**Entropy [bit]**	**Reduced Size [sample]**	**Mean**	**StDev**	**Reduction**	**Marginal Utility [bit]**
IAT	10809129	0.0023614 s	0.0023564 s	7.83	216183	0.0023491 s	0.0022617 s	98%	0.021
PSI	10809129	39.559 bytes	9.6741 bytes	4.93	972822	40.331 bytes	8.9248 bytes	91%	0.024
IDT	9190871	0.0027772 s	0.0062425 s	9.11	3768258	0.0028466 s	0.0064410 s	59%	0.045
PSO	9190871	127.68 bytes	100.42 bytes	7.89	459544	127.03 bytes	98.53 bytes	95%	0.036



(a) IAT

(b) PSI

(c) IDT

(d) PSO

Fig. 7. Autocorrelation plot of CS reduced and original data sets.

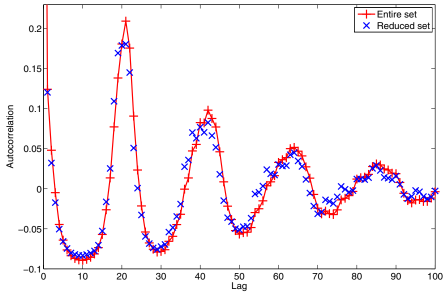


Fig. 8. Autocorrelation plot of IDT of CS reduced and original data sets.

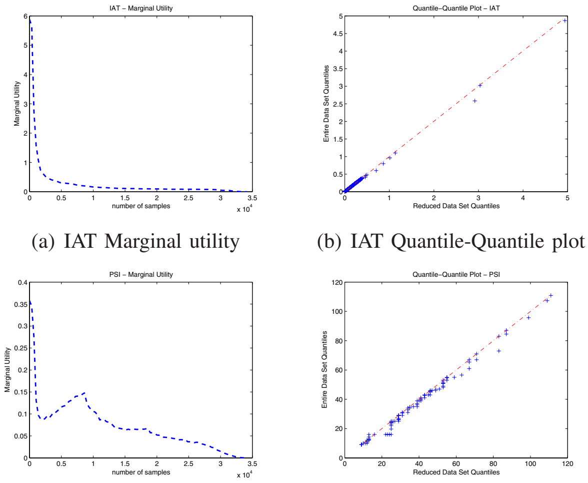
B. On the Reduction of Age of Mythology traffic traces

The traffic trace of Age of Mythology (UDP traffic on port 2500) has been provided, in Tcpdump format and with a time resolution $\leq 1us$, by the Worcester Polytechnic Institute (WPI), MA (USA) [14]. *Age of Mythology* is a popular game, representative of another category: real-time strategy (RTS) games. The trace (Table I) is related to 1 hour of traffic and it consists of packet sequences of a complete gaming session, between two players, captured in a LAN environment. The extracted data sets have been used in [24] and [18]. With *outbound* traffic (described by means of IDT and PSO) we mean 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). With *inbound* traffic (described by means of IAT and PSI) we refer to the opposite direction. The trace collection was limited to about $68,000$ packets (5 MB data) captured during *August 2003*.

It is worth noting that, due to space constrains, we present the AoM results following the same approach of the CS results but only focusing on the principal findings.

1) Entropy-based data sets reduction:

a) IAT: In Fig. 9(a) the *Marginal Utility* as a function of the number of samples is shown. In this case the data set reduction is only about 7%. The scarce reduction obtained in this case is due to the small number of samples present in the entire data set. This results in a high information content provided by each subset (i.e. each experiment).



(a) IAT Marginal utility

(b) IAT Quantile-Quantile plot

(c) PSI Marginal utility

(d) PSI Quantile-Quantile plot

Fig. 9. Reducing IAT and PSI Time Series of AoM.

A summary of the performed reduction is given in the first row of Table III, while in Fig. 9(b) the QQ-plot between the entire data set and the reduced one is shown. Given such a small reduction, the very good approximation of the IAT marginal distribution shown in Fig. 10, is easily expected.

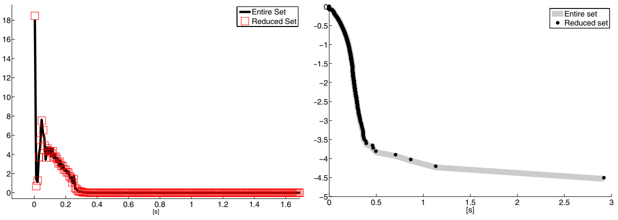


Fig. 10. AoM IAT (pdf left, ccdf right).

b) PSI: In Fig. 9(c) the *Marginal Utility* as a function of the number of samples is shown. In this case the data set reduction is of 10% as shown in the second row of Table III.

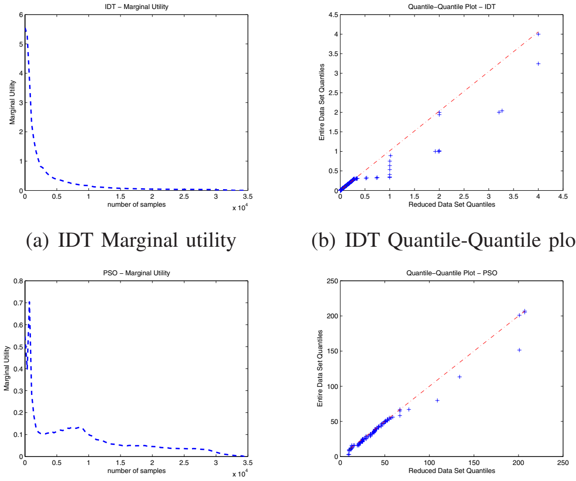
TABLE III

Age of Mythology DATA SET REDUCTION.

	Size [sample]	Mean	StDev	Entropy [bit]	Reduced Size [sample]	Mean	StDev	Reduction	Marginal Utility [bit]
IAT	34133	0.0986 s	0.0796 s	7.76	31744	0.0980 s	0.08 s	7%	0.031
PSI	34133	12.595 bytes	4.686 bytes	2.09	30720	12.592 bytes	4.835 bytes	10%	0.01
IDT	34480	0.0982 s	0.0727 s	7.96	22068	0.0959 s	0.0723 s	36%	0.039
PSO	34480	12.397 bytes	3.979 bytes	2.09	31722	12.386 bytes	4.064 bytes	8%	0.01

The QQ-plot (Fig. 9(d)) between the entire data set and the reduced one shows a good approximation.

c) IDT: In Fig. 11(a) the *marginal utility* as function of the number of samples is shown. As shown in Fig. 11(b), considering that the $99,9th$ percentile of the entire data set is equal to 0.3, the reduced dataset well approximates the entire one. For a summary of the analysis refer to the third row of Table III.



(a) IDT Marginal utility

(b) IDT Quantile-Quantile plot

(c) PSO Marginal utility

(d) PSO Quantile-Quantile plot

Fig. 11. Reducing IDT and PSO Time Series of AoM.

d) PSO: In Fig. 11(c) the *Marginal Utility* against the number of samples is sketched. A summary of the conducted analysis is shown in the fourth row of Table III. The QQ-plot (Fig. 11(d)) indicates a quite good approximation.

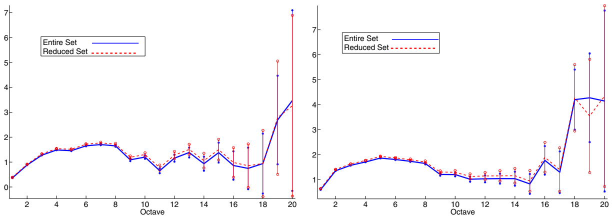
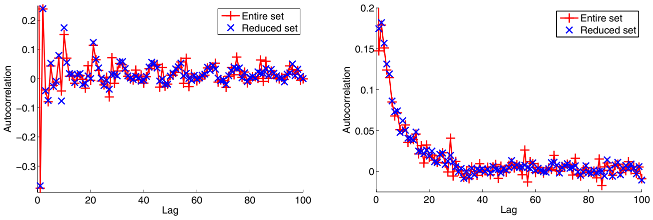


Fig. 12. Logscale Diagram comparison of *AoM* reduced and original data sets (output and input byte rate on the left and right diagram respectively).

2) Wavelet Analysis of AoM Reduced Data Sets: We performed a wavelet-based analysis of the inbound and outbound traffic data sets of Age of Mythology. The logscale diagrams related to the byte rate (calculated with a period of 1 *ms*) are shown in Fig. 12. It can be seen that, even in this case, we found consistency between the reduced and entire data sets.

3) Effects of the Reduction on the Autocorrelation: In order to verify the consistence between the autocorrelation functions of the entire and reduced set, Fig. 13 depicts the autocorrelation plots of the IAT and PSI series. As we can see, negligible differences exist. The IDT and PSO series autocorrelations are not shown because of space constraints, however, similar considerations apply.



(a) IAT

(b) PSI

Fig. 13. Autocorrelation plot of AoM reduced and original data sets.

V. DISCUSSION AND CONCLUDING REMARKS

In this paper we applied an off-line *Entropy*-based approach to traffic data set reduction (introduced in [30]). We showed its application, along with a critical analysis of the results and side-effects, in a framework of traffic characterization at packet level. We chose network games because they represent an important category of Internet applications generating novel traffic patterns. In the choice of the games we selected two different test cases. Indeed, even if both *Counter-Strike* and *Age of Mythology* run over UDP, the two considered traces differ in several aspects: game typology (RTS vs. FPS), network observed (LAN vs. WAN), time duration, size in bytes, etc. Moreover, in the case of *Counter-Strike* we have analyzed the traffic related to a gaming server, serving more than 20 players, whereas for *Age of Mythology* the traffic observed is related to the workstation of a single player fighting against one single opponent to which was directly connected (therefore traffic has symmetric properties). This broad range of different parameters allows a wider view on the applicability of the adopted reduction technique and to more easily spot pros and cons.

In the networking field, the most used techniques to reduce data sets are based on *sampling*. We believe that the presented approach is complementary to it. *Sampling* is suited for on-line applications aiming to produce reports that are quick and concise rather than accurate and complete. Our approach allows to characterize (and model) network traffic without losing sensible information. Moreover, *sampling* requires the data set to be strict-sense [10] or wide-sense [9] stationary.

Under such conditions, it can accurately approximate second order statistics like the *Hurst* parameter, but it could still fail to capture the mean [9]. Further, it is able to reconstruct the wavelet spectrum (at least at low frequencies) under particular conditions[10]. According to these considerations, as shown in the previous sections, the adopted Entropy-based off-line technique presents the advantages of correctly capturing mean, standard deviation, and marginal distributions. Moreover, in the analyzed traces we investigated side-effects on time properties, and we found they were not sensibly compromised by the reduction (even if the original data set is nonstationary). We investigated these properties by comparing the wavelet spectrum log-scale diagrams of the packet rate and byte rate series, and looking at auto-correlations. Also, as a further proof, the entire and reduced data sets were tested with the model proposed in [18] finding promising results (not shown here for the sake of brevity and because the proposed model is out of the scope of this paper). Such model, based on Hidden Markov Models, is specifically designed to take into account and reproduce - in a joint fashion - the temporal structures and the mutual dependencies (besides the marginal distributions) of the inter-packet time and packet size data sets. We found that, the model parameters did not significantly change when it was trained with the reduced data sets. However, in the present work, we also found that the reduction that can be obtained without relevant loss of information is not very effective in the case of smaller data sets. Moreover, a problem to be further investigated is about the amount of predictability or control on the quantity of reduction resulting from the application of the proposed methodology. Sampling techniques, for example, often allow to anticipate the amount of data to be stored.

Finally, in the case of large data sets, a loss of some information related to outliers or to very rare values is possible. Indeed, the threshold used as a stopping point implicitly defines the probability associated to the discarded samples. However, in the considered test-cases, the reduction did not affect important properties such as, for example, tail behaviors.

The experimental results have shown that the adopted approach reveals promising. In our ongoing work we plan to apply it to more typologies of network traffic data sets, also enlarging the set of statistical indicators to be studied.

References

[1] S. McCreary and k. Claffy, "Trends in Wide Area IP Traffic Patterns: A View from Ames Internet Exchange," 13th ITC Specialist Seminar on Measurement and Modeling of IP Traffic, pp. 111, Sep. '00.

[2] D.S. Jackson, "Video Games: A room Full of Doom," *Time Magazine* (US Edition), (153):20 May '99.

[3] `http://www.counter-strike.net/`

[4] Y. Liu, D. Towsley, J. Weng, D. Goeckel, "An Information Theoretic Approach to Network Trace Compression", UM-CS-2005-003 TR, Jan. '05

[5] L. Huang, M. Garofalakis, J. Hellerstein, A. Joseph, and N. Taft "Toward Sophisticated Detection With Distributed Triggers", ACM SIGCOMM Workshop on Mining Network Data (MineNet). September, 2006.

[6] C. Shannon "A Mathematical Theory of Communication", *Bell Systems Technical Journal*, (47):143-157, 1948.

[7] P. Barford, A. Bestavros, J. Byers, M. Crovella. "On the marginal utility of network topology measurements", IMW, Nov. '01.

[8] Y. Zhang, M. Roughan, C. Lund, D. Donoho, "Estimating Point-to-Point and Point-to-Multipoint Traffic Matrices: An Information-Theoretic Approach", *IEEE/ACM Trans. on Networking*, Mar. '04.

[9] G. He, J. C. Hou, "An In-Depth, Analytical Study of Sampling Techniques For Self-Similar Internet Traffic", ICDCS '05, pp. 404-413, Jun. '05.

[10] N. Hohn, D. Veitch, "Inverting Sampled Traffic", IMC, pp. 27-29, Oct. 03.

[11] `http://www.mshmro.com`

[12] D. Veitch, P. Abry, "A wavelet based joint estimator for the parameters of LRD", IEEE Trans. Info. Th. Vol. 45, No.3, Apr. '99.

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

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

[15] K. C. Claffy, G. C. Polyzos, and H. W. Braun, "Applications of Sampling Methodologies to Network Traffic Characterization". ACM SIGCOMM Conference, 1993.

[16] M. Ilvesmaki, S. Kaikkonen, "The length of measurement period to determine the application profile for traffic classification in the Internet". IEEE ICC 2001. Vol.6 pp. 1851-1855

[17] Y. Liu, D. Towsley, T. Ye, and J. Bolot, "An Information-theoretic Approach to Network Monitoring and Measurement" IMC'05

[18] A. Dainotti, A. Pescapé, P. Salvo Rossi, G. Iannello, F. Palmieri, G. Ventre, "An HMM Approach to Internet Traffic Modeling," *IEEE GLOBECOM*, Nov. 2006.

[19] A. Dainotti, A. Pescapé, G. Ventre, "A Packet-level Characterization of Network Traffic", 11th IEEE International Workshop on Computer-Aided Modeling, Analysis and Design of Communication Links and Networks (CAMAD 2006)

[20] A. Dainotti, A. Pescapé, G. Ventre, "A packet-level model of Starcraft traffic", Second International Workshop on Hot Topics in Peer-to-Peer Systems 2005 (co-located with IEEE Mobiquitous 2005) pp. 244-253, July 2005

[21] M. Claypool, D. LaPoint, and J. Winslow, "Network Analysis of Counter-strike and Starcraft". 22nd IEEE International Performance, Computing, and Communications Conference (IPCCC), April 2003.

[22] W. Feng, F. Chang, W. Feng, J. Walpole, "Provisioning On-line Games: A Traffic Analysis of a Busy Counter-Strike Server". SIGCOMM Internet Measurement Workshop, November 2002.

[23] W. Feng, F. Chang, W. Feng, J. Walpole, "A Traffic Characterization of Popular On-line Games", IEEE/ACM Transactions on Networking, vol. 13, no. 3, June 2005

[24] M. Claypool, "The Effect of Latency on User Performance in Real-Time Strategy Games", Technical Report WPI-CS-TR-03-32, Computer Science Department, Worcester Polytechnic Institute, Oct. 2003. Online at: ftp://ftp.cs.wpi.edu/pub/techreports/pdf/03-32.pdf

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

[26] N. Patwari, A. O. Hero, and A. Pacholski, "Manifold Learning Visualization of Network Traffic Data", SIGCOMM 2005 - Workshop on Mining Network Data, August 26, 2005.

[27] "Reagent sets and gene signatures for renal tubule injury", U.S. Patent Application n. 20060057066

[28] J.L. Eltinge, D.S. Jang and M.J. Cho, "Use of Generalized Variance Function Models in Inference from Social and Economic Survey Data", Statistics Canada International Symposium Series (2002)

[29] S. Gargolinski, C. St. Pierre, and M. Claypool, "Game Server Selection for Multiple Players", ACM Network and System Support for Games (NetGames), 2005

[30] A. Pescapé, "Entropy-Based Reduction of Traffic Data", IEEE Communications Letters, Vol.11, No.2, February 2007.