# Entropy-Based Reduction of Traffic Data

Antonio Pescapè, *Member, IEEE*

***Abstract*— This letter proposes an *Entropy*-based methodology to reduce large network traffic data sets obtained by measurements over real networks. The proposed off-line approach, based on the *Marginal Utility* concept, reveals interesting results when applied to real data captured over real networks: to show its applicability, results obtained with traffic traces from a popular network game, *Counter-Strike*, are presented.**

***Index Terms*— Communication system traffic, network operations and management, performance analysis.**

## I. Introduction

THE collection of traffic traces from real networks poses tremendous challenges due to the high speeds of current networks. In one hour, the collection of 60 byte packet headers on an OC-48 link can easily generate 600 GB of data [1]. After the collection stage, data need to be analyzed to obtain several information (e.g. packet size statistical properties, port distribution, ...). As the amount of data becomes larger, the time to analyze them increases. Finally, working with large data sets several problems can occur (e.g. *out of memory* errors under the software environment used for data processing and statistical analysis). To cope with this kind of issues, we need methods to reduce the original data set with an acceptable loss of statistical properties.

This letter proposes an *Entropy*-based methodology to reduce data sets of network traffic. The proposed approach has been applied to the traffic traces of *Counter-Strike* [2], a popular on-line game. In [3] it is reported that $3-4\%$ of all packets in a backbone could be associated with only 6 popular games, and, in USA alone, they are currently worth a significant fraction of the 7 billion computer games industry [4]. Also, in [5] it is reported that, by the year 2010, multiplayer network games will be likely responsible for over $25\%$ of LAN traffic. The dominance of Counter-Strike (with more than 20.000 active servers) goes back as far as year 2000, when measurements indicated that the application was generating a large percentage of all observed UDP traffic.

## II. Definitions and Objectives: An Analytical Basis

The proposed approach is based on the distance between two statistical distributions. The concepts of *Entropy*, *Kullback-Leibler distance*, and *Marginal Utility* are used. The *Entropy* of a discrete random variable is defined as

$$H(X) = -\sum_{x_i \in A} P(x_i) \cdot \log P(x_i) \tag{1}$$

where $A$ is the population space, i.e. the space composed of all possible outcomes $x_i$ of the random variable $X$. The quantity $-\log(P(x_i))$ is defined as the information content, measured in bits if $log$ is taken to base 2, revealed from the outcome $x_i$ of an experiment [6]. The *Kullback-Leibler Entropy* is defined as

$$K(f,g) = \sum_{x_i \in A} f(x_i) \cdot \log \frac{f(x_i)}{g(x_i)} \tag{2}$$

where $f(x)$ and $g(x)$ are two probability densities to be compared. This measure is more often referred as the "*Kullback-Leibler distance*" since it presents some distance characteristics.

The *Kullback-Leibler distance* was exploited in [7] to measure the *Marginal Utility* of adding new results to an aggregate data set of network topology measurements. Intuitively, the *Marginal Utility* of the experiment $S^m$ with $m > 1$, can be estimated considering the reduction in uncertainty provided by this experiment. The *reduction* in uncertainty for each outcome $x_i$, after the experiment $S^m$, is

$$-\log(P(x_i^{m-1})) + \log(P(x_i^m)) = \log(\frac{P(x_i^m)}{P(x_i^{m-1})}) \tag{3}$$

where $P(x_i^j)$ is the probability associated with outcome $x_i$ after the conclusion of experiments $S^1, S^2, ..., S^j$. The *Marginal Utility* is then defined as the *mean reduction* in uncertainty caused by the addition of the results of a new experiment to the aggregate set. This quantity can be therefore measured using the *Kullback-Leibler distance*. In [7] two alternatives for calculating the *Marginal Utility* are presented: an on-line and an off-line approach. The on-line *Marginal Utility* of $S^m$ is defined as

$$U(S^m) = \sum_{x_i \in A} P(x_i^m) \log(\frac{P(x_i^m)}{P(x_i^{m-1})}) \tag{4}$$

meaning that we determine the *Marginal Utility* of the experiment $S^m$ *before* performing any additional experiment $S^k$, $k > m$. Clearly, the utility of supplemental experimentation decreases as the additional experiments do not bring out new insights, meaning that, the probability distribution of the outcomes of an experiment converges. In the context of network traffic analysis, the on-line *Marginal Utility* can not be used: because of the *nonstationarity* of network traffic, we cannot *a priori* choose the instant $t$ at which the considered data set is statistically representative of the entire process. A different formulation of *Marginal Utility* - the off-line approach - appraises each experiment on an *ex post* basis, by measuring off-line the usefulness of each experiment after all of them have been conducted. The off-line *Marginal Utility* of the experiment $S^m$, with $m \leq z$, is defined as

$$U^z(S^m) = \sum_{x_i \in A} P(x_i^z) \log(\frac{P(x_i^z)}{P(x_i^m)}) \tag{5}$$

The main difference between on-line and off-line *Marginal Utility* is that the latter considers the *Marginal Utility* from the perspective of the complete set of experiments.

Manuscript received July 10, 2006. The associate editor coordinating the review of this letter and approving it for publication was Dr. Charalambos Charalambous.

The author is with the Dipartimento di Informatica e Sistemistica, University of Napoli "Federico II," Via Claudio, 21 - 80125, Napoli, Italy (email: pescape@unina.it).

Digital Object Identifier 10.1109/LCOMM.2007.061068.

This letter proposes a slightly modified version of the off-line *Marginal Utility* for the purpose of reducing network traffic data sets.

## III. Using *Marginal Utility* to Reduce Data Sets

In this work, the *Kullback-Leibler distance* is used with the aim to understand how the addition of a block of new traffic samples to an existing set improves our knowledge of its marginal distribution. This metric will quantitatively express the information gained adding the new block to the existing set of traffic samples. In [7], a group of $z$ identical (i.e. aimed at discovering a common property) successive experiments $S^1, S^2, ..., S^z$ is considered. The *Kullback-Leibler distance* is then applied to the results of such experiments. In the framework proposed in this letter we will consider, as experiment outcomes, blocks of length $N$ of data samples extracted from the same network traffic trace. Let $M$ be the size of entire data set to be reduced, we divide it into $z$ non-overlapping blocks of size $N = \lfloor M/z \rfloor$, where the $N$ samples of block $j$ represent the results of experiment $S^j$. Then, we compute the following expression for $m = 1, \ldots, z$

$$U^z(S^m) = \sum_{x_i \in A} P(x_i^z) \cdot Y_i^m \tag{6}$$

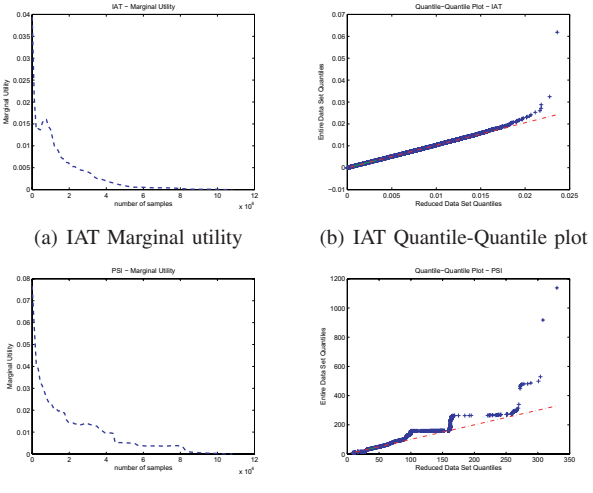
where:

$$Y_i^m = \begin{cases} -\log(P(x_i^z)), & \text{if } P(x_i^m) = 0 \\ \log(\dfrac{P(x_i^z)}{P(x_i^m)}), & \text{otherwise} \end{cases} \tag{7}$$

If $P(x_i^m) \neq 0 \ \forall i$, (6) becomes (5). Compared to network topology measurements [7], in the case of network traffic, this adjustment is necessary when a possible outcome $x_i^j$, with $j < z$, never occurs in the first $j$ experiments (the first $j$ blocks - of length $N$ - extracted from the entire data set, have not yet "discovered" such outcome). In this case the quantity of information gained considering this new outcome is just given by its *information content* and the reduced data set is composed by the first $j$ blocks when, for $j < z$, $U^z(S^j)$ becomes arbitrary smaller than the *Entropy* of the entire data set [8].

## IV. Experimental Results

We analyzed a traffic trace of a *Counter-Strike* server of one of the most popular on-line gaming communities in the Northwest region of USA, *mshmro.com*. Note that while the trace collection was limited to 20.000.000 packets (about 8 hours), traffic to and from the server exhibits similar behavior even for the rest of the day [9]. To show the applicability of the proposed reduction techniques, *Inter-Arrival Times* (IAT) and *Inter-Departure Times* (IDT) as well as *Packet Size In* (PSI) and *Packet Size Out* (PSO) data traces are considered. The point of view is that of the server, and 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 packets. In order to perform the data set reduction, for each considered variable, we have split the entire data set into 100 non-overlapping intervals, where each of them represents a new experiment. As for the reduction stopping point, we consider the *Marginal Utility* to be negligible when it is about $200 - 300$ times smaller than the *Entropy* of the entire data set. Table I contains a summary on the reduction results.




(a) IAT Marginal utility (b) IAT Quantile-Quantile plot

(c) PSI Marginal utility (d) PSI Quantile-Quantile plot

Fig. 1. Reducing IAT and PSI Time Series.

*a) IAT:* Fig. 1(a) shows that the off-line *Marginal Utility* of the IAT series, plotted as a function of the number of samples, drops down very fast. The *Entropy* of the entire IAT series is equal to 7.83 bits, and stopping at a *Marginal Utility* of 0.021, the reduced set is composed of just the first two experiments. In Fig. 1(b) the QQ-plot between the entire data set and the reduced one is shown (the approximation is quite good for over the 99.9% of the distribution). Also, the mean and standard deviation values reported for the entire and reduced data sets respectively are very close (Table I).

*b) PSI:* In Fig. 1(c) the off-line *Marginal Utility* as a function of the number of samples is shown. Note that in this case the *Marginal Utility* falls down more slowly than the IAT series. Despite this, we are able to reduce the original data set up to 91%, as shown in the second row of Table I. The QQ-plot (Fig. 1(d)) between the entire data set and the reduced one shows a good approximation of the entire data set until 90 bytes, which is the $99.8th$ percentile of the entire data set.

*c) IDT:* In this case the *Marginal Utility* tends to zero more slowly than in the IAT case (Fig. 2(a)). As we can observe from the third row of Table I, the reduction is equal to 59%. Also in this case (see Fig. 2(b)), the approximation is quite good for over the 99.9% of the distribution, and mean and standard deviation are well approximated.

*d) PSO:* In Fig. 2(c) the off-line *Marginal Utility* against the number of samples is sketched. A summary of the conducted analysis is shown in Table I. The QQ-plot (Fig. 2(d)) indicates a good approximation up to about 500 bytes, which accounts for 99.2% of the original data set. Therefore, by considering the size of the largest reduced data set, between IAT and PSI, we can approximate the incoming flow of traffic with about 1 million of samples, obtaining a net reduction of about 90%. Also, as for the outgoing traffic, it is well approximated by an IDT/PSO series of about 4.000.000 samples.

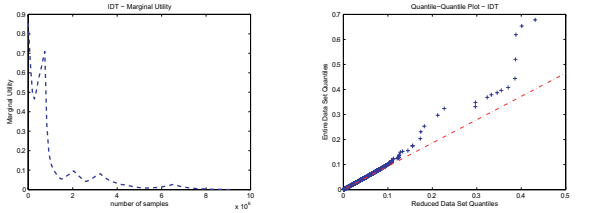
### A. Wavelet Analysis of Reduced Data Sets

The reduction criterion presented in this letter is based on the analysis of the marginal distributions of traffic data samples. Another aspect of network traffic is related to the temporal structures and dependencies (eg. *long range dependence* and *scaling* behavior properties of network traffic).

TABLE I

*Counter-Strike* DATA SET REDUCTION.

| | 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) IDT Marginal utility (b) IDT Quantile-Quantile plot

(c) PSO Marginal utility (d) PSO Quantile-Quantile plot

Fig. 2. Reducing IDT and PSO Time Series.

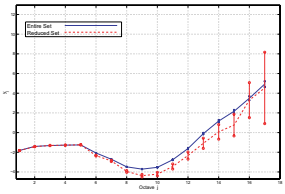



Fig. 3. Logscale Diagram comparison of reduced and original data sets.

In this Section, we briefly present a time-frequency analysis of the reduced data sets based on the *Wavelet Transform*, revealing behaviors similar to the entire ones. The estimation technique exposed in [10] is based on the *Discrete Wavelet Transform* of a random process $X$ of size $N$. A dyadic decomposition is applied, so that the number of considered scales is $J \approx log_2(N)$. The so-called *Logscale Diagram* (*LD*) shows the trend followed by (the logarithm of) the energy of the wavelet coefficients at each scale, allowing to estimate the scaling behavior of the process $X$ and the *Hurst* parameter.

From the *Counter-Strike* IAT data set, we calculated the packet rate time series, with a period of 1 ms. The relative *LD* is shown in Fig. 3. 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 $j = 1, ...17$, where the $=_{\sigma_j}$ operator reports if the energy estimates are consistent by taking into account their confidence intervals. This can be seen in the graph, where, at each scale, the confidence intervals of the two diagrams always intersect. We found the same result for the bitrate (obtained from IAT and PSI series), and also for the outgoing traffic (IDT/PSO). Thus, for the considered data sets, the reduction did not heavily affect the traffic temporal structures.

## V. CONCLUDING REMARKS

In this letter we proposed an off-line *Entropy*-based approach for reducing data sets, applied to *Counter-Strike* traffic traces. Basically, while sampling techniques work fine for on-line approaches aiming at producing network traffic statistics (and in which it is most important to have quick and concise reports even if with approximated values), our approach turns useful when large data sets are used to completely characterize (and model) network traffic without losing sensible information. Therefore, the presented approach is complementary to *sampling* approaches, whose main requisite is that the data set must be strict-sense [11] or wide-sense [12] stationary. In these hypothesis, *sampling* techniques can accurately capture second order statistics like the *Hurst* parameter, while they could fail to capture the mean [12]. Also, under particular conditions, they can reconstruct the wavelet spectrum of the original data set, at least at low frequencies [11]. According to these considerations, the proposed off-line technique to reduce traffic trace data sets presents the advantage of correctly capturing mean, standard deviation, and marginal distributions, without compromising time properties (even if the original data set is nonstationary). Further analysis is needed to better understand the behavior of other relevant statistical and temporal properties (e.g. tails behavior, autocorrelation, bivariate distributions, ...) after the reduction.

## REFERENCES

[1] Y. Liu, D. Towsley, J. Weng, D. Goeckel, "An information theoretic approach to network trace compression," UM-CS-2005-003 TR, Jan. 2005
[2] `http://www.counter-strike.net/`
[3] S. McCreary and K. Claffy, "Trends in wide area IP traffic patterns: a view from Ames Internet Exchange," in *Proc. 13th ITC Specialist Seminar on Measurement and Modeling of IP Traffic 2000*, pp. 111.
[4] D. S. Jackson, "Video games: a room full of doom," *Time Magazine* (US Edition), vol. 153, no. 20, May 1999.
[5] S. Gargolinski, C. St. Pierre, and M. Claypool, "Game server selection for multiple players," ACM NetGames 2005
[6] C. Shannon "A mathematical theory of communication," *Bell Systems Technical J.*, vol. 47, pp. 143-157, 1948.
[7] P. Barford, A. Bestavros, J. Byers, and M. Crovella. "On the marginal utility of network topology measurements," ACM IMW, Nov. 2001.
[8] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "Estimating point-to-point and point-to-multipoint traffic matrices: an information-theoretic approach," *IEEE/ACM Trans. Networking*, vol. 13, no. 5, pp. 947-960, Oct. 2005.
[9] W. Feng, F. Chang, W. Feng, and J. Walpole, "A traffic characterization of popular on-line games," *IEEE/ACM Trans. on Networking*, vol, 13, no. 3, pp. 488-500, June 2005.
[10] D. Veitch and P. Abry, "A wavelet based joint estimator for the parameters of LRD," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 878-897, Apr. 1999.
[11] N. Hohn and D. Veitch, "Inverting Sampled Traffic," IMC, pp. 27-29, Oct. 2003.
[12] G. He and J. C. Hou, "An in-depth, analytical study of sampling techniques for self-similar Internet traffic," in *Proc. ICDCS '05*, pp. 404-413.