STATISTICAL ANALYSIS OF LOCAL FEATURES IN NETWORK TRAFFIC PROCESSES

Giada Giorgi, Claudio Narduzzi, Paolo Attilio Pegoraro

Department of Information Engineering DEI – University of Padova
Via G. Gradenigo, 6/b – Padova, Italy – e-mail: {*giada, narduzzi, pego*}*@dei.unipd.it*

ABSTRACT

This work presents an approach to the detection of local features in network traffic, based on the analysis of short-time maximal rate envelopes, also called statistical arrival curves. In the proposed method, the time series representing a traffic trace is divided into non-overlapping segments, which are further divided into smaller blocks. The maximal rate envelope is estimated for each block and histograms of rate parameters are built over each segment. When significant local features are present in a trace segment, values of the maximal rates may change, resulting in the appearance of peaks or long tails in the corresponding histograms. These effects can be detected with remarkable sensitivity, since they are often evidenced by positive or negative peaks in skewness values of rate parameters histograms. The algorithm can be employed to detect such features on a reasonably fine-grained scale.

1. INTRODUCTION

Passive measurement in computer networks deals with the collection and analysis of traffic data, from which features of the analysed flow can be estimated. A significant area of research is concerned with modelling and prediction of network behaviour, which are relevant both to traffic engineering and to flow control and optimization. For these purposes, experimental data are represented by traffic traces. A traffic trace is a time series that gives the measured number of information units (i.e., a count of packets, bytes, etc.) for a succession of time slots. The aim of the analysis is to describe network traffic as accurately as possible by a suitable model.

Traffic has long been known to present similar features when observed at different time scales [1]. Experimental analysis showed that long-range dependence (LRD) is found in traffic traces, which means that, as the time scale gets larger, traffic is increasingly well modelled by a self-similar random process. This led to the widespread acceptance of self-similar processes as basic traffic models over time scales ranging from some tens of seconds upwards [2]. In turn, the scale-invariance feature of self-similar processes provided the rationale for the use of time-scale algorithms, namely, wavelet transforms, as a common analysis tool [3], [4].

More recently, research has turned to the analysis of shorter time scales, roughly from some tens of ms up to 1 s [5], where a more complex behaviour can be found. A number of works based on wavelet transform analysis have shown that multi-scaling actually occurs, suggesting that more complex models may be needed to describe traffic behaviour at these scales. However, the issue is still somewhat controversial, as it has also been shown that other features in the traffic process, namely, non-stationarities, may be mistaken for multiscaling [6], or may even affect second-order scaling behaviour estimation [7]. In this respect, it appears that wavelet analysis alone may not give a clear enough picture, particularly when the local features of a time series are concerned.

This work introduces a different approach, based on an alternative set of statistical features of traffic processes. The proposed method can be employed both to estimate traffic parameters at large time scales and to detect, on a reasonably fine-grained scale, local features that may affect the outcome of scaling analysis. As such, the approach complements wavelet algorithms and should prove useful particularly in the understanding of short time scale phenomena in network traffic.

2. fBm TRAFFIC MODEL AND RATE ENVELOPES

Network performances can be analysed from a variety of viewpoints. An important set of theoretical developments goes under the name of *Network Calculus* (e.g., [8]). In its deterministic form, this theory allows a synthetic description of a packet data stream by the so-called arrival curve, which gives an upper bound on the number of packets that flow through the monitored link (i.e., the "arrivals") over a given length of time. Although this may provide a suitable framework for analysis, it has to be remembered that, when measured traffic traces are considered, only an estimate of the theoretical arrival

curve can be obtained. Therefore, results are subjected to statistical variability. It is then more useful to recast the problem within the framework of stochastic network calculus [9].

Statistical arrival curves are also commonly called effective envelopes. A related function, the maximal rate envelope, provided the basis for the development of the statistical signal processing algorithm presented in this paper, which allows both the estimation of the parameters of a self-similar traffic model and the analysis of local features at the time scales where the LRD assumption may fail. It should be noted that the approach is closely related to the methods proposed in [10], [11] where, however, the main concern is measurement-based network admission control.

The literature broadly agrees on the fact that, at large time scales, traffic can be described by a mono-fractal self-similar behaviour. Furthermore, at these time scales a Gaussian approximation for traffic increments is roughly acceptable. The matter is actually somewhat subtler and has been discussed in some detail in [12] with regards to bit flows: basically, conditions for the Gaussian approximation to be acceptable depend on the level of aggregation, either horizontal (which results from the length of the time slot) or vertical (which is related to the number of traffic sources contributing to a flow). Therefore the attempt to identify some limiting time scale may be misleading, since link bandwidth and utilization conditions also affect and determine the "density" of the flow under analysis [13]. In the following it will at first be simply assumed that conditions for the application of an approximate Gaussian model hold. Under this assumption the statistical arrival curve (or effective envelope) can be calculated analytically for a given violation probability γ. In general terms, the arrival curve $\alpha_\gamma(n)$ is defined by the relationship:

$$\mathbf{P}[Z(n) \le \alpha_\gamma(n)] = 1-\gamma, \qquad (1)$$

where $Z(n)$ is a non-decreasing random process representing the total number of packets arrived in the interval [0, n]. It should be noticed that, throughout the paper, time will be represented by the discrete index n, reflecting the fact that a traffic trace is a time series of arrivals over finite-length time slots .

A comparatively simple but effective traffic model is a self-similar process with stationary increments and self-similarity parameter H (H-sssi) [5]. A H-sssi process $X(n)$ is called fractional Brownian motion (fBm) when the related increment process $y(m)=X(n+m)-X(n)$ is Gaussian; the latter is called fractional Gaussian noise (fGn) [5], [7]. Since fGn is a zero-mean process, whereas arrivals are always a non-negative quantity, a more realistic traffic model is obtained by adding to $y(m)$ a constant term equal to the mean traffic increment averaged over consecutive time slots, indicated by m_T. Then, the model for packet data traffic becomes the fBm process $X(n)$ with a linear component added.

$$Z(n) = X(n) + n \cdot m_T . \qquad (2)$$

Using the properties of self-similar processes and the hypothesis of Gaussian increments, it can be shown that the arrival curve for the process $Z(n)$ can be obtained from:

$$\Phi\left(\frac{\alpha_\gamma(n) - n \cdot m_T}{n^H \sigma_T} \right) = 1-\gamma , \qquad (3)$$

which yields:

$$\alpha_\gamma(n) = \sigma_T n^H \Phi^{-1}(1-\gamma) + n \cdot m_T , \qquad (4)$$

σ_T^2 being the variance of the zero-mean fGn increment process. The parameter H ($0.5 \le H \le 1$) is called the Hurst parameter and determines the degree of LRD of the fBm/fGn model.

The maximal rate envelope $R\gamma(n)$ represents the maximal arrival rate over a given interval, as a function of the interval length [10]. It is obtained from (4) by the ratio:

$$R_\gamma(n) = \frac{\alpha_\gamma(n)}{n} = \sigma_T n^{H-1} \Phi^{-1}(1-\gamma) + m_T . \qquad (5)$$

It should be noticed that, if the values of the mean m_T and variance σ_T^2 are estimated from traffic measurements, the Hurst parameter can be obtained by the following relationship:

$$H = 1 + \frac{1}{\log(n)} \log\left(\frac{R_\gamma(n) - m_T}{\sigma_T \Phi^{-1}(1-\gamma)} \right) . \qquad (6)$$

This approach will be used for the estimation of the Hurst parameter for real traffic traces, as shown in the following.

3. PROPOSED APPROACH

Let the analysed traffic trace be described by a sequence a_s, with $0 \le s \le S$, where a_s represents the number of packets arrived during the s-th time slot and S is the length of the series. The whole trace is divided into non-overlapping segments, which are then further divided into

K non-overlapping blocks, each containing M samples. For each block a maximal rate envelope is estimated, by the calculation of the maximal rates:

$$R_n = \frac{\max\limits_{n \le i \le M} \left[\sum\limits_{j=0}^{n-1} a_{i-j} \right]}{n} \qquad (7)$$

for all values of the interval lengths $n \le M$. It should be noticed that, for $n = M$, R_M is actually just an average. The estimates of the mean traffic increment m_T and the variance σ_T^2 are calculated over each segment.

Analysis of a segment provides a set of K estimates of the maximal rate envelope, obtained from measured data for finite increments of the interval length. This allows a statistical analysis, which implies considering the maximal rates R_n as random variables. The purposes of the analysis are twofold: on one hand, the Hurst parameter can be estimated from (6), providing a way to track its variation from one segment to another; on the other hand local features, such as heavy bursts or interruptions, can be detected by evidencing large differences in the maximal rate envelope estimates.

Equation (6) shows that the fit of the Hurst parameter estimate depends on the violation probability γ, for which a value has to be assumed. When dealing with experimental data, this probability is related to the confidence degree of the estimates. A probability γ means that the maximal rates R_n are assumed to estimate the maximal rate envelope with a confidence level of $(1-\gamma)\%$.

Fig. 1 refers to a simulated fBm process with mean value $m_T = 30$, variance $\sigma_T^2 = 20$ and Hurst parameter $H = 0.80$. In the figure, K=250 maximal rate envelopes are reported, each corresponding to a block of $M = 20$ samples and representing the maximum rates for each interval of duration n, with $n=1,\ldots,M$. Estimated values of R_n are largely superposed, although some trajectories may be discerned better than others. It may be noticed that R_1 only takes on integer values.

Points marked by a diamond in Fig. 1 are those of the experimental maximal envelope for the K estimated curves. Values represented by asterisks were instead calculated by formula (7), using the estimated values of mean, variance and Hurst parameter and with $\gamma = 0.01$. This value of γ is chosen so that the theoretical maximal rate envelope can include also little probable curves. It can be seen that the two plots are in good agreement. In this case the estimated value of H for a violation probability $\gamma = 0.01$ is equal to 0.799. It should be reminded, however that statistical variability should be taken into account to correctly assess the estimator accuracy.

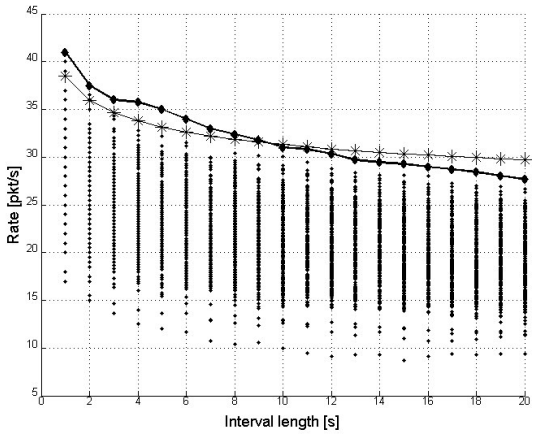

Fig. 1: experimental maximal rate envelopes for a simulated fBm traffic flow with $H = 0.80$.

Equation (6) can actually provide a sufficiently accurate estimate of the Hurst parameter. Fig. 2 shows the histogram of 100 estimated values of H, calculated, for a simulated fBm process with a theoretical Hurst parameter equal to 0.70, by applying (6) with $R_\gamma(n) = R_{20}$. Each value is obtained from the analysis of a segment of $K = 250$ blocks with $M = 20$ samples each. The violation probability γ is equal to 0.01. The resulting sample mean is 0.703 with a standard deviation of 0.032, although worst-case variations are higher.

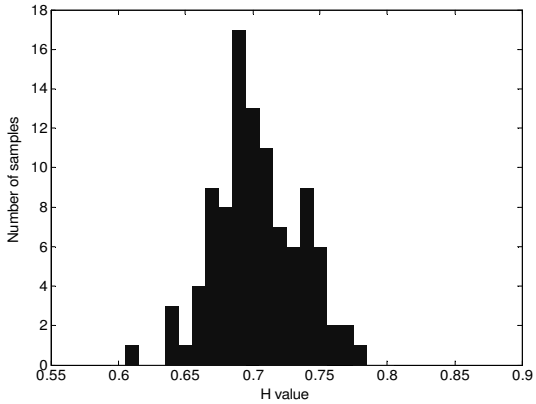

Fig. 2: histogram of 100 estimated values of the Hurst parameter H

The probability density function (pdf) of the maximal rate R_n can be estimated by a histogram of the K values of R_n calculated within a segment. Anomalies in a portion of the analysed traffic trace alter some values of the maximal rates in the affected segments: these appear as peaks or long tails in the corresponding histograms, which can be detected with remarkable sensitivity by looking at the skewness parameter of the pdf. This quantity is obtained from the second- and third-order central moments μ_2 and μ_3 as:

$$skew = \frac{\mu_3}{[\mu_2]^{3/2}} \ . \qquad (8)$$

Simulation analysis using the traffic model (2) showed that histograms of maximal rates are skewed also in regular traces; this is not unexpected since, by the theory of extreme values, maximal rates have a Gumbel limiting pdf [10]. However, typical phenomena that may occur within a trace at small time scales cause the skewness value to change significantly. In fact, if the traffic trace exhibits some non-stationarities, these appear as positive or negative peaks in the skewness values.

4. RESULTS

In this section results of the analysis of traces freely available from [14] will be presented. The traces were chosen because a complete analysis (by a wavelet-based approach) has already been carried out in [7], where several features have been carefully described. Therefore, a good reference is provided for the validation of the method proposed in this paper.

Results reported in Fig. 3 are related to a standard trace which is well modelled by fGn, as explained in [7]. At first, arrivals were aggregated over 1 s intervals, to represent a traffic trace measured over rather long time slots. In order to estimate the Hurst parameter a value of the percentile $\gamma = 0.01$ was again chosen. The estimated Hurst parameter is $H = 0.858$, which is in very good agreement with the value $H = 0.851$ obtained by the wavelet method in [7]. With a time slot length of 1 s, the traffic trace appears comparatively smooth and no significant local features are detected. This case can be regarded as a large time scale analysis, since the interval corresponding to a trace segment is 5000 s long and the regularizing effect of aggregation can clearly be felt.

The analysis was repeated on the same trace using the original time slot size of 1ms. Results, presented in Fig. 4, were also obtained with $M = 20$ and $K = 250$, which in this case correspond to a segment length of 5 s. It should be noticed that the segment length $M \cdot K$ represents the minimum interval length over which the traffic trace must be assumed to be stationary. On the other hand, the trace can be analysed for the detection of local features over much shorter intervals; in fact, phenomena as short as M time slots (20 ms for the experimental results reported here) are detectable if the associated maximal rates differ widely enough from regular behaviour. Generally, the suitability of the choice of M and K depends on several factors, including the nature of the analysed traffic. Criteria for optimal choice are still an open research issue.

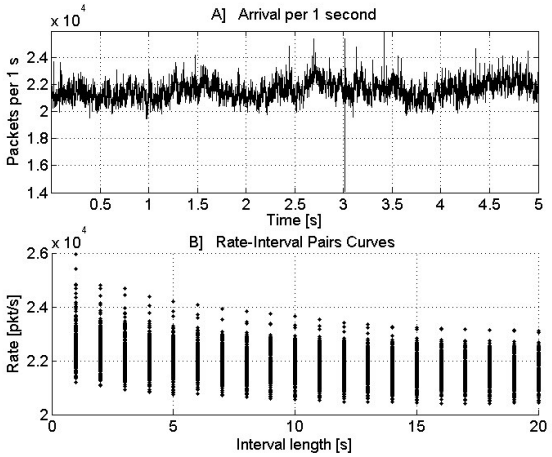

Fig. 3: A) time series of the number of packets arriving on a link in 1 s time intervals; B) maximal rate-interval curves calculated for M=20 and K=250.

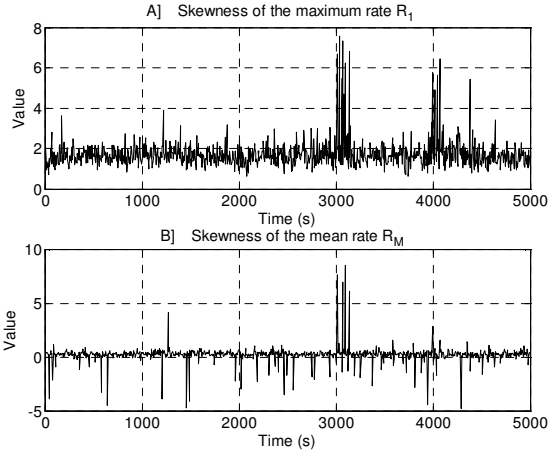

Fig. 4: A) skewness trace of the maximum rate R_1; B) skewness trace of the maximum rate R_{20}.

At the time scales of Fig. 4 both the skewness coefficient of the maximum rate (that is, R_1) and the skewness coefficient of the mean rate (that is, R_{20}), exhibit some positive and negative peaks. Observing Fig. 4, a large number of skewness peaks can be found around 3000 seconds from the origin, while the negative peaks in Fig. 4.b reveal the presence of a number of short interruptions in the traffic trace, as will be discussed later.

Analysis of a number of traffic traces suggested a relationship between some definite skewness peak configuration and the corresponding local feature of the trace. In particular, when the traffic segment presents an abrupt interruption of flow, the skewness value of the corresponding mean rate R_{20} exhibits a negative peak. On the other hand, a packet burst causes the skewness coefficient of the maximum rate R_1 to assume a large positive value within the affected segment.

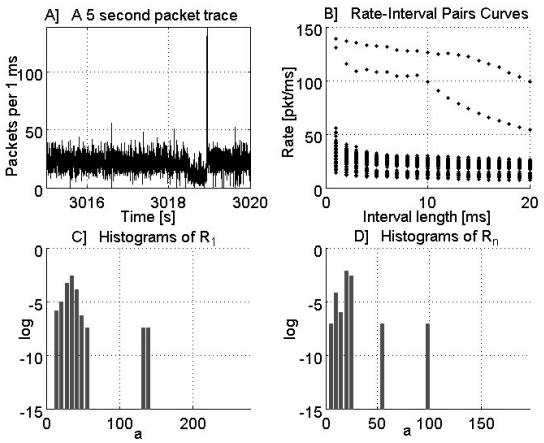

Fig. 5: A) time series of the number of packets arriving on a link in 1 ms time intervals; B) associated RIP curves; C) histogram of R_1 on a log-probability scale; D) histogram of R_{20} on a log-probability scale.

If there are positive peaks both for the mean and maximum rate skewness, then the presence of a droop in the local mean value of the traffic followed by a burst could be noted. In Fig. 5 analysis of this phenomenon is reported. The trends of maximal rate envelopes in Fig. 5.b show that some envelope trajectories differ widely from the others: these give rise to peaks on the rightmost part of both maximum rate and mean rate histograms (Fig. 5.c and Fig. 5.d). The presence of these peaks causes the skewness coefficients to assume very high positive values.

It should be noted that the plots presented in Figs. 5 and 6 refer to two different segments within the same trace, where skewness peaks have been found. In both cases the upper part of the figure shows a detailed view of the most interesting part of the 5 s trace segment.

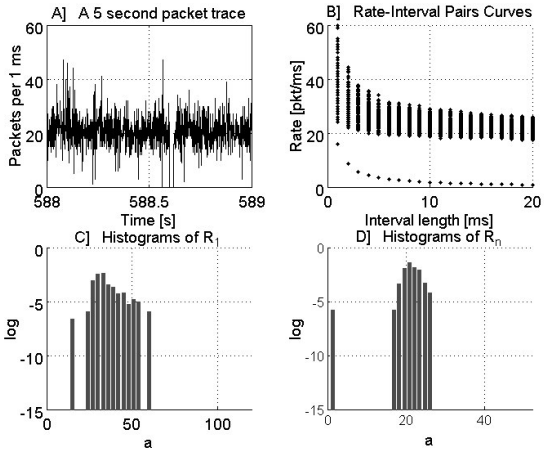

Fig. 6: same meaning as Fig. 5, but in this case a flow interruption has been located.

The analysis reported in Fig. 6 refers to a segment which includes an abrupt interruption. Also for this case, it can be seen that the phenomenon influences the maximal rate envelopes. In particular, the negative value of the mean rate skewness is caused by the lower trajectory of the rate envelopes in Fig 6.b. It should be remembered that the detected flow interruption only lasts a few tens of ms, which evidences the sensitivity of this kind of analysis to even short phenomena, as long as the change in trace features is detectable. It should also be noticed that, if the segments where skewness peaks occur are disregarded, the average value of the Hurst parameter is $H = 0.574$, which is again in good agreement with the plot reported in [7].

5. CONCLUSIONS

The proposed algorithm has been tested on a number of traces and provides a simple and fast method for the analysis of network traffic. Although it was initially conceived as a simple mean to detect anomalies in traffic traces, it has proved able to provide acceptably accurate estimates of the Hurst parameter for a self-similar traffic model. Furthermore, although the estimator is derived from an assumed fBm model, the resulting values generally agree with those obtained by wavelet-based methods, which do not rely on such modelling assumption.

Local features, which might be interpreted as non-stationary phenomena, are associated to maximal rate trajectories that are clearly distinguished from the common behaviour and can be easily discarded for the purpose of Hurst parameter estimation. This suggests that the approach can be made robust with respect to anomalies in the observed traffic. Detection capabilities of the proposed method are presently being further investigated with a view to on-line monitoring applications.

6. REFERENCES

[1] [W.E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, On the Self-Similar Nature of Ethernet Traffic (Extended Version), IEEE/ACM Trans. On Networking, vol. 2, no. 1, Feb. 1994, pp. 1-15.

[2] O. Cappé, E. Moulines, J.C. Pesquet, A. Petropulu, X. Yang, Long-Range Dependence and Heavy-Tail Modeling for Teletraffic Data, IEEE Signal Processing Magazine, vol. 19, no. 3, May 2002, pp. 14-27.

[3] P. Abry, D. Veitch, Wavelet Analysis of Long-Range-Dependent Traffic, IEEE Trans. On Information Theory, vol. 44, no. 1, Jan. 1998, pp. 2-15.

[4] D. Veitch, P. Abry, A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence, IEEE Trans. On Information Theory, vol. 45, no. 3, Apr. 1999, pp. 878-897.

[5] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, D. Veitch, Multiscale Nature of Network Traffic, IEEE Signal Processing Magazine, vol. 19, no. 3, May 2002, pp. 28-46.

[6] S. Uhlig, Non-stationarity and high-order scaling in TCP flow arrivals: a methodological analysis, ACM SIGCOMM Computer Communications Review, vol. 34, no. 2, Apr. 2004, pp. 9-24.

[7] S. Stoev, M.S. Taqqu, C. Park, J.S. Marron, Strenghts and Limitations of the Wavelet Spectrum Method in the Analysis of Internet Traffic, SAMSI Technical Report 2004-8, 26 March 2004.

[8] J.Y. Le Boudec, P. Thiran, "Network Calculus - A Theory of Deterministic Queuing Systems for the Internet", Springer, Berlin, 2001.

[9] C. Li, A. Burchard, J. Liebeherr, A Network Calculus with Effective Bandwidth, Technical Report, Univ. of Virginia, CS-2003-20, Nov. 2003. Submitted to IEEE/ACM Trans. on Networking.

[10] J. Qiu, E.W. Knightly, Measurement-Based Admission Control with Aggregate Traffic Envelopes, IEEE/ACM Trans. on Networking, vol. 9, no. 2, Apr. 2001, pp. 199-210.

[11] E.W. Knightly, H. Zhang, D-BIND: An Accurate Traffic Model for Providing QoS Guarantees to VBR Traffic, IEEE/ACM Trans. On Networking, vol. 5, no. 2, Apr. 1997, pp. 219-231

[12] J. Kilpi, I. Norros, Testing the Gaussian approximation of aggregate traffic, Proc. Internet Measurement Workshop, 6-8 November 2002, Marseilles, France.

[13] Z. Zhang, V.J. Ribeiro, S. Moon, C. Diot, Smal— Time Scaling Behaviors of Internet Backbone Traffic: An Empirical Study, Proc. IEEE INFOCOM 2003, 30 March – 3 April 2003, San Francisco, CA, USA, vol. 3, pp. 1826-1836.

[14] http://www-dirt.cs.unc.edu/ts/