Scaling Analysis of Wavelet Quantiles in Network Traffic

Giada Giorgi and Claudio Narduzzi

University of Padova, Dept. of Information Engineering,
via Gradenigo 6/B, I-35100 Padova, Italy

Abstract. The study of network traffic by flow analysis has been the subject of intense and varied research. Wavelet transforms, which form the core of most traffic analysis tools, are known to be robust to linear trends in data measurements, but may suffer from the presence of occasional non-stationarities.

This paper considers how the information associated to quantiles of wavelet coefficients can be exploited to improve the understanding of traffic features. A tool based on these principles is introduced and results of its application to analysis of traffic traces are presented.

1 Introduction

Statistical traffic analysis refers to the general properties of network traffic, aiming to describe them by suitable flow models. Traffic in packet networks has been the subject of intense and varied research, leading to progressive refinements of models and analysis tools.

When the statistical features of flow intensity in a traffic trace are analyzed, it can be seen that anomalies, associated to local changes in the distribution of traffic, frequently affect the tails of the empirical probability density function (pdf). Effects of a similar nature may also arise when a highly composite traffic trace is considered, in which case distribution changes may be attributed to the varying mix of contributions from flows having different statistical properties. These issues are directly related to the assumed traffic model: in a number of cases of practical interest, forcing a single-flow LRD random process model on measured data does not appear to suit the actual situation entirely [1], [2].

The well-known *Abry-Veitch* (A-V) wavelet-based tool has become a standard reference for most traffic analysis methods [3]. However, analysis of real traffic traces showed that, in the cases mentioned above, the tool may not provide meaningful measurements of the Hurst scaling exponent [4] and of other parameters. A reason why the A-V tool is not ideally suited to deal with these kinds of phenomena, is that it refers to a cumulative quantity, i.e., the energy of wavelet coefficients. From a statistical viewpoint this emphasises variance, which is sensitive to changes in empirical pdf's but does not allow a more detailed understanding of phenomena.

This paper will show that quantile analysis of wavelet coefficients, on the contrary, can provide very robust and acceptably accurate estimates of the Hurst

M. Papadopouli, P. Owezarski, and A. Pras (Eds.): TMA 2009, LNCS 5537, pp. 109–116, 2009.

parameter value, even in the presence of non-stationary disturbances in traffic time series. The probability level of quantiles represents an additional parameter, that can be tuned for the purposes of the analysis. Comparison between curves obtained for different confidence levels may provide additional information on the features of the analysed traffic.

2 Scaling and Wavelets

The proposed approach merges concepts from quantile analysis with the wavelet multiresolution approach, whose main features are briefly recalled in this Section.

Let $X(k)$ be a time series obtained by counting the number of packets (or bytes) flowing through a link during consecutive, non-overlapping time slots of duration T. Packet counts can be aggregated over larger time scales. Considering time intervals of progressively longer duration $2^j \cdot T$, the time series:

$$X^{(j)}(k) = \frac{1}{2^j} \sum_{i=0}^{2^j-1} X(k \cdot 2^j + i) \tag{1}$$

represents the aggregate version of the time series $X(k)$ at scale j. Under the hypothesis of self-similarity for $X(k)$, the following relationship can be found:

$$X^{(j)}(k) \stackrel{d}{=} 2^{j(H-1)} X(k), \tag{2}$$

where $\stackrel{d}{=}$ denotes equality of probability distributions and H is the Hurst exponent. It is well known, e.g., from the early pioneering studies presented in [5], that the correlation structure of the time series $X(k)$ can be assumed to decrease with a power law as the lag number increases. This statistical property is called *long-range dependence* (LRD). The Hurst parameter H quantifies the asymptotic self-similar scaling as well as the degree of long-range dependence. Under the common assumption that the underlying random process is fractional with stationary increments, H varies between 0.5 and 1, denoting respectively a non-correlated and a completely correlated time series.

For a self-similar process a scaling relationship among wavelet coefficients exists [6] and has the same form for both approximation coefficients $a_x(j,k)$ and detail coefficients $d_x(j,k)$. Using the symbol $c_x(j,k)$ to generically indicate either of the two set of coefficients, it can be given in the form:

$$c_x(j,k) \stackrel{d}{=} 2^{j(H+\frac{1}{2})} c_x(0,k), \tag{3}$$

where $\stackrel{d}{=}$ denotes equality of probability distributions. It should be remembered that, if the definition of aggregate process given in (1) is referred to, the relationship must be normalized by the number of samples considered in the summation, yielding:

$$c_x(j,k) \stackrel{d}{=} 2^{j(H-\frac{1}{2})} c_x(0,k). \tag{4}$$

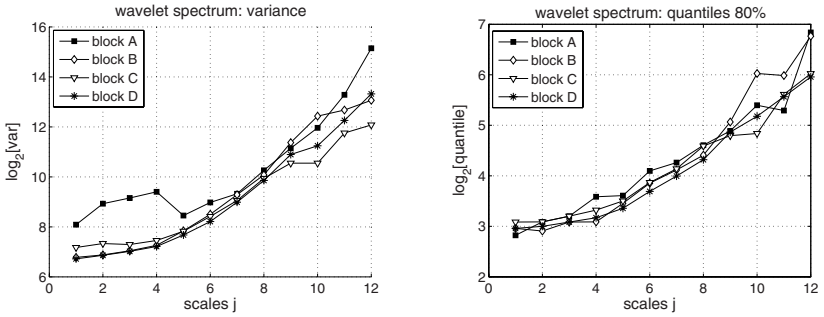
Recursive algorithms are initialized with $c_x(0,k) = X(k)$.

The *Abry-Veitch* estimator considers the energy of detail coefficients $d_x(j,k)$ at different time scales. This follows the scaling law:

$$\mathbb{E}\left[d_x(j,k)^2\right] = 2^{j(2H-1)}\mathbb{E}\left[d_x(0,k)^2\right], \tag{5}$$

which provides a means to identify the presence of long range dependence in data measurements and estimate the corresponding scaling exponent H. It can be noted that, since the mean of detail coefficients is zero: $\mathbb{E}\left[d_x(j,k)\right] = 0$, the energy (5) corresponds to the coefficient variance.

The tool has been largely used to identify the presence of scaling in data measurements and to estimate the value of the scaling exponent by a linear regression on the log-log wavelet spectrum diagram. Since the detail coefficients are uncorrelated, its variance is a function of the amount of data considered and does not depend on the unknown, actual value of the Hurst coefficient H. This very important property allows to improve estimation accuracy by increasing the number of samples and is one of the reasons for the success of the tool.



(a) A-V tool. (b) Quantile-based.

Fig. 1. Wavelet spectrum over consecutive non-overlapping blocks

3 A-V Analysis of a Non-stationary Trace

The A-V estimator is known to be robust to linear trends in data measurements, but may suffer from the presence of occasional non-stationarities. An example is provided by the following analysis of the AUCK [7] traffic trace captured on 06 April 2001, which presents a strong, localised non-stationarity. The raw traffic trace was initially aggregated over time intervals of duration $T = 50\ ms$. Analysis is restricted to measurements taken during the day working hours, by considering only the samples between the $(6.5E+05)$-th and the $(11.5E+05)$-th. This allows to disregard longer-term fluctuations of traffic on a daily scale. The discrete wavelet transform was applied over four non-overlapping blocks of $125,000$ samples each (roughly a two-hour lenght); the wavelet spectra obtained in each block are plotted together in Fig. 1(a). It can be seen that, at lower time

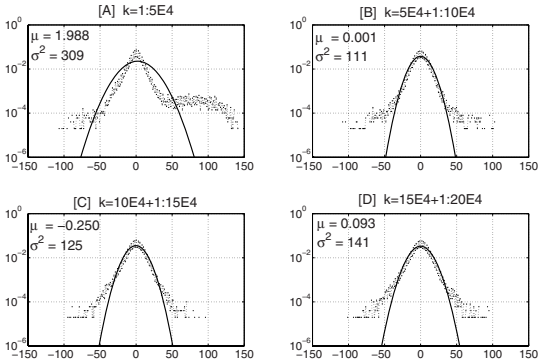


Fig. 2. Histograms of the wavelet detail coefficient $d_x(j,k)$ calculated over consecutive non-overlapping blocks. The analysis refers to the AUCKIV trace of the 06 April 2001.

scales, the curve related to *block A*, which entirely contains the non-stationarity, presents a strong discrepancy from the others.

Recall that the Hurst parameter characterizes the dependence of the traffic only a large scales. However the wavelet spectrum provides additional useful information about the dependence in the data also on small time scales. In this case, where an alignment can be found at the lowest scales [3], that is from $j_1 = 1$ to an upper bound j_2, the scaling indicates the fractal nature of the traffic.

To understand the influence of this local flow irregularity on wavelet spectra, the time series of the lowest-scale detail coefficients $d_x(1,k)$ have been considered for the same four blocks. Their histograms are presented in Fig. 2, where they are compared with Gaussian distributions having the same mean and variance. *Block A* is characterised by an asymmetric histogram with a much heavier tail for positive values of detail coefficients; the estimated variance is accordingly larger than in the other blocks.

As can be noted in Fig. 1(a), the non-stationarity affects the time series over time scales in the range between $j_1 = 1$ and $j_2 = 4$. The wavelet spectrum obtained by the A-V tool represents, over these scales, the behaviour of the non-stationarity and not that of the main process.

Similar effects are known and have been noted in a number of works, e.g., [8]. The consequences are that scaling analysis becomes harder, since alignments in a log energy-scale diagram are more difficult to find.

The analysis of quantiles provides additional information about the distribution of detail coefficients. Estimated quantile values for the four blocks of the AUCK traffic trace show that the local features in *block A* only affect quantiles associated with probability levels $\geq 99\%$. Lower probability levels are not affected by the presence of disturbances in the traffic time series.

It is important to investigate how this additional knowledge could be interpreted correctly. In this example, analysis of quantiles referring to a probability

level $< 99\%$ could provide more accurate scaling information. On the other hand, quantiles with higher probability levels might convey information about local features.

4 Quantile-Based Estimation

Let $r_\gamma(j)$ be the (1-γ)-quantile of coefficients at scale j. It provides a bound on the value that the samples of $c_x(j,k)$ can assume, which can be exceeded with a probability γ, called *violation probability*:

$$P\left[c_x(j,k) \leq r_\gamma(j)\right] = 1 - \gamma. \tag{6}$$

The self-similarity relationship between $c_x(0,k)$ and $c_x(j,k)$ extends to their quantiles, providing the following expression that links quantiles at different scales:

$$r_\gamma(j) - \mathbb{E}[c_x(j,k)] = 2^{j\left(H-\frac{1}{2}\right)}\left[r_\gamma(0) - \mathbb{E}[c_x(j,k)]\right]. \tag{7}$$

It should be remembered that for detail coefficients, i.e., when $c_x(j,k) = d_x(j,k)$, the mean value is null. In this case the scaling relationship between quantiles can be obtained in a straightforward manner by substituting (4) in (6). It results in:

$$P\left[2^{j(H-\frac{1}{2})} d_x(0,k) \leq r_\gamma(j)\right] = P\left[d_x(0,k) \leq r_\gamma(j) \cdot 2^{-j(H-\frac{1}{2})}\right] = 1 - \gamma. \tag{8}$$

where $P\left[d_x(0,k) \leq r_\gamma(0)\right] = 1 - \gamma$ for definition. This provides the expression (7) where $\mathbb{E}[d_x(j,k)] = 0$.

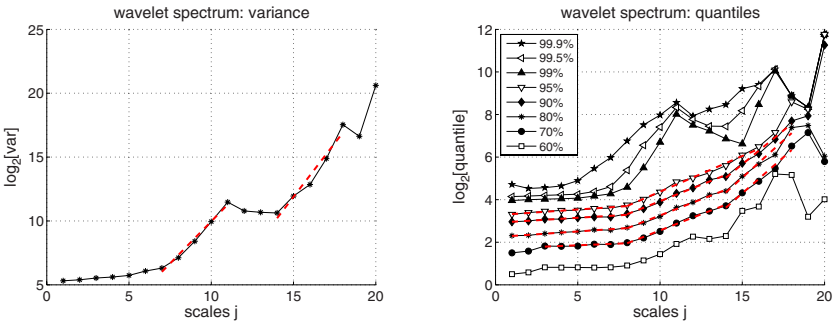
Rewriting expression (7) in a log-log scale shows that the scaling exponent can be obtained by a simple process.

Graphically, a plot of log-quantile versus scale is obtained; borrowing from [8], this will be called a *quantile-based wavelet spectrum*. A linear regression of this plot then yields the scaling exponent, from which an estimate of the Hurst parameter H follows immediately.

For the AUCK trace considered in Sec. 3, the quantile-based wavelet spectra have been plotted in Fig. 1(b), with a probability level $(1-\gamma) = 80\%$. The same partitioning scheme of Fig. 1(a) has been adopted. It can be noted that the quantile-based spectrum related to *block A* is very similar to the curves obtained from the other blocks. In fact, the non-stationarity located within that block does not affect quantile estimates at the 80% level of probability. As a consequence, variability in Hurst parameter estimation is much reduced.

To gain a better understanding of the potentiality of a quantile approach, it was tested on a large amount of traffic traces. In the following we will report the results obtained for one of the traffic traces collected by the DIRT research group at the University of North Carolina (UNC). These traffic traces are particularly useful because they have been thoroughly analyzed, identifying and localizing a number of features that made correct estimation of the Hurst parameter by the A-V tool quite difficult. Therefore, we employed them to test the effectiveness of the proposed approach.

The considered trace was captured on 09 April 2002; it has been aggregated over time intervals of $T = 1ms$. It presents a burst of about $300-400$ seconds duration. This burst gives rise to a strong non-stationarity that affects the medium time scales, as can be noted from the variance-based wavelet spectrum of detail coefficients in Fig. 3(a). In this case no alignment can be found, resulting in very poor estimates for H.



(a) A-V tool. (b) Quantile-based.

Fig. 3. Wavelet spectrum for the UNC02 trace, captured 09 Apr. 2002 from 19:30 to 21:30

The corresponding quantile-based wavelet spectrum for the same trace is plotted in Fig. 3(b), where the curves obtained for different probability levels are shown. If probability levels $\leq 95\%$ are considered, curves are not affected by the presence of the non-stationarity, therefore alignments can be found for certain scale ranges, as illustrated by dotted lines. At those time scales the scaling exponent can be correctly estimated. Interestingly, quantile-based wavelet spectra show the same familiar two-slope behavior that generally characterizes most traffic traces.

For lower probability levels, like 60%, the quantile spectrum presents a greater variability. To explain this matter, the uncertainty associated to the estimates of quantile must be taken into account. For a random process, having a probability density function (pdf) $f(\cdot)$, the estimation variance of the theoretical $(1-\gamma)-$quantile is:

$$\sigma^2_{r_\gamma(j)} = \frac{\gamma(1-\gamma)}{N_j \cdot f^2(r_\gamma(j))}. \tag{9}$$

where N_j is the number of samples considered for estimating the quantile.

Quantile properties therefore depend on the probability distribution of the process [9]. The uncertainty associated to quantile estimates presents a maximum for $\gamma = 50\%$ and minimum values for $\gamma = 0\%$ and $\gamma = 100\%$. For the purposes of uncertainty analysis, the distribution of measurement data can be approximated by a Gaussian process. It is important to remember the limits of

this idealization. In the case of actual processes, where the tails of the distribution are generally limited by some physical constraint, the Gaussian hypothesis no longer holds for values of γ close to 0 or to 1. This discrepancy can be overcome by considering values of γ for which the Gaussian hypothesis holds true, at least as an approximation. Analysis of experimental data by normal probability plots can help find suitable limiting values [10].

This explain the greater variability at lower probability levels as well as at higher levels, as can be noted in Fig. 3(b).

5 Conclusions

The study of quantiles is better suited to deal with the heavy-tail phenomena that characterize network traffic. Its application in quantile-based wavelet spectra, that can be referred to both detail and approximation coefficients of a wavelet transform is, to the authors' knowledge, a novel idea that appears quite promising. It is important, however, to approach the method with a degree of caution.

Results shown in this paper suggest that the accuracy of Hurst parameter estimates can be improved by tuning the choice of quantile probability level. It should be realised that, in so doing, an experimenter is deliberately discarding information contained in the heavy-tails. This choice has a considerable impact in determining what is actually being modelled in a traffic flow. For instance, if traffic irregularities are related to local phenomena, the network flow could be described by a "mainstream" process, whose statistical properties may be altered by occasional outliers. If the "contamination" is not self-similar, its presence would only be evident at well-defined time scales of influence, while for larger time scales it is smoothed out by aggregation. In a similar case, information obtained by considering wavelet spectra for higher probability quantiles and by tracking their evolution with time would be just as valuable.

In general, traffic analysis can present difficulties when complex and heterogeneous flows are considered. Then, a different modelling paradigm can be considered by decomposing the flow in the monitored link into a superposition of stochastic processes, each having its own specific correlation structure. In this case analysis of wavelet quantiles would provide a more detailed picture of traffic features and might prove to be a more flexible tool.

References

1. Sarvotham, S., Riedi, R., Baraniuk, R.: Network and user driven alpha-beta on-off source model for network traffic. Computer Networks 48(3), 335–350 (2005)
2. Giorgi, G., Narduzzi, C.: A study of measurement-based traffic models for network diagnostics. In: Proc. IEEE Instrum. Meas. Tech. Conf. IMTC 2007, Warsaw, Poland, May 01-03 (2007)
3. Abry, P., Taqqu, M.S., Veitch, D.: Wavelets for the analysis, estimation and synthesis of scaling data. In: Park, K., Willinger, W. (eds.) Self Similar Traffic Analysis and Performance Evaluation. Wiley, Chichester (2000)

4. Giorgi, G., Narduzzi, C.: Rate-interval curves: A tool for the analysis and monitoring of network traffic. Performance Evaluation 65(6-7), 441–462 (2008)
5. Leland, W., Taqqu, M., Willinger, W., Wilson, D.: On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Trans. on Information Theory 2(1), 1–15 (1994)
6. Pesquet-Popescu, B.: Statistical properties of the wavelet decomposition of certain non-gaussian self-similar processes. Signal Processing 75, 303–322 (1999)
7. National Laboratory for Applied Network Reasearch, U, `http://mna.nlanr.net`
8. Stoev, S., Taqqu, M., Marron, J.: On the wavelet spectrum diagnostic for hurst parameter estimation in the analysis of internet traffic. Computer Networks 48(3), 423–445 (2005)
9. Ivchenko, G., Medvedev, Y.: Mathematical Statistics. Mir, Moscow, Russia (1990)
10. Giorgi, G., Narduzzi, C.: Uncertainty of quantiles estimates in the measurement of self-similar processes. In: Proc. of inter. Workshop on Advanced Methods for Uncertainty Estimation in Measurement, AMUEM 2008, Sardagna, Trento, Italy, July 21-22 (2008)