Glyndŵr University
Glyndŵr University Research Online

Computing   Computer Science

10-1-2007

Identifying Long-range Dependent Network Traffic through Autocorrelation Functions

Karim Mohammed Rezaul

Vic Grout
Glyndwr University, v.grout@glyndwr.ac.uk

Follow this and additional works at: http://epubs.glyndwr.ac.uk/cair

Part of the Computer and Systems Architecture Commons, Digital Communications and Networking Commons, Hardware Systems Commons, and the Systems and Communications Commons

Identifying Long-range Dependent Network Traffic through Autocorrelation Functions

Abstract

For over a decade researchers have been reporting the impact of self-similar long-range dependent network traffic. Long-range dependence (LRD) is of great significance in traffic engineering problems such as measurement, queuing strategy, buffer sizing and admission and congestion control. In this research, in order to determine the existence of LRD, we apply three different robust versions of the autocorrelation function (ACF), namely weighted ACF (WACF), trimmed ACF (TACF) and variance-ratio of differences and sums, known as the D/S variance estimator (DACF), in conjunction with the sample ACF (which is moment based). Here we define the moment based ACF as MACF. In telecommunications, LRD traffic defines that a similar pattern of traffic persists for a longer span of time. Through ACF, it is possible to detect how long the traffic lasts. The aim of this research is to investigate the performance of ACF in identifying the existence of LRD traffic.

Disciplines

Computer and Systems Architecture | Digital Communications and Networking | Hardware Systems | Systems and Communications

Identifying Long-range Dependent Network Traffic through Autocorrelation Functions

Karim Mohammed Rezaul and Vic Grout

Centre for Applied Internet Research (CAIR), University of Wales, NEWI, Wrexham, UK

{karim\vic}@cair-uk.org

Abstract- **For over a decade researchers have been reporting the impact of self-similar long-range dependent network traffic. Long-range dependence (LRD) is of great significance in traffic engineering problems such as measurement, queuing strategy, buffer sizing and admission and congestion control. In this research, in order to determine the existence of LRD, we apply three different robust versions of the autocorrelation function (ACF), namely weighted ACF (WACF), trimmed ACF (TACF) and variance-ratio of differences and sums, known as the D/S variance estimator (DACF), in conjunction with the sample ACF (which is moment based). Here we define the moment based ACF as MACF. In telecommunications, LRD traffic defines that a similar pattern of traffic persists for a longer span of time. Through ACF, it is possible to detect how long the traffic lasts. The aim of this research is to investigate the performance of ACF in identifying the existence of LRD traffic.**

I. INTRODUCTION

Self-similar and *long-range dependent* (*LRD*) characteristics of Internet traffic have attracted the attention of researchers for over a decade. The condition of self-similarity is that the *autocorrelation function* (*ACF*) of the time-series declines as a power-law, leading to positive correlations among widely separated observations [1]. Note that a power law is a characteristic feature of heavy-tailed distributions. In the Internet, heavy-tailed distributions have been observed in the context of traffic characterization. The distributions having infinite variances are called heavy-tailed and the weight of their tails is determined by a parameter called the tail index, $\alpha < 2$ [2].

Self-similarity and heavy-tailedness are of great importance for network capacity planning purposes, in which researchers are interested in developing analytical methods for analysing traffic characteristics. The goal of traffic characterisation is to determine the nature of the traffic and develop tractable models that capture the important properties of data, which can eventually lead to accurate performance prediction. The uses of traffic characterisation include network planning, design, capacity management, performance prediction, real-time traffic management and network control.

The LRD property of traffic fluctuations has important implications on the performance, design and dimensioning of the network [3]. A simple, direct parameter, characterizing the degree of LRD, is the Hurst parameter. The Hurst exponent (or Hurst parameter, H), which more than a half-century ago was proposed for the analysis of reservoir long-term storage capacity [4], is used today to measure the intensity of LRD in network traffic. A number of methods have been proposed to estimate the Hurst parameter. Some of the most popular include aggregated variance time (V/T) [5], Rescaled-range (R/S) [3, 4] and the

Higuchi and wavelet-based methods [6, 7] although there are many others. In [8, 9] an estimator is introduced called the Hurst Exponent from the Autocorrelation Function (HEAF) and it is shown that (and why) lag 2 in HEAF (i.e. HEAF (2)) is considered when estimating LRD of network traffic. [10] considers the robustness of HEAF(2) when estimating the Hurst parameter of data traffic (e.g. packet sequences) with outliers and also the reliability of HEAF(2).

The remainder of this paper is organised as follows: Section II describes the relationship between self-similarity, long-range dependence and the autocorrelation function (ACF). Section III discusses different methods of autocorrelation function. Finally, conclusions are drawn in section IV.

II. THE RELATIONSHIP BETWEEN SELF-SIMILARITY, LRD AND ACF

A phenomenon that is self-similar looks or behaves the same when viewed at different degrees of magnification or different scales on a given dimension and is bursty over all time scales. Self-similarity is the property of a series of data points to retain a pattern or appearance regardless of the level of granularity used and is the result of LRD in the data series. If a self-similar process is bursty on a wide range of timescales, it may exhibit LRD. Often lagged autocorrelations are used in time series analysis for empirical stationary tests. Self-similarity manifests itself in the time series of arrivals' LRD (i.e., long memory). The evidence of very slow, linear decay in the sample lag ACF indicates nonstationary behaviour [11]. Long-range-dependence means that all the values at any time are correlated in a positive and non-negligible way with values at all future instants.

A continuous time process $Y = \{Y(t), t \geq 0\}$ is self-similar if it satisfies the following condition [12]:

$$Y(t) \overset{d}{=} a^{-H} Y(at), \quad \forall a > 0, \quad and \quad 0 < H < 1 \qquad (1)$$

where H is the index of self-similarity (the Hurst parameter) and the equality is in the sense of finite-dimensional distributions.

The stationary process X is said to be an LRD process if its ACF is non-summable [13], that is if $\sum_{k=-\infty}^{\infty} \rho_k = \infty$

Details of how the ACF decays with k are of interest because the behaviour of the tail of the ACF completely determines its summability. From [3], X is said to exhibit long-range dependence if

$$\rho_k \sim L(t) k^{-(2-2H)}, \quad as \; k \to \infty \qquad (2)$$

where $\dfrac{1}{2} < H < 1$ and $L(.)$ slowly varies at infinity, i.e.,

$$\lim_{t\to\infty} \frac{L(xt)}{L(t)} = 1, \quad \text{for all } x > 0.$$

Equation (2) implies that the LRD is characterized by an ACF that decays hyperbolically rather than exponentially fast.
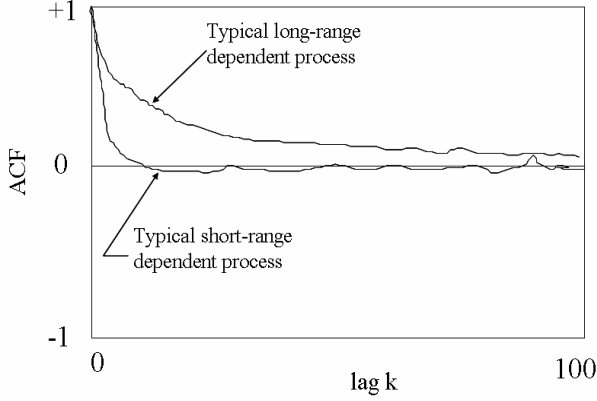

Fig. 1. Autocorrelation plot for LRD and SRD processes

LRD processes then are characterised by a slowly decaying covariance function that is non-summable. When network performance is affected by LRD, data are correlated over an unlimited range of time lags and this property results in a scale invariance phenomenon. Then no characteristic time scale can be identified in the process: they are all equivalent for describing its statistics - the part resembles the whole and vice versa. Figure 1 illustrates typical LRD and SRD (short-range dependent) processes.

III. DIFFERENT METHODS OF AUTOCORRELATION FUNCTION

A. Sample autocorrelation function (MACF)

As in [14], for given observed data X_i (i.e. $X_1,.........,X_n$), the sample autocorrelation function can be calculated by the following method:

Let $\hat{\mu}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (3.1)

and $\hat{\gamma}_n(k) = \dfrac{1}{n}\sum_{i=1}^{n-k}\left(X_i - \hat{\mu}_n\right)\left(X_{i+k} - \hat{\mu}_n\right)$,   (3.2)

where $k=0,1, 2, \ldots.., n$,

with $\hat{\sigma}_n^2 = \hat{\gamma}_n(0)$.   (3.3)

Then the sample autocorrelations of lag k are given by

$\hat{\rho}_k = \dfrac{\hat{\gamma}_n(k)}{\hat{\sigma}_n^2}$ (3.4)

(Equations (3.1), (3.2), (3.3) and (3.4) denote the sample mean, the sample covariance, the sample variance and the sample autocorrelation, respectively). As this ACF is moment based, for simplicity we name it as MACF.

B. Weighted Autocorrelation function (WACF)

For a given time series $z_1, \ldots, z_n$, let $w_1, \ldots, w_n$ be the corresponding weighting series. Then the weighted sample autocorrelation function (shorten as WACF) can be defined [15] as

$$\hat{\rho}_w(k) = \frac{\hat{\gamma}_w(k)}{\hat{\gamma}_w(0)}$$

where

$$\hat{\gamma}_w(k) = \frac{\sum_{t=k+1}^{n}(z_{t-k} - \bar{z}_w)(z_t - \bar{z}_w)w_{t-k}\,w_t}{\sum_{t=k+1}^{n} w_{t-k}\,w_t}$$

and $\quad \bar{z}_w = \sum_{t=1}^{n} z_t w_t \Big/ \sum_{t=1}^{n} w_t$

Here w_t is considered as Hampel's [15] weighted function.

$$w_t = \begin{cases} 1 & t < a, \\ a/t & a \le t < b, \\ a(c-t)/t(c-b) & b \le t < c, \\ 0 & t \ge c. \end{cases}$$

where a =1.7, b = 3.4 and c = 8.5

As the network traffic data (sample) can be larger than 8.5, we modify the Hampel's weighted function as in the following:

$$w_t = \begin{cases} 1 & t < a, \\ a/t & a \le t < b, \\ a(c-t)/t(c-b) & b \le t < c, \\ 1 & t \ge c. \end{cases}$$

C. Trimmed Autocorrelation function (TACF)

Let $\quad z_{(1)} \le z_{(2)} \le .... \le z_{(n)}$ be the ordered observations of the given time series $z_1, z_2, \ldots,$ and z_n. Chan and Wei [16] introduced the α-trimmed sample autocorrelation function (shorten as TACF) defined by

$$\hat{\rho}_T(k) = \frac{\hat{\gamma}_T(k)}{\hat{\gamma}_T(0)}$$

where

$$\hat{\gamma}_T(k) = \frac{\sum_{t=k+1}^{n}(z_{t-k} - \bar{z}^{(\alpha)})(z_t - \bar{z}^{(\alpha)})L_{t-k}^{(\alpha)}\,L_t^{(\alpha)}}{\sum_{t=k+1}^{n} L_{t-k}^{(\alpha)}\,L_t^{(\alpha)}}$$

$$\bar{z}^{(\alpha)} = \sum_{t=1}^{n} z_t L_t^{(\alpha)} \Big/ \sum_{t=1}^{n} L_t^{(\alpha)} \quad \text{and}$$

$$L_t^{(\alpha)} = \begin{cases} 0, & \text{if } z_t \le z_{(g)} \text{ or } z_t \ge z_{(n-g+1)} \\ 1, & \text{otherwise} \end{cases}$$

where $g = [\alpha n]$ is the integer part of αn and $0 \le \alpha \le 0.05$. Chan and Wei showed that TACF is, in general, very successful in removing the adverse effect of outliers on the estimation of ACF.

D. D/S variance estimator (DACF)

The moment estimator of ACF can be transformed into the following variance-ratio of differences and sums which is known as D/S variance estimator [17, 18].

$$r(k) = \frac{Var(S_{x_t}^k) - Var(D_{x_t}^k)}{Var(S_{x_t}^k) + Var(D_{x_t}^k)}$$

where $S_{x_t}^k$ and $D_{x_t}^k$ denote the k-th lagged sums and differences of the original time series.

$$S_{x_t}^k = x_t + x_{t-k} \text{ and } D_{x_t}^k = x_t - x_{t-k}$$

$$t = k+1,.......,N \text{ and } k = 1,.......,K$$

The $\alpha \%$ - trimmed ACF-estimator is defined [19] as

$$r(k) = \frac{V_\alpha(y_i) - V_\alpha(z_i)}{V_\alpha(y_i) + V_\alpha(z_i)} \text{ where } i = 0,.......,N-k$$

$$y_i = S_{x_t}^k$$

$$z_i = D_{x_t}^k$$

$$V_\alpha(y_i) = \frac{\sum_{i=0}^{N-k}(y_i - \overline{y}^{(\alpha)})^2}{\sum_{i=0}^{N-k} L_i}$$

$$V_\alpha(z_i) = \frac{\sum_{i=0}^{N-k}(z_i - \overline{z}^{(\alpha)})^2}{\sum_{i=0}^{N-k} M_i}$$

$$\overline{y}^{(\alpha)} = \sum_{t=1}^{N} y_t L_t^{(\alpha)} / \sum_{t=1}^{N} L_t^{(\alpha)}$$

$$\overline{z}^{(\alpha)} = \sum_{t=1}^{N} z_t M_t^{(\alpha)} / \sum_{t=1}^{N} M_t^{(\alpha)}$$

$L_t^{(\alpha)}$ and $M_t^{(\alpha)}$ are weighting functions [20] and can be defined as

$$L_t^{(\alpha)} = \begin{cases} 0, & if \quad y_t \le y_{(g)} \quad or \quad y_t \ge y_{(N-g+1)} \\ 1, & otherwise \end{cases}$$

$$M_t^{(\alpha)} = \begin{cases} 0, & if \quad z_t \le z_{(g)} \quad or \quad z_t \ge z_{(N-g+1)} \\ 1, & otherwise \end{cases}$$

where $y(\cdot) = y_{(1)} \le y_{(2)} \le .... \le y_{(N)}$; ordered observations of the sums

$z(\cdot) = z_{(1)} \le z_{(2)} \le .... \le z_{(N)}$; ordered observations of the differences

$g = [\alpha N]$ is the integer part of αN and $0 \le \alpha \le 0.05$

The results found from these methods are presented here. Because of space limitation we cannot provide our all results. Figure 1 shows the traffic burstiness and autocorrelation plots for UNC data. These real data were taken from the UNC archive [21] with trace collections obtained from a Gigabit Ethernet link. The data set considered here are numbers of TCP packets per 1 ms bin. The ACF plots by different methods confirm that the traffic is LRD. Clearly, traffic persists for more than 3.5 seconds provided by MACF, WACF and TACF. The DACF plot in Figure 1 depicts that it can capture the LRD traffic for a longer period of time. The estimated Hurst parameter for UNC data is 0.63. Figure 2 illustrates the traffic burstiness and autocorrelation plots for video traffic [22]. The frame lengths (in bytes) of intracoded frames (I-frames) per 41.7 ms bin from the movie "Kiss of the Dragon" are considered here. For both video traffic and UNC data, the sample length (N) considered for the dataset is 65000. In Figure 2, MACF, WACF and TACF plots show that the LRD traffic lasts for 150 seconds. The DACF plot shows that the traffic lasts for more than 2500 seconds which is indicative of heavy network load conditions. The estimated Hurst parameter for this video traffic is 0.91 which agrees with the autocorrelation plots. We used the HEAF(2) estimator to measure the degree of LRD (i.e. Hurst parameter).

IV. CONCLUSIONS

The autocorrelation function is a very useful tool in traffic engineering problems, especially when measuring the duration of existing traffic in the network. In this research, we investigate the performance of four different versions of ACF. The experimental results show that variance-ratio of differences and sums estimator (DACF) can capture a longer duration of traffic whereas MACF, WACF and TACF provide similar performance, showing the duration of traffic captured for less period of time. It is particularly important to understand the link between self-similarity and long-range dependence of network traffic and the performance of the networks because such characterization can be potentially applied for essential control purposes such as traffic shaping, load balancing and other strategies of the future.

The autocorrelation function of the traditional Markovian (or memoryless) traffic process decays exponentially, whereas self-similar, long-range dependent traffic decays more slowly and tends to be hyperbolic, implying a strong correlation structure of such a process. Such traffic behaviour (e.g. self-similarity, long-range dependence,

heavy-tailedness) is of great importance for network capacity planning purposes, in which researchers are interested in developing analytical methods or mathematical models for analysing traffic characteristics. There is a significant impact observed when the underlying network traffic exhibits self-similar, long-range dependent behaviour, such as packet drops due to buffer overflow, which, in turn, causes delays in packet transmission. As a result, it makes the network system slower, which leads to a degrading of the quality of service. The autocorrelation function is an efficient tool, which can be used to evaluate network performance by using offline data traces as well as monitoring real time network traffic characteristics. Results from performance evaluation can be used to identify existing problems, guide network re-optimization and aid in the prediction of potential future problems.
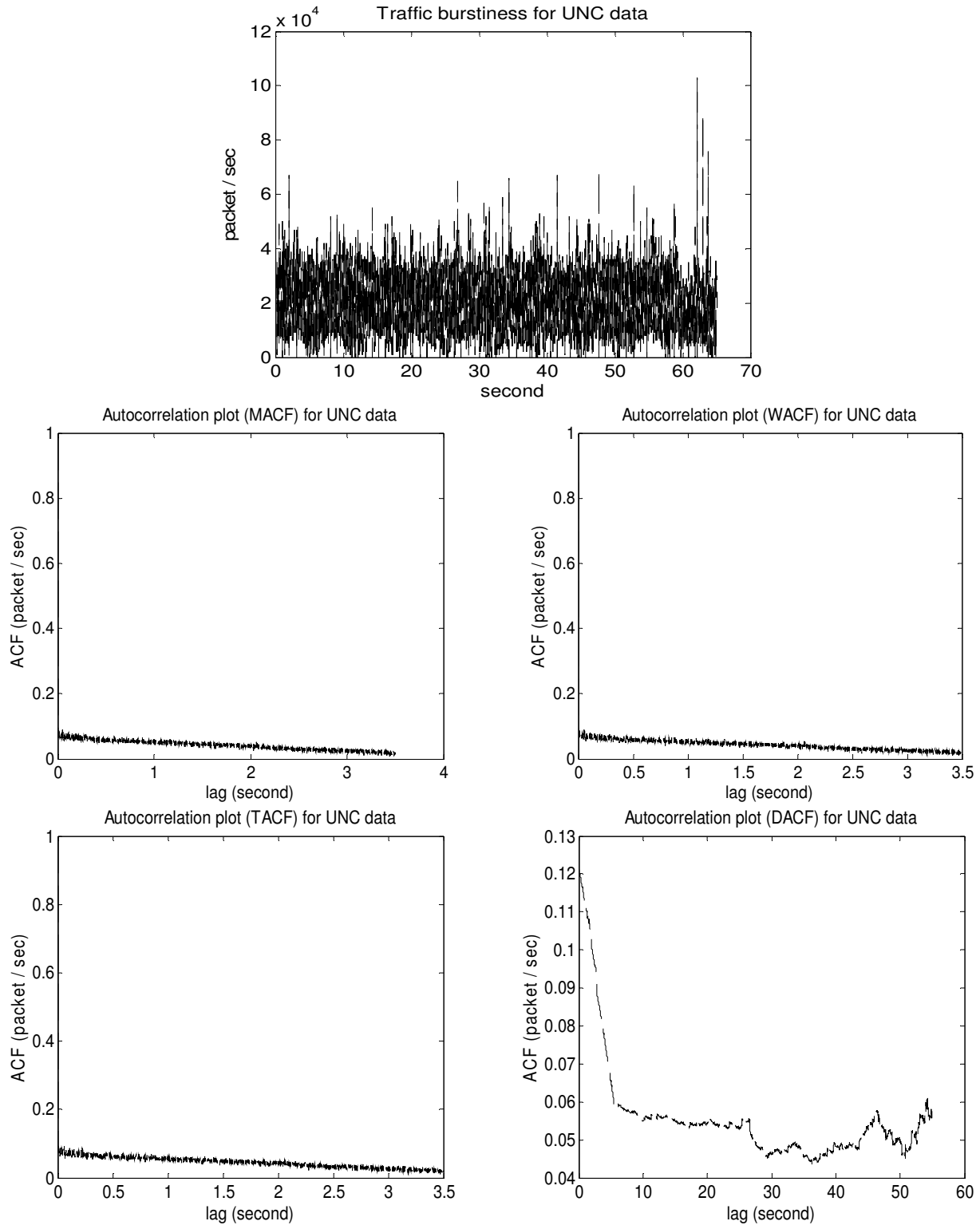

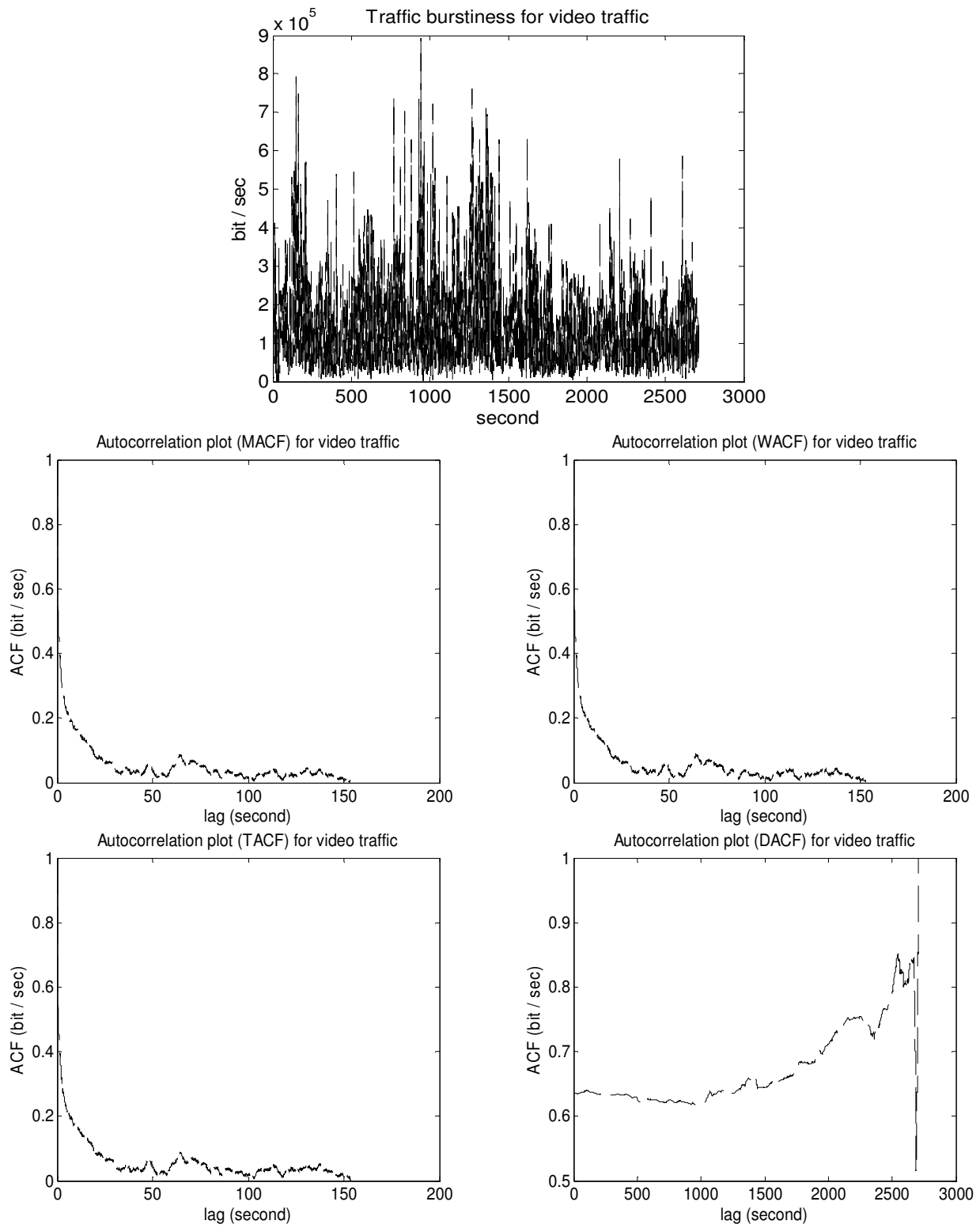Fig. 1. Traffic behaviour by UNC data

Fig. 2. Traffic behaviour by the movie 'Kiss of the dragon'

REFERENCES

[1] Mark E. Crovella and Lester Lipsky, Long-Lasting Transient Conditions in Simulations with Heavy-Tailed Workloads, In Proceedings of the 1997 Winter Simulation Conference, 1997. pp.1005-1012.

[2] Mark E. Crovella and Azer Bestavros, Explaining World Wide Web Traffic Self-Similarity, October 12, 1995, Boston University, Technical Report TR-95-015.

[3] Will E. Leland, Murad S. Taqqu, Walter Willinger, Daniel V. Wilson, On the self-similar nature of Ethernet traffic (extended version), IEEE/ACM Trans. Networking, Vol. 2, No. 1, February 1994, pp. 1-15.

[4] Hurst H. E., Long-term storage capacity of reservoirs, Transactions of the American Society of Civil Engineers, 1951, vol.116, pp 770-808.

[5] Ton Dieker, Simulation of Fractional Brownian Motion, Masters Thesis, Department of Mathematical Sciences, University of Twente, The Netherlands, 2004.

[6] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D, 1988, vol.31, pp. 277-283.

[7] P. Abry, P. Flandrin, M. S. Taqqu and D. Veitch, Wavelets for the Analysis, Estimation, and Synthesis of Scaling Data, K. Park and W. Willinger (editors), Self-Similar Network Traffic and Performance Evaluation. John Wiley & Sons, New York, 2000, pp. 39-88.

[8] Karim M. Rezaul , Algirdas Pakštas, Robert Gilchrist, Thomas M. Chen, HEAF: A Novel Estimator for Long-Range Dependent Self-similar Network Traffic, Y. Koucheryavy, J. Harju, and V.B. Iversen (Eds.): Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN), May 29 - June 2, 2006, LNCS 4003, pp. 34 – 45.

[9] Karim M. Rezaul , Algirdas Pakštas, Robert Gilchrist, Investigation of the Properties of the HEAF Estimator Using Simulation Experiments and MPEG-encoded Video Traces, 10th IEEE International Conference on Intelligent Engineering Systems (INES 2006), London, UK, June 26-28, 2006, pp. 276-281.

[10] Karim M. Rezaul and Grout, V., Exploring the Reliability and Robustness of HEAF(2) for Quantifying the Intensity of Long-Range Dependent Network Traffic, International Journal of Computer Science and Network Security, Vol. 7, No. 2, February 2007, pp221-229.

[11] Brocklebank J. and D. Dickey. SAS System for Forecasting Time Series. SAS Institute Inc. Cary NC. 1986.

[12] Walter Willinger, Vern Paxson, and Murad Taqqu, Self-similarity and Heavy Tails: Structural Modeling of Network Traffic, Adler, R., Feldman, R., and Taqqu, M.S., (editors), In A Practical Guide to Heavy Tails: Statistical Techniques and Applications, Birkhauser, 1998.

[13] Cox D., Long-Range Dependence: a Review. H. A. David and H. T. David (eds.), In Statistics: An Appraisal, Iowa State Statistical Library, The Iowa State University Press, 1984, pp.55-74.

[14] Box, G. E. P., and Jenkins, G. (1976), Time Series Analysis: Forecasting and Control, Holden-Day.

[15] Wang W. and Wei W.S.W., ASA proceedings of Business and Economic Statistics Section, pp. 175-180, 1993.

[16] Chan, W. S. and Wei, W. W. S., A comparison of some estimators of time series autocorrelations, Computational Statistics & Data Ananlysis, vol. 14, 1992, pp. 149-163.

[17] Gnanadesikan, R. - Kettenring, J.R. (1972) Robust estimates,residuals, and outlier detection with multiresponse data. Biometrics 28, pp. 81-124.

[18] Polasek W., Robust and Resistant measures for the autocorrelation function, mimeo, Institute of Statistics and Informatics, University of Vienna, 1982.

[19] Polasek, W. - Mertl, R. (1990) Robust and jackknife estimators for the autocorrelation function. Österreichische Zeitschrift für Statistik und Informatik 20, pp. 351-364.

[20] Wang W. and Wei W.S.W., ASA proceedings of Business and Economic Statistics Section, pp. 175-180, 1993.

[21] http://www-dirt.cs.unc.edu/unc02_ts/, visited on January 2005.

[22] http://trace.kom.aau.dk/preencoded/index.html, visited on February 2005.