J. Indian Inst. Sci., Jan.–Apr. 2004, **84**, 31–39

On the prediction of packet process in network traffic using FARIMA time-series model

CHANDRASHEKHAR G. DETHE[1]* AND D. G. WAKDE[2]
[1]Department of Electronics, [2]Principal, Shri Sant Gajanan Maharaj College of Engineering, Shegaon 444 203, Maharashtra, India.
emails: [1]cgdethe@ssgmce.ac.in; [2]principal@ssgmce.ac.in; Phones: [1]+91-7279-252478, Ext. 303/310 (O) and 253090 (R); [2]+91-7279-252216 (O) and 252208 (R); Fax: +91-7279-252346.

Received on December 3, 2003.

Abstract

The simultaneous existence of short- and long-range dependence in the network traffic has exposed the limitations of conventional traffic models. In this paper, we suggest fractionally integrated autoregressive moving average process (FARIMA) to model the packet process observed in network traffic. We have used different levels of aggregations for computing differencing parameter '*d*'. We also give the complete procedure for modeling and obtaining the predictions for packet process in network traffic using the FARIMA (*p*, *d*, *q*) model.

Keywords: SRD, LRD, FARIMA.

1. Introduction

The analyses of LAN traffic [1] and of wide area network traffic [2] have challenged the commonly used models like Poisson process for packet arrivals. Also it is shown that the traffic is bursty on many time scales and can be statistically described using the notion of self-similarity [1]. Another related notion that has been successfully used in describing real packet arrival process is that of long-range dependence (LRD). The importance of LRD in network traffic has been studied in detail [2–4]. We believe that the use of sophisticated analytical models for packet traffic will help improve the design, analysis and control of real networks. LRD [1], in network traffic, challenges the traffic models such as Markov processes and autoregressive moving average (ARMA) models that are suitable for short-range dependence (SRD) processes. However, SRD cannot be completely ruled out. On the contrary, we need a model which is equally good for capturing SRD and LRD processes. We will explore FARIMA time-series model for capturing LRD as well as SRD [5, 6] and forecast the process $N(t)$ which is described below.

In this paper, we deal with continuous time, non-negative integer-valued process; for example, the number of packet arrivals in an interval (t_1, t_2). Specifically, we will construct forecasting models for a process $N(t)_{t=0}^{\infty}$ where, $N(t)$ = number of instances of the event in the interval $(0, t]$. In Section 3, we define the concept of LRD and in Section 4, we develop the fractional ARIMA (FARIMA) time-series model as a means of modeling network

*Author for correspondence.

traffic and apply it to some real network traffic. Section 5 presents the results and conclusions.

2. Related work

Similar studies were carried by Liu *et al.* [7], Shu *et al.* [8], Basu [9] and Ilow [10]. However, our study is different with respect to certain points. For example, Liu *et al.* [7] and Shu *et al.* [8] have estimated the Hurst parameter using periodogram-based analysis while we have used log–log correlogram. None of the above studies has reported diagnostic checks, while we have included Q–Q plot and ACF (autocorrelation function) plot of residues. These studies do not develop understanding about packet time process $N(t)$ generated from the real traces, which we have described in this study. Also we have used different levels of aggregations for generating such processes and have demonstrated that the differencing parameter d does not change with the change in aggregation scale (Table I).

3. Background: Long-range dependence

Traffic models traditionally used to analyze telephone networks exhibit a correlation structure characterized by an exponential decay. Recent analysis of the packet traffic suggests that the autocorrelations decay at a rate slower than the exponential. Long-memory processes have been discussed in detail by Beran [11] and time-series models that can be used to model such processes by Brockwell and Davis [12]. We summarize the relevant details from these sources. $X_1, X_2, \ldots, X_n$ are sampled observations of the given process $X(t)$ if population mean and variance are $m = E(X_i)$ and $s^2 = \text{var}(X_i)$, respectively. Then we have autocorrelation between X_i and X_j as shown below,

$$r(i, j) = \frac{v(i, j)}{s^2}, \tag{1}$$

where $n(i, j) = E[(X_i - m)(X_j - m)]$ is the autocovariance between X_i and X_j.

Now, if the following equation is true,

$$\sum_{k=-\infty}^{\infty} r(k) = \infty, \tag{2}$$

the correlations decay to zero so slowly that they are not summable. We interpret it as the process having long memory or existence of the long-range dependence in a given stationary process. More formally, long-range dependence is defined as given in the following definition.

Definition 2.1: *Let X_t be a stationary time series with autocorrelation function $r(k)$, for which the following holds. There exists a real number $a \in (0, 1)$ and a constant $c > 0$ such that*

$$\lim_{k \to \infty} r(k)/ck^{-a} = 1; \; k = 1, 2, \ldots. \tag{3}$$

Then X_t is called a stationary process with long memory or LRD.

LRD has a number of implications [13]. First, the variance of n samples from such a series does not decrease as a function of n (as predicted by basic statistics for uncorrelated data sets). Second, the power spectrum of such a series is hyperbolic, rising to infinity at zero frequency reflecting the 'infinite' influence of the LRD in the data. Typical sample paths of such processes appear qualitatively the same, irrespective of the scale of observation.

For historical reasons, parameter H is called Hurst parameter [1]. Using the parameter, for a self-similar process, the autocorrelation is expressed as follows.

$$r(k) \sim ck^{2H-2}, \tag{4}$$

where c is a non-negative constant. The parameter H relates to a given above as $H = 1-a/2$ [11]. Thus, in terms of H, long-memory occurs for $1/2 < H < 1$.

4. Time-series models

As demonstrated by You and Chandra [14], we model the Internet data traffic using time-series models. Literature shows that many other models like fractional Brownian and Gaussian noise capture the long-range behavior of packet network traffic. The reader is encouraged to refer Beran [11] for a detailed discussion on these models. Here we are interested in modeling the LRD behavior of packet traffic and also in predicting the general traffic pattern. In this section, we give the complete procedure for modeling and obtaining predictions using the FARIMA models.

4.1. *Fractional ARIMA*

An autoregressive moving average model of order (p, q), denoted as ARMA (p, q) has the form:

$$X_t = \sum_{i=1}^{p} f_i X_{t-i} + Z_t + \sum_{i=1}^{q} q_i Z_{t-i}, \tag{5}$$

which can equivalently be represented as $f(B)X_t = q(B)Z_t$, where B is the backward difference operator and f and q are polynomials of orders p and q, respectively. If the above equation holds for d^{th} difference $(1-B)^d X_t$, then X_t is called an ARIMA (p, d, q) process. FARIMA generalizes this notion by allowing d to be fractional. For stationarity and LRD we get $0 < d < 1/2$. The parameter d determines the long-term behavior, whereas p, q and the corresponding parameters in $f(B)$ and $q(B)$ allow for more flexible modeling of short-range properties [11]. The process is formally defined as,

Definition 3.2: *The ARIMA (p, d, q) process with* $d \in (-0.5, 0.5)$ *is said to be a fractionally integrated ARMA (p, q) process if* X_t *is stationary and satisfies the difference equation,*

$$f(B)\nabla^d X(t) = q(B)Z_t, \tag{6}$$

where $\{Z_t\}$ is white noise process $(0, s^2)$ and f, q are polynomials in B of degree p, q, respectively. FARIMA processes are asymptotically self-similar with parameter $d-1/2$, where $d = H-1/2$. The ACF for FARIMA$(0, d, 0)$ is given by, $r(k) = ck^{2d-1}$.

4.2. *Preliminaries*

Following conventions are commonly followed in time-series literature: B is the backward-shift operator such that $BX_t = X_{t-1}$. ∇ is the difference operator such that $\nabla X_t = X_t - X_{t-1}$ or in terms of B, $\nabla = (1 - B)$. p, q are non-negative integers and $f(B)$ and $q(B)$ are given by

$$f(B) = 1 - f_1 B - f_2 B^2 - \cdots - f_p B^p, \tag{7}$$

and

$$q(B) = 1 + q_1 B + q_2 B^2 + \cdots + q_q B^q. \tag{8}$$

$q(B)$ has no zeros in the unit disk $\{B:|B| \le 1\}$ and $q(B)$ and $f(B)$ have no common zeros. Note that for an *ARMA* (p, q) process $(p + q + 2)$ parameters, $(f_1, \ldots, f_p, q_1, \ldots, q_q, m, s_z^2)$ need to be estimated. For $0 \le d \le 1/2$, $\nabla^d = (1 - B)^d$ is defined by means of binomial expansion,

$$\nabla^d = (1-B)^d = \sum_{j=0}^{\infty} \binom{d}{j} (-B)^j = \sum_{j=0}^{\infty} p_j B^j, \tag{9}$$

where

$$p_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}, \tag{10}$$

and $\Gamma(.)$ is the gamma function

$$\Gamma(x) = \begin{cases} \int_0^{\infty} t^{x-1} e^{-t} \mathrm{d}t, & x > 0 \\ \infty & x = 0 \\ x^{-1}\Gamma(1+x), & x < 0 \end{cases}, \tag{11}$$

and ∇^{-d} is obtained similarly by replacing d in eqn (6) by $-d$.

4.3. *Estimation of Hurst parameter (H)*

The value of H for any given series X_t can be estimated using any of the methods given in Beran [11]. More importantly, Karagiannis *et al*. [15] have shown that no single method can be trusted to find the LRD in a given time series. However, they do not talk about log–log correlogram plot which we have exploited in this study for estimating the value of H. Importantly, the suitability of log–log correlogram for detecting LRD in the given process is described in Berac [11]. Also our study differs from that of Shu *et al*. [8] who estimate parameter H using periodogram-based analysis. The log–log correlogram is a natural and simple diagnostic tool for estimating H. The log of the estimated ACF against the log of the lag is used for estimating H as follows:

The estimates of the ACVF, $\hat{g}(k)$ of the series X_t are estimated as

$$\hat{g}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (X_t - m)(x_{t+k} - m), \tag{12}$$

where $X_1, X_2, \ldots, X_n$ are elements of the series $\{X_t\}$, $m = (1/n)\sum_{t=1}^{n} X_t$, is the estimated mean and n, the total number of elements in the series. From the ACVF as defined above, the

ACF are estimated as $\hat{r}(k) = \hat{g}(k)\,\hat{g}(0)$. The points in the plot of $\log|r(k)|$ against $\log k$ should be scattered around the straight line with negative slope approximately equal to $2H-2$. This can be easily derived from eqn (4). Thus, from the slope of the best fit line, b, we get H as $H = 1 + b/2$. d of the FARIMA process is then obtained as $d = H-1/2$ or in terms of b, $d = (b+1)/2$. Note that d lies in the interval $(-0.5, 0.5)$ and LRD is defined for $0 < d < 0.5$. Processes for which d lies in the range $-0.5 < d < 0$ are termed intermediate memory processes, the practical implications of which are still not understood. We, however, restrict ourselves to only LRD processes. Thus, if the estimate of d lies in the interval $(0, 0.5)$, we say that the series exhibits LRD else we consider it as a short-memory process.

4.4. *Model identification and initial estimation of model parameters*

The value p is obtained by observing the partial autocorrelation functions (PACF) and the value of q is obtained by observing the ACF of the series X_t. As a simple rule of thumb [12] one draws two horizontal lines at the levels $\pm 2/\sqrt{N}$, N being the total number of the terms in the series. Correlations outside this band are considered significant. We do not suggest that p and q obtained using the above method give the best possible model for the data. The final model is chosen according to the model selection criterion known as the Akaike Information Criterion (AICC). Interestingly, the previous studies of Shu *et al.* [8] and Ilow [10] do not refer AICC for model selection. More information on AICC and its computation can be found in Brockwell and Davis [12]. The pair (p, q) which gives minimum AICC value is chosen and the estimates thus obtained are used as initial estimates for the MLE.

4.5. *Diagnostic checking*

Typically, the goodness of fit of a statistical model to a set of data is judged by comparing the observed values with the corresponding predicted values obtained from the fitted model. If the fitted model is appropriate, then the residuals should behave in a manner that is consistent with the model. The maximum likelihood estimates $\hat{f}, \hat{q}$ and $\hat{s}^2$ of the parameters f, q and s^2 are obtained. The predicted values $\hat{X}_t(\hat{f}, \hat{q})$ of X_t based on $X_1, \ldots, X_{t-1}$ are computed for the fitted model. The residuals are then defined by

$$\hat{W}_t = \frac{X_t - \hat{X}_t(\hat{f},\, \hat{q})}{r_{t-1}^{1/2}}, \tag{13}$$

where $r_{t-1} = E(X_t - \hat{X}_{t-1})^2/s^2$. If we were to assume that the maximum likelihood ARMA (p, q) model is the true process generating X_t, then we could say that $\hat{W}_t \sim WN(0,\, \hat{s}^2)$. The properties of the residuals $\{\hat{W}_t\}$ should reflect those of white noise sequence Z_t generating underlying ARMA (p, q) process. In particular, the sequence $\{\hat{W}_t\}$ should be approximately (i) uncorrelated if $Z_t \sim WN(0,\, s^2)$, (ii) independent if $Z_t \sim$ IID $(0,\, s^2)$, and (iii) normally distributed if $Z_t \sim N(0,\, s^2)$.

The rescaled residuals, $\hat{R}_t$ are obtained by dividing $\hat{W}_t$ by the estimated white noise standard deviations as $\hat{R}_t = \hat{W}_t/\hat{s}$. If the fitted model is appropriate, then the $\hat{R}_t$ will have properties similar to those of a WN (0, 1) sequence or of an IID (0, 1) sequence, if we make stronger assumption that the white noise (Z_t) driving the ARMA process is independent.

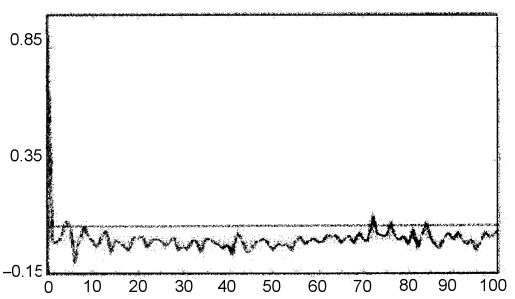


FIG. 1. ACF of residues.

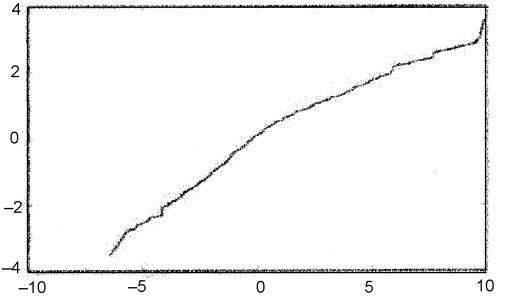


FIG. 2. Q–Q plot of residues.

We have used the following diagnostic checks:

- *The sample ACF test:* For large n, sample autocorrelation of an IID sequence Y_1, Y_2, …, Y_n is a realization of such an IID sequence. About 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{(n)}$. If we compute the sample autocorrelations up to lag 40 and find that more than two or three values fall outside the bounds, we reject the IID hypothesis (Fig. 1).
- *The Q–Q plot:* Let $\hat{Z}_t$, $t = 1, \ldots, n$ be the order of statistics of the rescaled residuals $\hat{R}_t$, $t = 1, \ldots, n$. A simple way to make a Q–Q plot is to simulate n independent realization Y_t from the standard normal distribution. Let Y_t be the corresponding order statistics. Plot $(\hat{Z}_t, \hat{Y}_t)$, $t = 1, \ldots, n$. Gaussianity of the rescaled residuals would lead to the plot being approximately linear (Fig. 2).

4.6. *Forecasting using FARIMA*

Numerous prediction algorithms exist for obtaining the forecasts for an ARIMA process [12]. Obtaining the h-step forecast for a FARIMA model is an extension of the forecasting method used to obtain h-step forecasts for an ARIMA model. We however use the following approach to obtain forecasts for a FARIMA model.

Recall that a FARIMA (p, d, q) model is given by $f(B)\nabla^d X_t = q(B)Z_t$. Denote $\tilde{X}_{n+h}$ as the best linear predictor of X_{n+h} in terms of $X_1, \ldots, X_n$, which is also called the h-step predictor of X_n. Since we are assuming causality and invertibility we can write X_t as,

$$X_t = \sum_{j=0}^{\infty} y_j Z_{t-j}, \tag{14}$$

and

$$Z_t = \sum_{j=0}^{\infty} p_j X_{t-j}, \tag{15}$$

where $\sum_{j=0}^{\infty} y_j B^j = q(B) f^{-1}(B)(1-z)^{-d}$. Then extending the prediction algorithm for a ARIMA process we can write $\hat{X}_{n+h}$ as

$$\hat{X}_{n+h} = \sum_{j=1}^{n+h-1} p_j \hat{X}_{n+h-j}, \tag{16}$$

Table I
Values of *d* for different aggregations

Aggregation	*d*
0.1	0.289860
1.0	0.289732
10.0	0.282991
100.0	0.270460

Table II
Values of *d* for traces pAug1 to pAug5

Data set	*d*
pAug1	0.247660
pAug2	0.231502
pAug3	0.251495
pAug4	0.259395
pAug5	0.286913

Table III
Model parameters for traces pAug1 to pAug5 (except pAug3)

Data set	(*p*, *q*)	Level of differencing	*fs*
pAug1	(0, 1)	0	–
pAug2	(0, 1)	0	–
pAug4	(0, 1)	0	–
pAug5	(0, 2)	0	–

and

$$\hat{s}_n^2(h) = E(X_{n+h} - \hat{X}_{n+h})^2 = s^2 \sum_{j=0}^{h-1} y_j^2 \tag{17}$$

are the mean squared errors of the predictors. For the purpose of computation, $p_j s$ are computed by equating coefficients of same degree on either side of the equation $q(B)p(B) = f(B)(1-x)^d$. Similarly, the coefficients y_js are computed from equation $q(B)y(B) = f(B)(1-z)^{-d}$. The upper and lower probability limits of the forecasts are:

$$\hat{X}_{n+h} = X_n + u\sqrt{E(x_{n+h} - \hat{X}_{n+h})^2}, \tag{18}$$

where u = 0.68, 1.65, 1.96 or 2.58 depending on the probability that a future value lies in the interval is 0.50, 0.90, 0.95 or 0.99, respectively.

5. Results and conclusions

We use the above techniques to obtain models for the publicly available Bellcore network traffic traces [16]. We first obtain the values of *d* for different aggregations of the packet arrival process for the trace *pAug.TL* (Table I).

We see that there is little change in the value of *d* for various aggregations of the trace. We then form the data set of time length 100 s, with 10,000 samples, each representing packet arrivals during 0.01 s (Table II). The estimated model parameters for the traces pAug1, pAug2, pAug4 and pAug5 are shown in Tables III and IV. We illustrate the detailed analysis of the modeling procedure for the trace pAug1 in Figs 1 to 9.

Figure 3 is a plot of the trace of 200 samples. Figure 4 is a plot of ACF of the data set and we see that the ACF is significant for large lags. The limit $2/\sqrt{(n)}$ is also shown in the plot. Figure 5 is the log–log plot of the ACF of the data. A straight line is fit to Fig. 5 and the slope *b* of the line is used in calculating the value of *d*. Figure 6 is a plot of the differ-

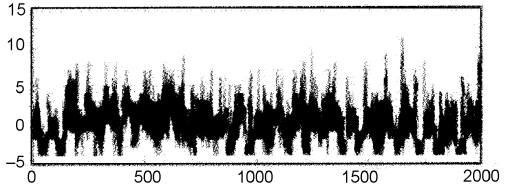


FIG. 3. PAug4 trace.

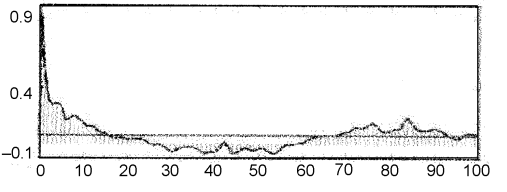


FIG. 4. ACF of data.

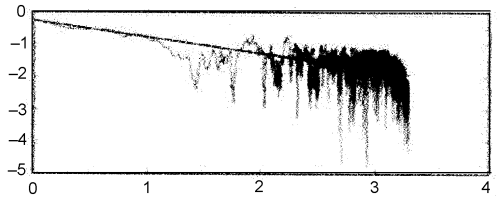


FIG. 5. Log–log plot of ACF.

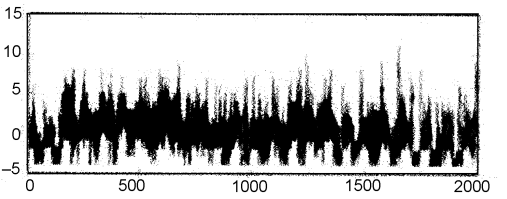


FIG. 6. Trace of the differenced data.

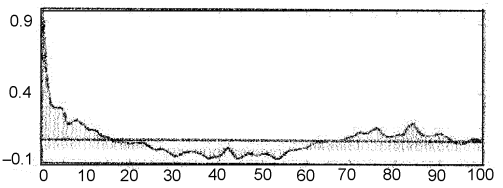


FIG. 7. ACF of differenced data.

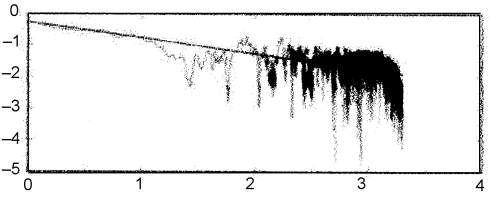
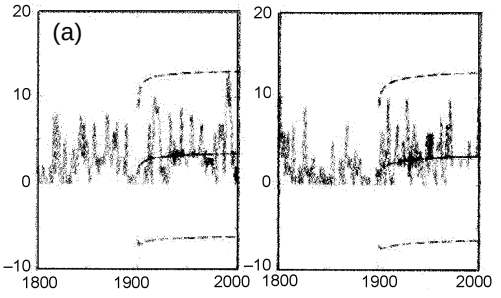


FIG. 8. PACF of the differenced data.



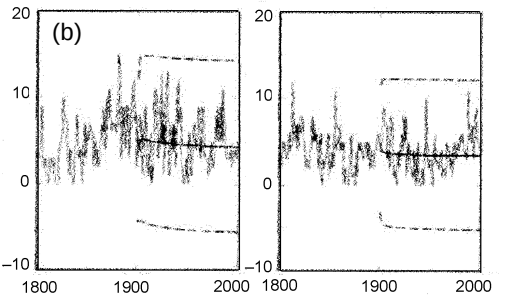


FIG. 9. *h*-step forecasts for (a) pAug1 and pAug2, and (b) pAug4 and pAug5, $h = 100$.

enced series obtained after fractional differencing of the data set. Figure 7 is the plot of ACF of the differenced series and Fig. 8 is a plot of the PACF. We observe that ACF is significant up to lag 2 compared to data set of the ACF plot of the samples. Figure 2, a Q–Q plot of the data set, shows that the plot is approximately a straight line confirming our assumption that the residues $\hat{Z}_t$ are derived from a $WN(0, s^2)$ distribution. Figure 1 clearly shows that the ACF of the residues dies down rapidly. As mentioned earlier, we are interested in forecasting the general behavior of the packet arrival process. Once the FARIMA model has been fit to data as described in the previous sections, the *h*-step forecasts can be obtained. Figure 9 shows the forecasts obtained for the trace files pAug1, 2, 4 and 5. The

Table IV
Model parameters for traces pAug1 to pAug5 (except pAug3)

Data set	qs	AICC	mx	σ_z^2
pAug1	0.335	4459.158	3.899	21.958
pAug2	0.232	4494.183	3.769	23.957
pAug4	0.259	4610.653	3.948	24.953
pAug5	0.148	4368.004	3.271	19.940

forecasts are obtained with 90% upper and lower confidence limits. These slowly die down to the mean of the series indicating LRD nature. Another important conclusion is the overestimation in the variance of the forecasts. This is clearly seen in the large difference between the lower and upper bounds of 90% confidence and also because the values of most in the series lie within this bound. We may also say that the 90% confidence intervals form an envelope for the series being modeled.

References

1. W. E. Leland, M. S. Taqqu, W. Willnger and D. V. Wilson, On the self-similar nature of ethernet traffic (extended version), *IEEE/ACM Trans. Networking*, **2**, 1–15 (1994).
2. V. Paxson, Empirically derived analytical models of wide area TCP connections, *IEEE/ACM Trans. Networking*, **2**, 316–336 (1994).
3. W. E. Leland and D. V. Wilson, High resolution measurements and analysis of LAN traffic applications for LAN interconnection, *Proc. IEEE Infocom*, pp. 1360–1366 (1991).
4. M. Grossglauser and J.-C. Bolot, On the relevance of long-range dependence in network traffic source, *IEEE/ACM Trans. Networking (TON)*, **7**, 629–640 (1999).
5. M. Ghaderi, *On the relevance of self-similarity in network traffic prediction*, Tech. Rep., CS-2003-28, School of Computer Science, University of Waterloo, Waterloo (2003).
6. M. S. Taqqu, *Self-similarity, fractional Brownian motion and long-range dependence*, Tech. Rep., Boston University (2000).
7. J. Liu, Y. Shu, L. Zhang, F. Xue and O. W. W. Yang, Traffic modeling based on FARIMA models, *IEEE Can. Conf. on Electrical and Computer Engineering*, pp. 162–167 (1999).
8. Y. Shu, Z. Jin, L. Zhang, L. Wang and O. W. W. Yang, Traffic prediction using FARIMA models, *IEEE Int. Conf. on Commun.*, **2**, 891–895 (1999).
9. S. Basu, A. Mukherjee and S. Klivansky, Time series models for internet traffic, *IEEE Infocom Conf.*, San Franciso, pp. 611–620 (1996).
10. J. Ilow, Forecasting network traffic using FARIMA models with heavy tailed innovations, *IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, **6**, 3814–3817 (2000).
11. J. Beran, Statistics for long-memory processes, Chapman and Hall (1994).
12. P. J. Brockwell and R. A. Davis, *Time series: Theory and methods*, Springer-Verlag (1991).
13. D. Wischik, *Implications of long-range dependence*, Tech. Rep., Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge (2001).
14. C. You and K. Chandra, Time series models for internet data traffic, *Proc. 24th Conf. on Local Computer Networks, LCN-99*, Lowel, Massachusetts, USA (1999).
15. T. Karagiannis, M. Faloutsos and R. Riedi, Long-range dependence: Now you see it now you don't!, *Global Internet*, Taiwan (2002).
16. http://ita.ee.lbl.gov/html/traces.html.