Markov chain-based method for generating long-range dependence

Richard G. Clegg* and Maurice Dodson†
Department of Mathematics, University of York, York YO10 5DD, United Kingdom
(Received 7 October 2004; published 16 August 2005)

This paper describes a model for generating time series which exhibit the statistical phenomenon known as long-range dependence (LRD). A Markov modulated process based on an infinite Markov chain is described. The work described is motivated by applications in telecommunications where LRD is a known property of time series measured on the Internet. The process can generate a time series exhibiting LRD with known parameters and is particularly suitable for modeling Internet traffic because the time series is in terms of ones and zeros, which can be interpreted as data packets and interpacket gaps. The method is extremely simple, both computationally and analytically, and could prove more tractable than other methods described in the literature.

DOI: 10.1103/PhysRevE.72.026118 PACS number(s): 02.50.−r, 89.75.Da, 05.10.−a, 95.75.Pq

I. INTRODUCTION

Long-range dependence (LRD) is a statistical phenomenon that is used to describe a process that exhibits significant correlations even between widely separated points. A more formal definition is given in Sec. I A. Roughly speaking, a process with a high degree of LRD can be thought of as correlated at all scales. A good introduction to the topic of LRD is provided by [1], and a discussion in the context of telecommunications traffic is given by [2]. LRD is most often characterised by the Hurst parameter H, which is in the range $(1/2,1)$ for a time series that exhibits LRD. If $H=1/2$ then this indicates the data is independent or has only short-range correlations. The topic of LRD has attracted a great deal of interest since LRD has been observed in time series measured in fields as diverse as finance, Internet traffic, and hydrology.

This paper presents and tests a mechanism for generating LRD based on an infinite Markov chain. The traffic stream generated is binary in nature, and the model has only two parameters, the mean and the Hurst parameter of the generated traffic. This section provides a brief introduction to the topic of LRD in the context of telecommunications networks and discusses currently used generation mechanisms for modeling LRD and also methods that are currently used to measure LRD in a time series. Section II describes the infinite Markov model. In Sec. III it is proven that the model does, in fact, generate a time series with a given mean and with LRD having a given Hurst parameter. Finally, Sec. IV tests the model against other standard LRD generation models and discusses the advantages of the model.

A. Brief introduction to LRD

A number of different (and not necessarily equivalent) definitions of LRD are in use in the literature. A commonly used definition is the one given here.

Definition 1. *A weakly stationary time series exhibits LRD if the absolute value of its autocorrelation function (ACF) $\rho(k)$ does not have a finite sum. That is,*

$$\sum_{k=-\infty}^{\infty} |\rho(k)| = \infty.$$

It is often assumed that the ACF has the specific asymptotic form

$$\rho(k) \sim c_\rho k^{-\alpha}, \tag{1}$$

for some positive constant c_ρ and some real $\alpha \in (0,1)$. Note that this is equivalent to a functional form for the spectral density $f(\lambda)$ defined by

$$f(\lambda) = \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) e^{ik\lambda}.$$

Equation (1) is equivalent to

$$f(\lambda) \sim c_f |\lambda|^{-\beta},$$

as $\lambda \to 0$, where σ^2 is the variance, c_f is some positive constant and $\beta \in (0,1)$. The Hurst parameter is then given by $H=(1+\beta)/2$.

It should be noted that here, and throughout this paper, $f(x) \sim g(x)$ is used to mean $f(x)/g(x) \to 1$ as $x \to \infty$; sometimes, in the literature, this symbol is used to mean *asymptotically proportional to* or $f(x)/g(x) \to k$ for some constant k as $x \to \infty$.

The constant α in (1) is sometimes expressed in terms of the Hurst parameter $H=1-\alpha/2$. The Hurst parameter as defined by this relation and (1) is the most commonly used measure of LRD in the telecommuncations literature.

The reason for the interest in the subject within the field of telecommunications is the fact that LRD has been observed in various time series related to Internet traffic [3–5]. It is widely recognized that the engineering implications of LRD on queuing performance can be considerable. If Internet traffic is not modeled well by independent or short-range dependent models, then much traditional queuing theory work based on the assumption of Poisson processes is no longer appropriate. Traffic, which is long-range dependent in nature can have a queuing performance that is significantly

*Electronic address: richard@richardclegg.org
†Electronic address: mmd1@york.ac.uk

 ©2005 The American Physical Society

worse than Poisson traffic. Modeling has shown how phase transitions can arise in computer networks [6], and how this phase transition can be related to LRD [7,8].

In general, it has been found that a higher Hurst parameter often increases delays in a network, increases the probability of packet loss, and affects a number of measures of engineering importance. In fact, Erramilli *et al.* [9] claims that the Hurst parameter is "…a dominant characteristic for a number of packet traffic engineering problems… ." Some of the effects on queuing performance are given by [10,11]. However, [12] shows that although the Hurst parameter is important to queuing, the relationship is not a simple one; in some cases a high Hurst parameter may improve performance or have no effect. The issue of the scale and nature of the effect of LRD on queuing remains contentious.

B. Current generation mechanisms for LRD

A number of modeling techniques are currently used for generating traffic streams exhibiting LRD. Of these, the most commonly encountered in the telecommunications literature are fractional Gaussian noise processes (FGN), fractional autoregressive integrated moving average models (FARIMA, also refered to as ARFIMA), iterated chaotic maps, and wavelet modeling.

The FGN process is usually defined as increments of the fractional Brownian motion (FBM) process. An FBM process $B_H(t)$ is defined by

$$\mathbb{P}[B_H(t+k) - B_H(t) \leqslant x] = (2\pi)^{-\frac{1}{2}} k^{-H} \int_{-\infty}^{x} \exp\left(\frac{-u^2}{2k^{2H}}\right) du,$$

where $\mathbb{P}[X]$ is a probability of an event X and $H \in (1/2, 1)$ is the Hurst parameter. This can be seen as a generalization of the more common Gaussian white-noise process. A number of authors have described methods for generating FGN and FBM [13–15].

The FARIMA model is an obvious modification of the traditional ARIMA(p,d,q) model from time series analysis, allowing $d \in (-1/2, 1/2)$ instead of $d \in \mathbb{Z}_+$. FARIMA processes were proposed by [16], and a description in the context of LRD can be found in pp. 59–66 of Ref. [1]. As might be expected the d parameter relates to the Hurst parameter. The relation is simply $H = d + 1/2$. Note that this only produces legitimate values for H when $d \in (0, 1/2)$.

Iterated chaotic maps, which exhibit intermittency, are also commonly used to generate time series exhibiting LRD. Given a starting value $x_0 \in (0,1)$, then a time series $\{x_n : n \in \mathbb{N}\}$ can be generated by the following map:

$$x_{n+1} = \begin{cases} x_n + \dfrac{1-d}{d^{m_1}} x_n^{m_1}, & 0 < x_n < d, \\ x_n - \dfrac{d}{(1-d)^{m_2}} (1-x_n)^{m_2}, & d < x_n < 1, \end{cases}$$

where $d \in (0,1)$ and $m_1, m_2 \in (3/2, 2)$. If $x_0 \in (0,1)$ then $x_n \in (0,1)$ for all $n \in \mathbb{N}$. If this time series is used to generate a binary time series $\{y_n : n \in \mathbb{N}\}$ by the rule $y_i = 0$ if $x_i < d$ and $y_i = 1$ otherwise, then the series can be shown to exhibit LRD with a Hurst parameter given by $H = (3m-4)/(2m-2)$. This map is illustrated in Fig. 1. An explanation for the presence of LRD in this map is provided by examining the behavior of the orbits at x_i near zero or one. The escape from points near zero or one is extremely slow, and this causes long sequences of zeros or ones in the generated y_i series. Pioneering work in this area is [17] with early applications to telecommunications being given by [18]. This mechanism is particularly suited for generating data for modeling of packet networks, since the ON $(y_i = 1)$ state can be considered to be a packet and the OFF $(y_i = 0)$ state as an interpacket gap.

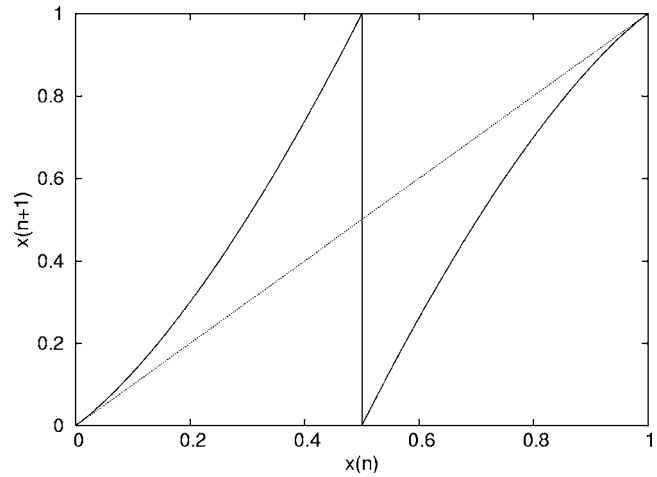

FIG. 1. Graph of a map that can be used to generate LRD.

In fact, the work described in [17] relates to the Markov chain-based work described in this paper as it approximates the chaotic map approach as a piecewise linear map, which can, in turn, be modeled as a Markov chain with the topology described later. A number of other papers have used Markov chains to model linear approximations to intermittent maps [19–22]. The papers [17,18,23] relate the piecewise linear approximations of intermittency maps to LRD and show how certain parameters for Markov chains give rise to LRD in a process arising from the chain. However, we could find no reference to papers that relate intermittency, in general, to LRD.

A technique gaining favor in modeling (and also in measuring) LRD is wavelet analysis. This allows the LRD hypothesis to be generalized to multifractals. LRD defines a single scaling behavior for the system (which applies in the tail of the ACF). If this scaling behavior was the same at any scale, then the process defined would be a monofractal. However, if the scaling behavior differs across scales, then the process is multifractal. There is some evidence that Internet traffic exhibits different scaling behavior at different time scales. A general description of multifractal processes and wavelets is found in [24], and a description of how wavelets can be used to create models with the same multifractal spectrum as a given data set can be found in [25].

C. Measurement techniques for LRD

A number of techniques are known for estimating the Hurst parameter from real data. There is no single technique

that can be considered perfect. Computer code and analysis of various techniques can be found at [26]. Comparisons of measurement techniques can be found in [27–29]. In this paper, five techniques are used: the R/S statistic (in two variants), the aggregated variance, the periodogram, Whittle's local estimator, and a wavelet-based technique.

The R/S statistic (also known as rescaled adjusted range) is one of the oldest and best known techniques for estimating H. It is a time-domain method that relies on considering the way that $R/S(n)$ varies with n where R is the range, S is the sample variance, and n is a scale (sample size) within the time series. It is discussed, in detail, in [30] and also on pp. 83–87 of [1]. There are several problems with this technique that are cited in the literature. The estimate produced is highly sensitive to the range of scales examined. In this paper two versions of the estimator are used that choose the scales to investigate in different ways. The estimator is known to be biased and also slow to converge. It is included in this paper mainly for its historic importance since it has become a standard measure despite its known weaknesses.

The aggregated variance estimator produces an estimate for the Hurst parameter by considering how the variance of the time series scales as the series itself is aggregated into blocks. Again this is a time-domain technique with known weaknesses—jumps in the mean and slowly decaying trends, in particular, can be issues. A fuller description can be found in p. 92 of [1].

The periodogram is one of the oldest frequency-domain-based estimators and is described in [31]. It involves producing an estimate for the spectral density $I(\lambda)$ of the time series and considering the slope of this as $|\lambda| \to 0$. Theoretically, for LRD, a log-log plot of the periodogram should have a slope of $1-2H$ close to the origin.

Whittle's estimator [32] is a frequency-domain technique that uses an approximate maximum likelihood estimator and an estimated functional form for the spectral density $I(\lambda)$ based on an assumed underlying model. Here, the local Whittle variant is used [33], which is a semiparametric version assuming a functional form for $I(\lambda)$ only as $|\lambda| \to 0$.

Wavelet analysis has already been mentioned as a modeling technique and has been used for the estimation of the Hurst parameter. In addition this has the benefit of providing an estimate of the multifractal spectrum of the data [24,25]. This method is based on considering the behavior of the frequency spectrum although wavelets themselves are a technique to allow insight into both frequency and time-domain behavior, simultaneously.

D. Need for a parsimonious and tractable LRD generation method

Given the large (and not exhaustive) list of modeling techniques already mentioned, it might be asked whether there is a need for another model. However, the model here is specifically designed to be the simplest possible computational model that produces LRD.

Fractional Gaussian noise and FARIMA are relatively simple to analyze from a statistical point of view (though the model described here is arguably simpler). However, these processes cannot easily be calculated in an ongoing manner; that is, the entire time series is usually generated "at once" and, having generated n points, the user must effectively start again to generate the $(n+1)$th point.

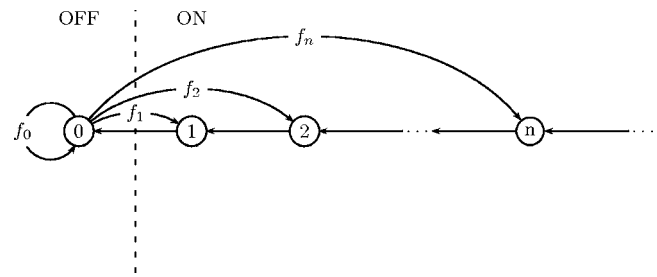

FIG. 2. An infinite Markov chain, which generates a time series exhibiting LRD.

Iterated chaotic maps are computationally parsimonious but are analytically problematic because no closed form for the invariant density of the map described in Sec. I C is known. Therefore, it is difficult to generate traffic with a given mean using the iterated map method, and progress theoretically is difficult. An intermittent map with a known invariant density is given by [34]; however, it is not known if this map would generate LRD and other barriers exist to computational implementation.

The generation mechanism given here is extremely simple, theoretically sound, and has only two parameters, the mean and the Hurst parameter. The data produced are produced in a stream of ones and zeros and can be simply used with simulation models of networks—the one representing a data packet and the zero representing an interpacket gap.

II. THE MARKOV MODEL FOR LRD

Figure 2 shows an infinite Markov chain that can be used to generate a time series exhibiting LRD. This particular chain with different transition probabilities has been studied by a number of authors, notably, in this context Wang [17] and Barenco and Arrowsmith [23] (the latter also investigates the double-sided version). The parameters f_i are the transition probabilities for reaching a given state i from state 0. Also π_i is defined as the equilibrium probability of state i. It is clear that $\Sigma_{i=0}^{\infty} f_i = 1$ and also that $\Sigma_{i=0}^{\infty} \pi_i = 1$. More details and expanded versions of the proofs included here can be found in Chap. 2 of [2].

The chain shown in Fig. 2, given a starting state $X_0 \in \mathbb{Z}_+$, produces a Markov time series $\{X_i : i \in \mathbb{N}\}$ where all the $X_i \in \mathbb{Z}_+$. In turn, this chain can generate a time series $\{Y_i : i \in \mathbb{N}\}$, where $Y_i = 0$ if $X_i = 0$ and $Y_i = 1$ otherwise.

It can be easily shown that the chain above is ergodic (and hence the equilibrium distribution exists) if $\Sigma_{i=0}^{\infty} i f_i < \infty$ and also $\forall i \in \mathbb{N}, \exists j > i : f_j > 0$; the first condition ensures that the mean return time to the zero state is finite and the second ensures that any state in the chain can be reached from the zero state (obviously the zero state will be reached from any state i in exactly i steps). For the rest of this paper it will be assumed that any chain discussed meets these conditions for ergodicity.

Theorem 1. *The equilibrium distribution of the ith state is given by*

$$\pi_i = \pi_0 \sum_{j=i}^{\infty} f_j.$$

Proof. For a state i then at equilibrium, the inputs to a state will sum to π_i. That is,

$$\pi_i = \pi_{i+1} + \pi_0 f_i. \tag{2}$$

Substituting the same equation for π_{i+1} gives,

$$\pi_i = \pi_0 f_i + \pi_0 f_{i+1} + \pi_{i+2},$$

and repeating this subsitution recursively gives the proof. ■

Note that since $\Sigma_{j=0}^{\infty} f_j = 1$, then for $i=0$ this equation simply says $\pi_0 = \pi_0$. Since all the π_i must sum to one then, in addition,

$$\pi_0 = 1 - \sum_{i=1}^{\infty} \pi_i = 1 - \sum_{i=1}^{\infty} i f_i,$$

which, as has already been discussed, is finite.

Introducing LRD into the model

LRD with Hurst parameter H can be guaranteed if the ACF $\rho(k)$ meets the condition given by (1). The most obvious way to induce a correlation for a lag k into such a model is to choose the f_i in such a way that unbroken sequences of k or more ones occur in the Y_n series with the required frequency. Therefore, it would be suspected that the condition,

$$\mathbb{P}[Y_i = 1, Y_{i+1} = 1 \cdots Y_{i+k} = 1] \sim Ck^{-\alpha},$$

where $\alpha \in (0,1)$ will produce LRD with $H = 1 - \alpha/2$. To meet this requirement, the following strict condition is introduced for $k > 0$,

$$\sum_{i=k}^{\infty} \pi_i = Ck^{-\alpha},$$

where C is a constant. Note that there is no guarantee that this is a valid Markov chain (conditions for this will be given later). By setting $k=1$ it is immediate that $C = 1 - \pi_0$. This gives,

$$\sum_{i=k}^{\infty} \pi_i = (1 - \pi_0)k^{-\alpha} \quad k > 0.$$

Subtracting the equation for $k+1$ gives,

$$\pi_k = (1 - \pi_0)[k^{-\alpha} - (k+1)^{-\alpha}] \quad k > 0.$$

From (2) for k then,

$$\pi_0 f_k = \pi_k - \pi_{k+1} \quad k > 0,$$

and therefore for $k > 0$,

$$f_k = \frac{1 - \pi_0}{\pi_0}[k^{-\alpha} - 2(k+1)^{-\alpha} + (k+2)^{-\alpha}], \tag{3}$$

and also, since $f_0 = 1 - \Sigma_{i=1}^{\infty} f_i$,

$$f_0 = 1 - \frac{1 - \pi_0}{\pi_0}\left(\sum_{i=1}^{\infty} i^{-\alpha} - 2\sum_{i=2}^{\infty} i^{-\alpha} + \sum_{i=3}^{\infty} i^{-\alpha}\right).$$

Most of the terms of the sum cancel leaving

$$f_0 = 1 - \frac{1 - \pi_0}{\pi_0}(1 - 2^{-\alpha}). \tag{4}$$

Equations (3) and (4) form the model for LRD. The model is defined by two parameters π_0 and α. The α parameter is related to the Hurst parameter as shown. The π_0 parameter is the equilibrium probability of state zero. Hence $1 - \pi_0$ is the sum of all other equilibrium probabilities and, therefore, the probability that any given $Y_i = 1$. Therefore, the expectation value of Y_i is given by, $E[Y_i] = 1 - \pi_0$. It remains to be proven that the model does generate LRD with the required Hurst parameter, and this is shown in Sec. III.

It can be easily shown that this model meets the conditions for ergodicity established earlier. However, it should be noted that it is not valid for every possible combination of π_0 and α. In particular, for values of π_0 near zero then the term $(1 - \pi_0)/\pi_0$ becomes large and values of f_i from (3) will be negative, a contradiction since the f_i are probabilities. The fact that the model is invalid for some combinations of π_0 and α simply means that for practical experiments the model must be confined to the valid region. Rearranging Eq. (4) shows that for $\alpha, \pi_0 \in (0,1)$ then $f_0 \in (0,1)$ if

$$\pi_0 > \frac{2^{\alpha} - 1}{2^{\alpha+1} - 1},$$

and this defines a valid region for the model.

III. ACF OF THE MARKOV MODEL

It must now be shown that the model described in Sec. II does produce traffic with a given mean and Hurst parameter. To recap, the model relies on a Markov chain of the form shown in Fig. 2 and with transition probabilities given by (3) and (4). The parameters of the model are π_0 and α. Given some starting $X_0 \in \mathbb{Z}_+$, the Markov chain produces a time series $\{X_i : i \in \mathbb{N}\}$ where X_i is the state of the chain at the ith iteration. This is used to produce another time series $\{Y_i : i \in \mathbb{N}\}$ where $Y_i = 0$ if $X_i = 0$ and $Y_i = 1$ otherwise. (The time series $Y_i = 1$ if $X_i = 0$ and $Y_i = 0$ otherwise also produces a series with LRD and mean π_0.) This series has LRD with mean $1 - \pi_0$ and Hurst parameter $H = 1 - \alpha/2$. That $E[Y_i] = 1 - \pi_0$ has already been shown. It remains to be shown that the series has an ACF, which follows the form in (1), and this requires a result due to Feller [35] and is based on Wang [17].

Proof that the chain generates LRD

The event ε occurs whenever $X_i = 0$. It can easily be seen that the number of samples between successive occurrences of ε is an independent and identically distributed variable and hence meets the definition in [35]. A "trial" in the terms of [35] is equivalent to one iteration of the Markov chain in this model.

Definition 2. *If ε occurs at the zeroth trial, then let the number of occurrences of ε in n trials be N_n. Let $F(n)$ be the distribution function of the number of trials between one event ε and the next. (Note that these definitions are those used by [35]).*

If the event ε has just occurred (at the zeroth trial), then the chain is in state 0. If the chain makes the transition to state $i-1$, then the event ε will occur in i steps. Therefore, the distribution function is given by

$$F(n) = \sum_{i=1}^{n} f_{i-1}. \tag{5}$$

The results from [35,17] both assume that the distribution function $F(n)$ obeys

$$1 - F(n) \sim An^{\gamma}, \tag{6}$$

for some positive constant A and some γ. This will now be shown for the specified infinite chain.

From Eq. (5),

$$1 - F(n) = 1 - \sum_{i=1}^{n} f_{i-1} = \sum_{i=n+1}^{\infty} f_{i-1} = \sum_{i=n}^{\infty} f_i.$$

Substituting f_i from (3)

$$\begin{aligned} 1 - F(n) &= \left(\frac{1-\pi_0}{\pi_0}\right)\sum_{i=n}^{\infty}[i^{-\alpha} - 2(i+1)^{-\alpha} + (i+2)^{-\alpha}] \\ &= \left(\frac{1-\pi_0}{\pi_0}\right)[n^{-\alpha} - (n+1)^{-\alpha}] \\ &= \left(\frac{1-\pi_0}{\pi_0}\right)\frac{(n+1)^{\alpha} - n^{\alpha}}{(n+1)^{\alpha}n^{\alpha}} \\ &= \left(\frac{1-\pi_0}{\pi_0}\right)\frac{(1+1/n)^{\alpha} - 1}{n^{\alpha}(1+1/n)^{\alpha}}. \end{aligned}$$

Expanding $(1+1/n)^{\alpha}$ using the binomial theorem gives

$$(1+1/n)^{\alpha} = 1 + \alpha/n + O(n^{-2}).$$

Substituting this expression top and bottom gives

$$\begin{aligned} 1 - F(n) &= \left(\frac{1-\pi_0}{\pi_0}\right)\frac{1 + \alpha/n + O(n^{-2}) - 1}{n^{\alpha}(1 + \alpha/n + O(n^{-2}))} \\ &= \left(\frac{1-\pi_0}{\pi_0}\right)\frac{n^{-\alpha}(\alpha/n + O(n^{-2}))}{(1 + \alpha/n + O(n^{-2}))} \sim \left(\frac{1-\pi_0}{\pi_0}\right)\alpha n^{-(1+\alpha)}, \end{aligned}$$

where $f(n)=O(g(n))$ for functions $f(n)$ and $g(n)$ means that $|f(n)|<Ag(n)$ for some positive constant A and all $n>0$. This is the form required by Eq. (6) with $\gamma=(1+\alpha)$ and $A=\alpha(1-\pi_0)/\pi_0$.

From Theorem 10 of [35], given that the probability distribution satisifies $1-F(x)\sim Ax^{-\gamma}$, where A is a positive constant and $1<\gamma<2$, then

$$\mathrm{var}(N_n) \sim \frac{2A}{(2-\gamma)(3-\gamma)\mu^3}n^{3-\gamma}.$$

In the case of the chain under investigation, $\gamma=1+\alpha$ and $A=\alpha(1-\pi_0)/\pi_0$. Since the chain is ergodic, the mean recurrence time of state zero for the infinite chain is $1/\pi_0$. Therefore,

$$\mathrm{var}(N_n) \sim \frac{2\alpha\pi_0^2(1-\pi_0)}{(1-\alpha)(2-\alpha)}n^{2-\alpha}. \tag{7}$$

From [17] (Eqs. 2.26a and 2.26c) if

$$\mathrm{var}(N_n) \sim Kn^{2-\alpha},$$

for some positive constant K and some $\alpha \in (0,1)$ then the autocorrelation function is given by

$$\rho(n) \sim Cn^{-\alpha},$$

where C is some positive constant. This is the form required by (1).

IV. TESTS ON THE MARKOV MODEL

In this section, two standard models for generating LRD are compared with the Markov model described in this paper. The computational performance of the algorithm is compared against other algorithms.

A. Practical implementation of the model

The only difficulty in modeling the situation on a computer comes in calculating X_{n+1} when $X_n=0$. In this case, a random-number generator and the transition probabilities f_j must be used to find the next state. A naive approach to this would be to generate a random number r uniformly distributed in (0, 1) and say that X_{n+1} is the smallest i such that $\Sigma_{j=0}^{i} f_i < r$. This is fine for low values of i, but as i increases

TABLE I. Procedure for finding X_{n+1} from X_n in the infinite chain.

Steps	Descriptions
1	If $X_n>0$ then $X_{n+1}=X_n-1$. Exit here.
2	Explicitly calculate $\mathbb{P}[X_{n+1}>j]$ for values of $j\leqslant N$ where N is some small integer. Use the procedure for the finite state model to find a value for X_{n+1} if $X_{n+1}<N$.
3	Generate a new random number R in the range [0,1].
4	Calculate $\mathbb{P}[X_{n+1}\in[N,2N-1]\|X_{n+1}\in[N,\infty]]$ from Eq. (8). If R is less than or equal to this probability, then X_{n+1} is in the required range. Otherwise, go to step six.
5	If X_{n+1} is in the required range, then refine down by generating a new random number and seeing if X_{n+1} is in the range $[N,(3/2)N]$. Continue refining by a binary search (with a new random number each time) until X_{n+1} is found. Exit here.
6	Increase the value of N to $2N$ and go to step 3.

TABLE II. Hurst parameter estimates on simulated data.

Source	H	R/S	Mod. R/S	Agg. var.	Periodogram	Local Whittle	Wavelets
FGN							
	0.625	0.637	0.624	0.623	0.626	0.639	0.635
	0.625	0.632	0.624	0.622	0.624	0.638	0.635
	0.625	0.645	0.633	0.620	0.622	0.638	0.635
	0.75	0.728	0.738	0.741	0.747	0.774	0.767
	0.75	0.741	0.736	0.749	0.755	0.776	0.769
	0.75	0.694	0.719	0.741	0.754	0.774	0.768
	0.875	0.784	0.837	0.858	0.877	0.908	0.897
	0.875	0.750	0.823	0.850	0.876	0.908	0.897
	0.875	0.747	0.835	0.860	0.876	0.908	0.898
Iterated map							
	0.625	0.635	0.590	0.604	0.630	0.719	0.706
	0.625	0.608	0.595	0.604	0.627	0.716	0.703
	0.625	0.637	0.594	0.610	0.637	0.718	0.707
	0.75	0.828	0.666	0.717	0.746	0.813	0.800
	0.75	0.725	0.650	0.712	0.739	0.813	0.801
	0.75	0.678	0.694	0.765	0.768	0.814	0.803
	0.875	0.703	0.779	0.851	0.876	0.925	0.910
	0.875	0.779	0.802	0.854	0.877	0.924	0.910
	0.875	0.846	0.817	0.861	0.874	0.925	0.912
Markov							
	0.625	0.526	0.597	0.611	0.621	0.703	0.691
	0.625	0.593	0.645	0.700	0.684	0.710	0.702
	0.625	0.632	0.603	0.646	0.650	0.707	0.698
	0.75	0.663	0.684	0.744	0.760	0.793	0.784
	0.75	0.670	0.667	0.751	0.759	0.793	0.783
	0.75	0.671	0.671	0.724	0.736	0.786	0.776
	0.875	0.724	0.732	0.816	0.848	0.884	0.873
	0.875	0.757	0.754	0.830	0.859	0.885	0.874
	0.875	0.656	0.781	0.852	0.866	0.885	0.875

then this procedure becomes inaccurate because of the finite precision arithmetic used by computers. The problem is that, as i increases the sum gets nearer to one, but the f_i gets nearer to zero (since adding numbers approaching zero to numbers approaching one is likely to produce severe rounding error problems). Hence, the errors in each stage of addition get larger. However, by the very nature of LRD, large values of i are very likely to come up.

It can simply be shown that if $X_n=0$ and $0<k\leqslant i\leqslant j$,

$$\mathbb{P}\{X_{n+1}\in[i,j]|X_{n+1}\in[k,\infty]\} = \frac{i^{-\alpha}-(i+1)^{-\alpha}-(j+1)^{-\alpha}+(j+2)^{-\alpha}}{k^{-\alpha}-(k+1)^{-\alpha}}. \quad (8)$$

Using this equation, Table I shows a procedure for generating the sequence $\{X_n:n\in\mathbb{N}\}$ given some randomly chosen X_0.

B. Hurst parameter estimates

Three generation mechanisms for LRD are compared, fractional Gaussian noise (FGN), iterated maps, and the Markov method developed in this paper. For each method, three different Hurst parameters are investigated and for each of these, three data realizations are created. For each realization, 10^6 points were generated (in the case of the iterated map and Markov method, each of those points was an aggregate of 100 zeros and ones). The Hurst parameter was estimated using the previously discussed measurement techniques to check the match between theory and experiment.

The three methods were implemented in the C programing language. On a 2 GHz processor PC running Debian linux, to generate 10^6 points took 55 s for the Markov method, 60 s for the iterated maps method, and 6 s for the fractional Gaussian noise method. However, it is debatable whether this is a fair comparison since the first two methods could be considered to be generating 10^8 points and aggre-

gating into groups of 100. No C code to generate FARIMA-based data was available, and the R code available took 188 s to generate only 10^5 points. The running time did not seem to scale linearly, and the test to generate 10^6 points was stopped after several hours.

It would naturally be expected that the FGN model is the easiest to estimate, and this shows in the results in Table II. All the estimators were relatively close to correct with the possible exception of the R/S plot on traffic with a Hurst parameter of 0.875, where the underestimate of H was quite severe.

Estimates on the iterated chaotic map traffic were not as successful. The raw R/S plot proved inconsistent and had a hard time estimating higher Hurst parameters. It should be noted, for example, that for $H=0.75$ estimates varied from 0.678 to 0.828. The performance for $H=0.875$ was similarly poor. The modified R/S parameter was better in that it was more stable across runs but tended to overestimate. Local Whittle and wavelets tended to overestimate the Hurst parameter. It should also be noted that the true result was regularly outside the 95% confidence intervals for the wavelet estimator.

Estimates for the Markov-based method were, in many ways, similar to the iterated-map method. If anything, the results from the estimators are slightly closer to the theory and this is particularly notable for the wavelet and local Whittle case. The evidence provided by the estimators is hard to interpret. However, it can certainly be said that the results for the Markov method are as close as the results for the iterated-map method.

Generally, considering the estimators themselves, the R/S method seemed unreliable (and this agrees with theory, which shows it to be a biased estimator with poor convergence). The local Whittle and wavelets methods (which have better theoretical backing) seem to have a better agreement with theory, but it is worrisome that the true Hurst parameter for the data lay outside 95% confidence for the wavelet estimator in many cases.

V. CONCLUSIONS

The method for generating LRD shown here is computationally efficient, extremely simple, and produces a data stream with a given mean and Hurst parameter. The data stream can be generated in an online manner; that is, the method can be started without knowing how many points must ultimately be generated unlike, for example, FGN. The method has been proved theoretically to generate LRD with the required parameters, and this has been tested against a variety of known estimators for the Hurst parameter. It is interesting to see how quite poorly certain estimators perform, even against very standard LRD-generation mechanisms.

Compared to existing methods of generating LRD, this procedure has a number of extremely attractive properties. It is computationally and mathematically extremely simple. Although other models may have more flexibility for precisely representing the nature of the time series being simulated, it is hard to imagine a simpler model for generating LRD. It is hoped, therefore, that this model will be tractable analytically for further developments, for example, analysis of queuing performance of traffic generated by such a model.

ACKNOWLEDGMENTS

The authors would like to thank David Arrowsmith and Martino Barenco for their contributions to this research.

[1] J. Beran, *Statistics For Long-Memory Processes* (Chapman and Hall, New York, 1994).
[2] R. G. Clegg, Ph.D. thesis, University of York, 2004 (unpublished), available online at www.richardclegg.org/pubs/thesis.pdf
[3] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, in *Proc. ACM SIGCOMM*, *San Francisco*, edited by D. P. Sidhu (ACM, New York, 1993), pp. 183–193.
[4] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, IEEE/ACM Trans. Netw. **5**, 71 (1997).
[5] R. Morris and D. Lin, in *Proc. IEEE INFOCOM* (IEEE, New York, 2000), pp. 360–366.
[6] T. Ohira and R. Sawatari, Phys. Rev. E **58**, 193 (1998).
[7] S. Valverde and R. V. Sole, Physica A **289**, 595 (2001).
[8] M. Woolf, D. K. Arrowsmith, R. J. Mongragón, and J. M. Pitts, Phys. Rev. E **66**, 046106 (2002).
[9] A. Erramilli, O. Narayan, and W. Willinger, IEEE/ACM Trans. Netw. **4**, 209 (1996).
[10] I. Norros, Queueing Syst. **16**, 387 (1994).
[11] Z. Sahinoglu and S. Tekinay, IEEE Commun. Mag. **Jan**, 48 (1999).
[12] A. L. Neidhardt and J. L. Wang, in *Proceedings of the 1998 ACM Sigmetrics Joint International Conference on Measurement and Modeling of Computer Systems* (ACM, New York, 1998), pp. 222–232.
[13] B. B. Mandelbrot, Water Resour. Res. **7**, 543 (1971).
[14] R. B. Davies and D. S. Harte, Biometrika **74**, 95 (1987).
[15] V. Paxson, Comput. Commun. Rev. **27**, 5 (1997).
[16] C. W. J. Granger and R. Joyeux, J. Time Ser. Anal. **1**, 15 (1980).
[17] X. J. Wang, Phys. Rev. A **40**, 6647 (1989).
[18] A. Erramilli, R. P. Singh, and P. Pruthi, Queueing Syst. **20**, 171 (1995).
[19] H. Suetani and T. Horita, Phys. Rev. E **60**, 422 (1999).
[20] T. Horita and H. Suetani, Phys. Rev. E **65**, 056217 (2002).
[21] P. Ashwin, A. M. Rucklidge, and R. Sturman, Physica D **194**, 30 (2004).
[22] M. Giampieri and S. Isola, Discrete Contin. Dyn. Syst. **12**, 115 (2005).
[23] M. Barenco and D. Arrowsmith, Dyn. Syst. **19**, 61 (2004).
[24] R. H. Riedi, in *Theory And Applications Of Long-Range Dependence*, edited by P. Doukhan, G. Oppenheim, and M. S. Taqqu (Birkhäuser, Germany, 2003), pp. 625–716.
[25] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk,

45, 992 (1999).
[26] M. S. Taqqu (unpublished), math.bu.edu/individual/murad/home.html.
[27] M. Taqqu, V. Teverovsky, and W. Willinger, Fractals **3**, 785 (1995).
[28] M. S. Taqqu and V. Teverovsky, Scand. J. Stat. **13**, 723 (1997).
[29] J.-M. Bardet, G. Lang, G. Oppenheim, A. Phillipe, S. Stoev, and M. S. Taqqu, in *Theory and Applications of Long-Range Dependence*, edited by P. Doukhan, G. Oppenheim, and M. S. Taqqu (Birkhäuser,Germany , 2003), pp. 557–577.
[30] B. B. Mandelbrot and J. R. Wallis, Water Resour. Res. **5**, 228 (1969).
[31] J. Geweke and S. Porter-Hudak, J. Time Ser. Anal. **4**, 221 (1983).
[32] R. Fox and M. S. Taqqu, Ann. Stat. **14**, 517 (1986).
[33] P. M. Robinson, Ann. Stat. **23**, 1630 (1995).
[34] R. Artuso and G. Cristadoro, Phys. Rev. Lett. **90**, 244101 (2003).
[35] W. Feller, Trans. Am. Math. Soc. **67**, 94 (1949).