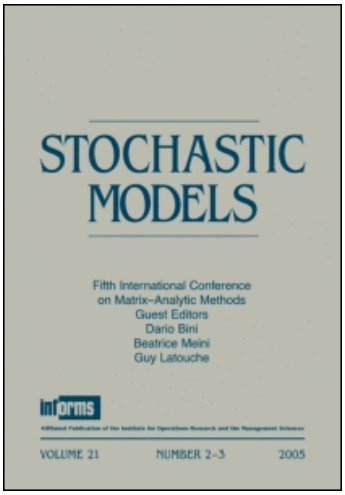


Stochastic Models
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597301

On the Beneficial Impact of Strong Correlations for Anomaly Detection

Matthew Roughan[a]
[a] School of Mathematical Sciences, University of Adelaide, Australia

To cite this Article Roughan, Matthew(2009) 'On the Beneficial Impact of Strong Correlations for Anomaly Detection', Stochastic Models, 25: 1, 1 — 27
To link to this Article: DOI: 10.1080/15326340802640917
URL: http://dx.doi.org/10.1080/15326340802640917

Stochastic Models, 25:1–27, 2009

ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326340802640917

Taylor & Francis
Taylor & Francis Group

ON THE BENEFICIAL IMPACT OF STRONG CORRELATIONS FOR ANOMALY DETECTION

Matthew Roughan

School of Mathematical Sciences, University of Adelaide, Australia

□ *It is now widely accepted that packet network traffic exhibits long-range dependence (LRD), and this has been shown to be harmful to network performance. LRD also reduces the effectiveness of estimators of traffic parameters. For instance, it is much harder to estimate the mean of a LRD process than that of a process with only short-term correlations. One might intuitively expect that LRD would be detrimental to most networking tasks. One important network task is anomaly detection. Anomalies often correspond to problems, for instance, denial-of-service attacks or outages, and so rapid detection is important for maintaining a reliable network. In this article we demonstrate that, counter to the above intuition, LRD is actually beneficial to the detection of anomalies, as in fact are other forms of strong correlations in the observed process. We provide both theoretical proofs and simulation examples to show that LRD in traffic measurements actually improves the probability of detection of anomalies in that traffic.*

Keywords Anomaly detection; Long-range dependence; Network traffic; Self-similarity.

Mathematics Subject Classification Primary 62M; Secondary 94C.

1. INTRODUCTION

Network managers require a number of tools to build highly reliable networks. Apart from hardware and protocols that create redundancy, they also need methods for detecting unexpected problems quickly. Problems such as denial of service (DoS) attacks, network worms, flash crowds, and network configuration problems are hard to prevent in the current Internet, but can often be corrected if they are detected rapidly. There are two approaches to detecting such problems: (i) signature detection, where one tries to detect specific, known signatures of a particular problems, and (ii) anomaly detection, where the problems one wishes to detect are

Received May 29, 2005; Accepted June 14, 2006
Address correspondence to Matthew Roughan, School of Mathematical Sciences, University of Adelaide, South Australia 5005, Australia; E-mail: matthew.roughan@adelaide.edu.au

unknown *a priori*. Thus anomaly detection is ill-defined as the anomalous events one wishes to detect are not prespecified. However, network problems often result in network measurements that have different characteristics from those when the network is operating correctly. Intuitively, the processes that generated anomalies are other than the normal network processes, and this leads to techniques for finding these anomalies by estimating the normal behavior of the processes, and looking for deviations from this behavior. For instance, DoS attacks will often result in outliers in traffic or performance measurements, as do other types of anomalies (see, for instance, Refs.[3,8,18]). This article concentrates on detection of anomalous outliers in network measurements.

One standard example of such a method is Holt–Winters (see, for example, Ref.[5]). However, the analysis that has accompanied work on network anomaly detection is often limited, particularly in the context of network data, which violates the standard statistical assumption of weak correlations between data points. Data network traffic shows properties consistent with self-similarity and long-range dependence (LRD) over a wide range of timescales[9,12]. Long-range dependence refers to the fact that correlations in the traffic volumes, while decaying to zero over longer lags, decay so slowly that their impact never becomes negligible. A typical form for LRD is a power-law decay in the auto-covariance of the traffic rate process, with the exponent being such that the sum of the auto-covariance function over all lags diverges. The result is that many standard statistical results do not hold, and hence we must re-evaluate statistical techniques in this context.

Such re-evaluation has generally given LRD a negative connotation. It has been shown to be harmful to network performance[6,11], by leading to more extended queueing, and persistent transient congestion conditions; and LRD also reduces the performance of estimators of traffic parameters. For instance, it is much harder to estimate the mean of a LRD process, than that of a process with only short-term correlations[4].

One might, therefore, intuitively expect that LRD would be harmful in most networking tasks. In this article we demonstrate that, counter to the above intuition, LRD is actually beneficial to the detection of anomalies. Intuitively, the result arises because processes with high Hurst parameter are actually smoother[15] (under some fairly intuitive definitions of the term smoothness). This is in contrast to much of the literature on self-similar processes, which refers to the Hurst parameter, H, as a parameter of burstiness. However, for higher H, the correlations result in runs—portions of the data that look like smooth trends. These persist for time intervals that scale (in a self-similar way), with longer runs being more common for higher H, leading to a process that is smoother in some respects than, for instance, white noise, in which all of the data points are uncorrelated. An anomaly stands out more clearly against this smoother process.

Despite the fact that it is impractical to suggest that one would deliberately increase the Hurst parameter of traffic in order to make anomaly detection easier, this result is not just of theoretical interest. The result has both a quantitative and qualitative impact on the design of anomaly detection algorithms. Firstly, the theory and results presented here are the first (known to the author) in which quantitative means are presented for the design of anomaly detection algorithms where LRD is present. Other articles, for instance[5] present recommendations for parameter settings, but only in general terms, with no quantitative assessment of the results, in particular in the presence of LRD. Secondly, and more importantly, there is a qualitative result revealed here, of some interest. When LRD is present in traffic (and other measurements), the optimal length of the filter used in anomaly detection may be quite short. In fact, for typical traffic Hurst parameters around $H = 0.8$ the optimal filters are of length 11–17, in contrast to the standard intuition that longer filters would always perform better. This is important, because shorter filters, are

- computationally less expensive,
- have less lag (and so detection may occur more quickly),
- require shorter initialization (i.e., they have fewer edge effects).

The results are in fact more widely applicable than just LRD processes. As one can see from the optimal filter lengths, short-range correlations are still important. The reason for this articles focus on LRD lies in the ubiquity of such processes in network measurements, and the fact that typical LRD process also have strong short-range correlations.

The article presents in Section 2 background definitions and theory for LRD processes, and in Section 3, basic information on anomaly detection including specific definitions of the approach used in this article. Section 4 presents simulation results that demonstrate the basic contribution of this article: that LRD in fact helps anomaly detection. The underlying reason for LRD being beneficial is derived from a theoretical analysis of anomaly detection in Section 5, which also provides an approach for quantifying the gains. Finally, Section 6 uses these results to demonstrate the practical benefits in terms of filter lengths.

2. LONG-RANGE DEPENDENCE

We start by defining precisely what we mean by LRD[9,12]. Given a discrete-time (second-order) stationary process X_i we can define the

constant mean as $\mu = E[X_i]$, and the variance as $\sigma^2 = E[(X_i - \mu)^2]$. The auto-covariance is given by

$$R(k) = E[(X_i - \mu)(X_{i+k} - \mu)] = E[X_i X_{i+k}] - \mu^2. \tag{1}$$

Notice that there is no dependence on i, because of the stationarity assumption. The Fourier transform of $R(k)$ (for a mean zero process) is known as the spectral density and we denote it by $f(v)$. The autocorrelation $r(k)$ of the (stationary) process is defined by $r(k) = R(k)/\sigma^2$.

Long-range dependence is commonly defined by the slow, power-law decrease in the auto-covariance function of a second-order stationary process:

$$R(k) \sim c_r |k|^{-(1-\alpha)}, \quad \text{as } k \to \infty, \tag{2}$$

for $\alpha \in (0, 1)$, where $h(t) \sim g(t)$ denotes asymptotic equivalence, which means $\lim_{t\to\infty} |h(t)|/|g(t)| = 1$. Equivalently, we may define LRD as the power-law divergence at the origin of its spectrum: $f(v) \sim c_f |v|^{-\alpha}$, $|v| \to 0$[4]. The power-law decay is such that the sum of all correlations is always appreciable, even if individually the correlations are small. The past therefore exerts a long-term influence on the future, exaggerating the impact of traffic variability. The main parameter of LRD is the dimensionless scaling exponent α. It describes the qualitative nature of scaling—how behavior on different scales is related. The second parameter, c_r or c_f, is a quantitative parameter which gives a measure of the magnitude of LRD-induced effects. The two are related by $c_f = 2(2\pi)^{-\alpha} c_r \Gamma(\alpha) \sin((1-\alpha)\pi/2)$, where $\Gamma(\cdot)$ is the Gamma function.

There is a direct relationship between LRD and statistical self-similarity. In Ref.[9], statistical self-similarity is defined in terms of aggregates $X_i^{(m)}$ of the original process X_i over nonoverlapping blocks of size m. The aggregates are defined by

$$X_i^{(m)} = \frac{1}{m}(X_{im-m+1} + \cdots + X_{im}), \tag{3}$$

for each $m = 1, 2, 3, \ldots$. The process is called exactly self-similar with Hurst parameter $H = (1 + \alpha)/2$ if the scaled aggregated process $m^{(1-H)} X_i^{(m)}$ has the same statistical properties as X_i. For such a process, the autocorrelation function for the aggregated process is the same as that for the original process X, i.e.,

$$r^{(m)}(k) = r(k), \quad \forall m = 1, 2, 3, \ldots, \ \ k = 1, 2, 3, \ldots \tag{4}$$

Note that Beran[4] defines self-similarity somewhat differently, but the definition used here is more often applied within the context of Internet

traffic measurements. Also notice that real processes are rarely exactly self-similar, but exhibit self-similarity over a range of time-scales, for instance, in network traffic over scales from 10's of milliseconds up to hours (and possibly more). Asymptotic definitions of self-similarity exist[9], but do not provide much more insight for our purposes here.

Self-similarity is often confusingly equated with burstiness in a process, because a self-similar process exhibits burstiness over a range of time-scales. However, this should not be confused with the degree of burstiness at a particular time-scale, which may be significantly smaller for a LRD process than a comparable short-range dependent (SRD) process. For instance, Figure 1 shows two examples of nonstationary processes, one with $H = 0.5$ and the other with $H = 0.9$. Both processes have identical variance, and mean—the only difference is in the autocorrelation function. The curve with larger H appears subjectively smoother, and this smoothness can be also quantified[15].

2.1. LRD Traffic

In this article we shall often consider second-order traffic modeling, that is Gaussian models where the auto-covariance function and the mean specify the model completely. In general, terms the results of the article are also valid for non-Gaussian processes; however, in that case the second-order statistics cannot specify the processes fully. The models will

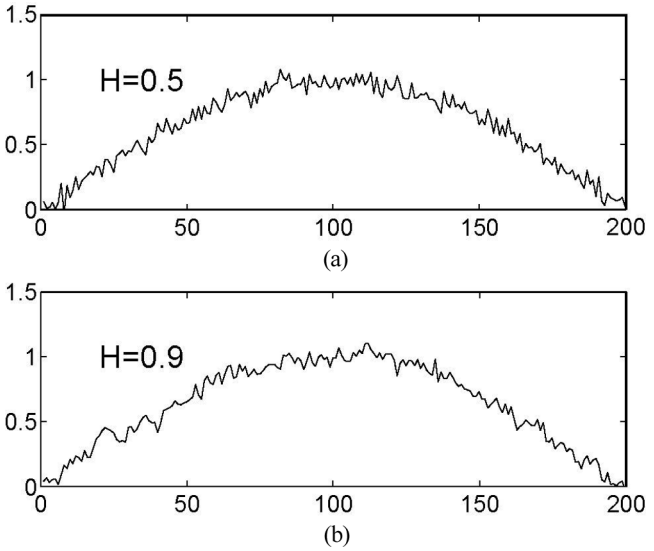


FIGURE 1 Plots of white Gaussian noise ($H = 0.5$), and fractional Gaussian noise ($H = 0.9$) added to a sine function. In each case the standard deviation of the noise is 0.025.

be defined in discrete time, corresponding to the discrete time-series obtained from real measurements.

A simple second-order LRD process is fractional Gaussian noise (FGN), which is the increment process of the self-similar generalization of Brownian motion: fractional Brownian motion. The discrete time FGN X_i has auto-covariance function

$$R(k) = \frac{\sigma^2}{2}(|k+1|^{2H} - 2k^{2H} + |k-1|^{2H}), \tag{5}$$

for $k \geq 0$. Note that if $H = \frac{1}{2}$ then $R(k) = 0$ for all $k \geq 1$, corresponding to white noise, but when $H \in (0.5, 1)$

$$R(k) \sim \sigma^2 H(2H-1)k^{2H-2}, \quad k \to \infty. \tag{6}$$

There is a useful relation[21] between the variance of a FGN and c_f, namely, $c_f = \sigma^2 2(2\pi)^{1-2H} H(2H-1)\Gamma(2H-1)\sin(\pi(1-H))$. Further details of the FGN process can be found in Refs.[4,20]. Fractional Gaussian noise is an appealing model because of its simplicity; we need to specify only three parameters—the mean, variance, and Hurst parameter. We generate approximate FGN sequences here using the spectral synthesis method also used in Ref.[19]. The resulting sequences are therefore not perfect samples of FGN, and in particular the auto-covariance function observes a slight deviation from that in equation (5) for very low lags.

2.2. The Known Impacts of LRD

Long-range dependence has an impact of the variance of estimators, for example, the sample mean $\widehat{X} = 1/N \sum_{i=1}^{N} X_i$ (an unbiased estimate of the true mean of the process). For conventional (stationary) processes, the sample mean quickly converges to the true mean. The law of large numbers, and the central limit theorem (CLT) describe this convergence precisely. The estimator for a series containing only short-range correlations is unbiased (i.e., $E[\widehat{X}] = \mu$), and its variance

$$\mathrm{Var}(\widehat{X}) \sim \sigma^2/N\left(1 + 2\sum_{k=1}^{\infty} r(k)\right) \quad \text{as } N \to \infty. \tag{7}$$

Note that for an uncorrelated series $r(i) = 0$ for $i > 0$, and so $\mathrm{Var}(\widehat{X}) \sim \sigma^2/N$. However, for a LRD process the sum above diverges[4] and so the CLT cannot be applied. For a LRD process, the sample mean of a process

is still unbiased, and converges to the true mean, and the CLT can be generalized to give

$$\mathrm{Var}(\widehat{X}_H) \sim \frac{c_r N^{2H-2}}{H(2H-1)} \quad \text{as } N \to \infty. \tag{8}$$

In contrast to the standard CLT, the variance decreases much more slowly with N. In fact the rate of decrease in the variance is now a function of H: as H increases the variance decreases more slowly, until in the extreme case $H \to 1$ the variance would not decrease no matter how much data we collect. Notice that the aforementioned result applies directly to computing the variance of the aggregated process $X_i^{(m)}$ previously described, and hence the direct relationship between LRD and self-similarity.

Additionally, there have been a number of articles (see, for example, Refs.[6,11]) pointing out that larger values of H would result in much longer queues. The resulting queues often display heavy-tails, where there is significant mass in the tail of the distribution no matter how far one goes into the tail, resulting, potentially in distributions with infinite variance, or even infinite mean. There is some argument about the degree of impact of LRD on performance (e.g., see Ref.[7]), but it is clear that under certain conditions LRD can have a detrimental impact.

3. ANOMALY DETECTION

A number of techniques for anomaly detection in Internet data have been suggested. Typically one assumes that during an anomalous event the network in question will have rather different characteristics from the network under normal circumstances. This is usually manifest in network measurements, for instance, traffic data. A typical implementation might apply anomaly detection to simple network management protocol (SNMP)[10] traffic data giving the traffic volumes (in bytes or packets) on a link during 5-minute intervals[5,18]. An anomaly in such data might arise when a large denial-of-service (DoS) attack occurs. The traffic that arises from such an attack will be quite different from the majority of traffic data resulting from the normal downloading and exchange of files, and so we may look for the differences, rather than a specific signature of a particular attack. Other types of anomalies might arise from Internet viruses or worms, flash crowds, or network malfunctions, though it is important to note that we hope to also be able to detect anomalous events that have not ever been seen before, and so we cannot characterize the anomalous measurements except by their deviation from normal measurements. In addition to SNMP and other types of traffic measurements (e.g., flow records giving volumes of traffic from origin to

destination[3]), one may also observe anomalies in network performance, or routing data[18], though this is perhaps less straightforward.

The majority of anomaly detection techniques that have been presently applied to internet protocol (IP) networks are outlier detection methods. Anomalies may not result in simple outliers, e.g., they may manifest in the frequency domain, but we consider the simple and common case that the anomaly manifests as an outlier in the time-domain. These methods basically work by estimating the normal behavior of the traffic, and detecting outliers that lie well away from this normal behavior.

A simple example of such a technique is to look for values that lie more than δ standard deviations from the sample mean of the data, where the threshold parameter δ determines the false alarm rate. If the data is Gaussian, we can choose the δ to give a particular fixed false-alarm probability, but note that the results presented here are not otherwise dependent on Gaussian data. For any anomaly detection method, there is a tradeoff between probability of detection (sensitivity), and the false alarm probability (related to specificity), which are directly related to the probabilities of type II and I errors, respectively. For instance, in the case mentioned, a larger value of δ will reduce the number of false alarms, but may result in missed detections, whereas for a small δ we will detect most anomalies but at the cost of many false alarms (due to natural statistical variation in the traffic). Any comparison of methods should consider these tradeoffs, with the usual approach being to plot a receiver operating characteristics (ROC) curve, showing the probability of detection versus the false-alarm probability, for a range of values of δ.

The aforementioned approach has an implicit assumption that the process under investigation is stationary, but traffic measurements are certainly not stationary—they exhibit cyclical behavior over periods of days and weeks, as well as long-term trends. In this case, a common approach is to assume approximate stationarity over shorter time periods, for instance, a few hours, and then use the above approach on subsequences of the data of this length. This approach may be implemented using a moving average (MA) of the process over a symmetric rectangular window of length $2M + 1$, i.e.,

$$\widehat{X}_i = \frac{1}{2M+1} \sum_{k=-M}^{M} X_{i+k}, \tag{9}$$

at each time interval i. We then estimate the variance of the data about this mean by

$$\hat{s}^2 = \frac{1}{N} \sum_{i=1}^{N} [X_i - \widehat{X}_i]^2, \tag{10}$$

and we apply thresholds at

$$X_i > \widehat{X}_i + \delta\sqrt{\hat{s}^2} \quad \text{and} \quad X_i < \widehat{X}_i - \delta\sqrt{\hat{s}^2} \tag{11}$$

to signal anomalies.

One advantage of using a symmetric window in the MA mentioned is that it passes linear trends, and so the data in question need not even be stationary. We only require that the data approximately form a linear system over the interval specified by the MA window, i.e., $2M + 1$. However, there are alternatives to rectangular windows, including nonrectangular windows, recursive estimators of the mean, such as the exponentially weighted moving average[18], or methods that use a model for nonstationarity in the data such as the Holt–Winter's method presented in Ref.[5]. The common theme in all of these techniques is that we estimate the local average, and look for large deviations from this average. More sophisticated approaches may look for statistical characterizations of the "normal" mode of behavior of the process other than the mean, for example, Ref.[3] uses the wavelet transform of the data.

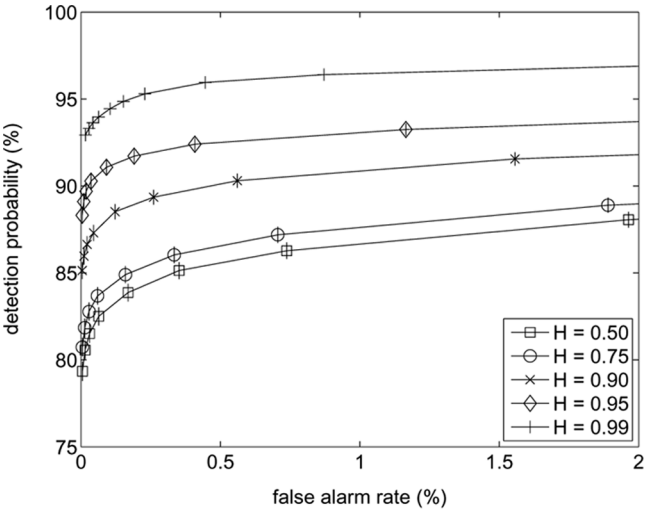
In the results in this article we concentrate on the rectangular-windowed MA previously described, because of its analytic tractability. However, in our simulation results, we also test more sophisticated methods such as Holt–Winters as described in Ref.[5] and the decomposition method used in Ref.[18].

4. SIMULATION RESULTS

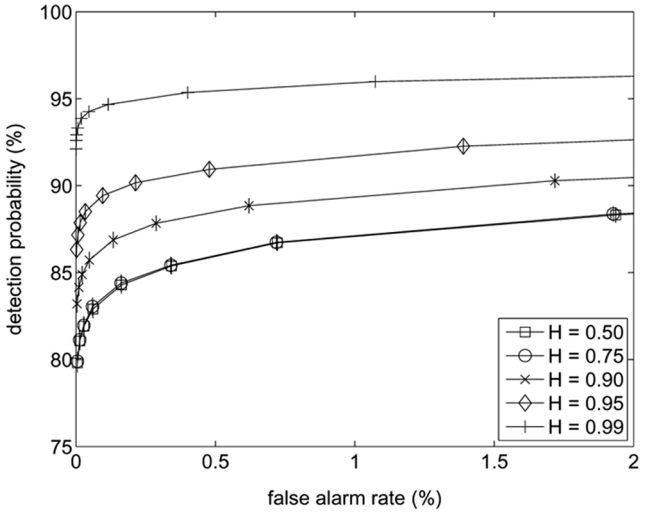
In this section, we present simulation results. While it would be preferable, in some respects, to base our results on real traffic data, this is not practical for this work, because we need to be able to: (i) control the parameters of the analysed data precisely, and (ii) have complete knowledge of the location of anomalies in the data. Unfortunately, even with carefully collected data-sets (as in Ref.[3]) there are some ambiguities in the data complicating precise determination of the false alarm rate[14].

We initially simulated FGN sequences with mean $\mu = 1$, standard deviation $\sigma = 0.1$, and length $N = 10000$ data points for a variety of values of H. Into each sequence we inject $N_A = 4$ anomalies at randomly chosen time points. We generated $S = 5000$ such sequences for each value of H, to allow for accurate assessment of the detection and false-alarm probabilities. The injected anomalies were drawn from a uniform distribution on the interval $[0, 4]$. The range was chosen so that many of these anomalies would fall well away from the FGN process, but a significant number (around 10%) would fall within two standard deviations of the mean of the FGN process, and therefore be difficult to detect. Note though, that all anomalies were uncorrelated with the FGN process. We repeat the

experiment for a range of threshold parameters δ (ranging from 0.06 to 0.4), which allows us to study the tradeoff between sensitivity and specificity. For each simulation i, and value of δ we total the number of correctly identified anomalies C_i^δ and divide by the total number of injected anomalies ($N_A S$) to get the detection probability $p_{det}^\delta = \sum_{i=1}^{S} C_i^\delta / N_A S$. The false-alarm probability was obtained by dividing the total number of



(a)



(b)

FIGURE 2 Detection probability vs. false alarm rate (ROC curves) for varying Hurst parameter. The short vertical lines are approximate 95th percentile confidence bounds for the results.

incorrectly identified anomalies E_i^δ by the total number of nonanomalous data points to get $p_{FA}^\delta = \sum_{i=1}^{S} E_i^\delta/(N - N_A)S$. We plot the ROC curves in Figure 2 by plotting the values of the two probabilities against each other as δ varies. A separate curve is shown for each value of H. Note that one is interested in the cases where the false-alarm probability is quite small (given the number of samples being taken). Figure 2 also shows approximate 95th percentile confidence intervals for the detection probabilities as short vertical lines (the confidence intervals for the false alarm probabilities are inconsequential).

The results are shown for two different sized windows, $M = 10$ and $M = 100$, in Figure 2(a) and 2(b), respectively. The ROC curves show the detection probability versus the false-alarm probability for a range of values of H. Points in the upper-left-hand quadrant of the plot indicate high detection probabilities combined with low false-alarm rates. Notice that we see significant improvements for larger values of H, for both values of M. Even for high H there are some anomalies (around 3%) that are very hard to detect, because they lie close to the value they would have had otherwise. Thus, none of the curves has a probability of detection near 100%, a fact that would not be discernible without *a priori* knowledge of exactly which data points are generated by anomalies.

Note that in these results, one may observe that in many cases the shorter MA filter (smaller M) results in higher detection probabilities for a given false alarm rate. We will show in Section 6 that shorter filters may often work better for LRD data, and derive optimal filter lengths.

A more realistic traffic model would include some cyclical variations corresponding to the daily cycles observed in real SNMP data. We build such a model by incorporating a simple sinusoidal mean into the traffic, using the model of Refs.[16,21]. That is, we take the traffic at time t to be given by $x_t = m_t + \sqrt{am_t}w_t$, where the mean includes a sinusoidal component (with period = 24 hours), and an exponential trend, and w_t is a FGN process with Hurst parameter H. Given a model containing cyclical components, one would naturally seek an anomaly detection technique that explicitly allows for such, examples of which are Holt–Winters, as described in Ref.[5], and a decomposition technique that is designed to separate the m_t and w_t components above[18]. The second method has a number of advantages due to better matching of its behavior to the above model for traffic data. Figure 3 shows results given this new cyclical traffic, and the two detection methods. We see that the decomposition method performs better than Holt–Winters, but the notable fact, in this context, is the improvement in both techniques for higher values of H. Notice that the improvement is not as marked as for the simple example above, a fact which we explain below. However, note that there is still a significant improvement in the results for high values of H.

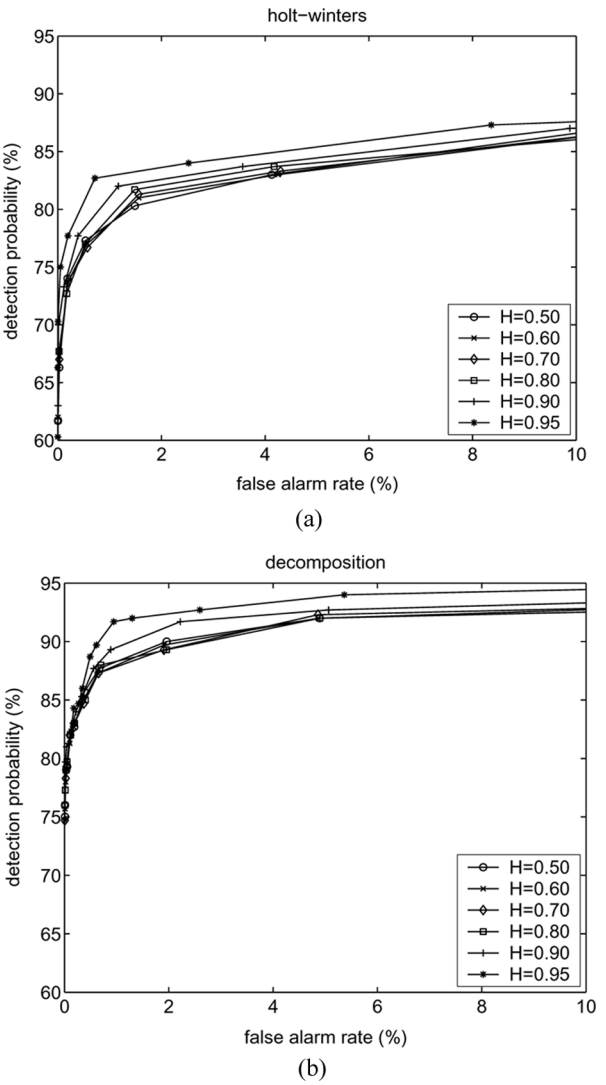


FIGURE 3 ROC curves for varying Hurst parameter for more sophisticated detection methods.

These results are hardly a comprehensive summary of anomaly detection. There are many parameters we could vary, and different models we could use here. However, the point of this article can be clearly seen in these results, namely, that the presence of LRD in the traffic is beneficial when it comes to detecting anomalies. We have performed more extensive simulations (omitted for brevity) and this property is consistently observed. In the next section we examine why this property occurs, and show that it is a generic feature related to the correlations in the normal traffic.

5. EXPLANATION

Usually a higher value of H leads to problems, so it appears counterintuitive that larger values of H would make our detection algorithms perform better. However, intuitively, higher values of H result in data points that are more closely correlated, and will thus lie closer together. Outliers will then "stick out" more clearly. We quantify the above argument here.

5.1. Variance of Estimators

Let us first look at the variance of the estimator $\widehat{X}_i$ from equation (9). For a second-order stationary process X_i, with mean μ, variance σ^2, and auto-correlation function $r(k)$, then the moving-average estimate $\widehat{X}_i$ with window size $K = 2M + 1$ is an unbiased estimator of μ and has variance

$$\mathrm{Var}(\widehat{X}_i) = \sigma^2 g_r(M), \tag{12}$$

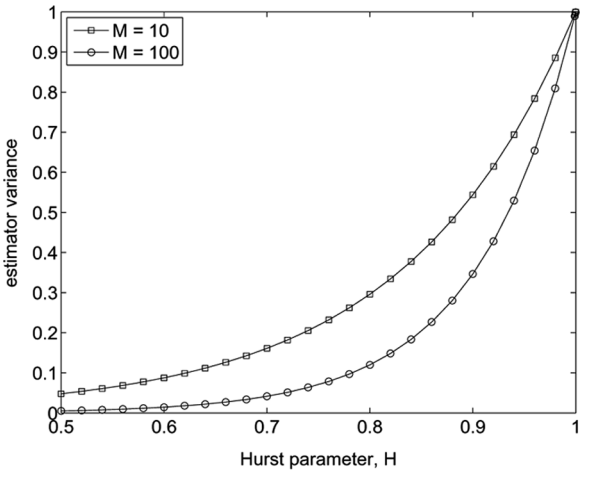
where

$$g_r(M) = \frac{1}{K} + \frac{2}{K}\sum_{i=1}^{2M} r(i)\left(1 - \frac{i}{K}\right). \tag{13}$$

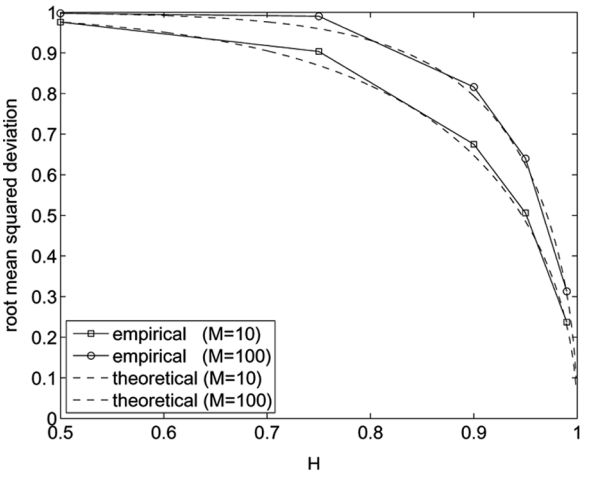
The above result is well known—see, for instance, Ref.[4]. Correlations in the data mean that each additional data point contributes less information than it would if the data were independent. Notice that the aforementioned results are not dependent on a process being LRD or otherwise SRD, because the above formula are concerned with finite sums. The results depend only on the value of the sum in equation (13), but one would often expect that this sum would be larger for a LRD process, and that it would be larger for larger values of H. Figure 4(a) shows this estimator variance for two values of M. Notice that it increases to unity (the variance of the process is one) for larger H.

5.2. Mean-Squared Deviation

The fact that the variance of an estimator of the mean is increased in the presence of strong correlations is not directly relevant to anomaly detection. Anomaly detection techniques may make use of estimates of the mean (which has higher variance), but not directly. In the simple methods previously described the deviation of a data point from the mean estimate is used to signal an anomaly. However, where the process has strong correlations the mean estimate will actually lie closer to the measured data,



(a)



(b)

FIGURE 4 Moving-average estimator variance and mean squared deviation for an FGN process ($\sigma^2 = 1$).

and so the residual difference between the mean estimate and the data points will be reduced. Hence it is easier to distinguish an anomaly.

This result can be quantified. For anomaly detection the important quantity is not the variance of the estimator of the mean, but rather the mean-squared deviation of the estimate away from the data point itself, i.e., the value of $\hat{s}^2$, as given in equation (10). Anomalies are detected when they lie outside a threshold $\delta\hat{s}$, so the value of $\hat{s}^2$ is a real indication of the variance of the noise process which forms the background against which we must estimate whether a data point is an anomaly. A smaller value of $\hat{s}^2$

makes anomaly detection easier. The expected value of $\hat{s}^2$ is given by the following theorem.

Theorem 5.2.1. *Take a wide-sense stationary process X_i, with mean μ, variance σ^2, autocorrelation function $r(k)$, and moving average estimate of the mean $\widehat{X}_i = \frac{1}{2M+1}\sum_{k=-M}^{M} X_{i+k}$ then $\hat{s}^2$, defined in equation (10), has expected value given by*

$$E[\hat{s}^2] = E\left[(X_i - \widehat{X}_i)^2\right] = \sigma^2 f_r(M), \tag{14}$$

where

$$f_r(M) = \left(1 - \frac{1}{K}\right) + \frac{2}{K}\sum_{k=1}^{2M} r(k)\left(1 - \frac{k}{K}\right) - \frac{4}{K}\sum_{k=1}^{M} r(k), \tag{15}$$

and $K = 2M + 1$.

Proof. We can expand $E\left[(\widehat{X}_i - X_i)^2\right]$ to get

$$E\left[(\widehat{X}_i - X_i)^2\right] = E\left[\widehat{X}_i^2\right] + E\left[X_i^2\right] - 2E\left[\widehat{X}_i X_i\right]. \tag{16}$$

for which the first term is given by equation (12), $E\left[X_i^2\right] = \sigma^2 = R(0)$, and

$$\begin{aligned} E\left[\widehat{X}_i X_i\right] &= \frac{1}{K}\sum_{k=-M}^{M} E[X_{i+k}X_i] \\ &= \mu^2 + \frac{R(0)}{K} + \frac{2}{2M+1}\sum_{k=1}^{M} R(k). \end{aligned} \tag{17}$$

Substituting the result of equations (12) and (17) into equation (16), we get

$$\begin{aligned} E\left[(X_i - \widehat{X}_i)^2\right] &= \mathrm{Var}(\widehat{X}_i) + 2\mu^2 + \sigma^2 - 2\mu^2 - 2\frac{\sigma^2}{K} - \frac{4}{K}\sum_{k=1}^{M} R(k) \\ &= \sigma^2\left(1 - \frac{1}{K}\right) + \frac{2\sigma^2}{K}\sum_{k=1}^{2M} r(k)\left(1 - \frac{k}{K}\right) - \frac{4}{K}\sum_{k=1}^{M} R(k), \end{aligned} \tag{18}$$

and note once again that $R(k) = \sigma^2 r(k)$. □

Once again, this result does not assume LRD, but we would expect the summation terms to be larger in this case. However, note that there are both positive and negative terms summed across the autocorrelation, and so it is not immediately obvious whether LRD will increase or decrease

the mean-squared deviation. Figure 4(b) shows the theoretical results (as dashed lines) for two values of M, using the autocorrelation function for FGN, given in equation (5). We can see that $\hat{s}^2$ decreases strongly for larger H. Figure 4(b) also shows (as solid lines) the experimental values of $\hat{s}^2$, derived from the above simulation experiments. We can see that, while not a perfect match to the theory, they observe the same trends. The fact that they do not match precisely may be a result of edge effects, and distortions from the anomalies. Further deviations may arise because of slow convergence of the measured value of $\hat{s}^2$ to its theoretical value, similarly to the slow convergence previously described for the sample mean.

The result is applicable to any process with strong correlations, such as an auto-regressive (AR) process, with $r(k) = p^k$ for p near one. In this case, only the first few terms will contribute to $\hat{s}^2$, because of the geometric decay, and so we could approximate the expression for $\hat{s}^2$ in Theorem 5.2.1 by

$$\hat{s}^2 \simeq \sigma^2\left(1 - \frac{1}{K}\right) - \frac{2\sigma^2}{K}\sum_{k=1}^{j} p^k, \tag{19}$$

where M is large enough that we can choose $j \ll M$, such that p^{j-1} is very close to zero. We can then approximate

$$\hat{s}^2 \simeq \sigma^2\left(1 - \frac{1}{K}\right) - \frac{2\sigma^2}{K}\frac{p}{1-p} = \sigma^2\left(1 - \frac{1}{K}\left[\frac{1+p}{1-p}\right]\right), \tag{20}$$

so that, for p near one, the second term will be large enough to cancel some of the first term (the approximation could become negative, but this would simply be an indication of a regime where the approximation did not hold). The result would be a quite small value of $\hat{s}^2$ even for a short-range correlated process: a simple AR(1) process. The important distinction between SRD and LRD processes is that, in the case of a LRD process, the autocorrelation function does not change with aggregation, and so the only change to the above results with aggregation is the change in the variance σ^2. However, when aggregating a non-LRD process, the autocorrelation will eventually decay to a degenerate function, i.e., $r^{(m)}(k) \simeq 0$ for all $k \neq 0$, for m large enough. Hence, for such a process, enough aggregation will eventually wash away the above effect. A similar effect impacts both Holt–Winters and the decomposition technique. These techniques assume a periodic nonstationary component in the traffic. They therefore use moving averages over data points separated by the period. Such data points will be less correlated (even for process containing LRD), and hence the summations above will be smaller, and $\hat{s}^2$ will be larger. Hence the results shown in Figure 3.

We have previously argued that the critical quantity for estimating the performance of anomaly detection techniques is $\hat{s}^2$. Smaller values of $\hat{s}^2$ results in easier anomaly detection. More precisely, for a smaller values of $\hat{s}^2$, we may obtain a higher detection probability for any fixed false-alarm rate. The exact degree of improvement is quantified below.

5.3. Detection Probabilities

Notice first that we can rewrite the condition for detection of an anomaly (11) as an anomaly will be indicated if

$$|\widehat{X}_i - X_i| > \delta\sqrt{\hat{s}^2}. \tag{21}$$

Assume that the process X_i is Gaussian with mean μ, variance σ^2, and autocovariance $R(k)$. Then the MA $\widehat{X}_i$ is formed from a sum of Gaussian random variables, and therefore is itself a Gaussian random variable, with mean μ and variance given in equation (12). Likewise, $\widehat{X}_i - X_i$ will also be Gaussian, with mean zero, and variance given by $E[\hat{s}^2]$. Therefore, the probability of a false alarm (where no anomalies are present) is the standard probability that a Gaussian random variable with mean zero, and variance $E[\hat{s}^2]$ will fall outside a particular interval. We can therefore determine δ to fix this false alarm rate by choosing $\delta = z_{\beta/2}$, where $z_{\beta/2}$ is the $1 - \beta/2$ quantile of the normal distribution, and $p_{FA} = 1 - \beta$ is the desired probability that a normal measurement will trigger a false alarm. Notice here that δ is independent of $E[\hat{s}^2]$, and only depends on the desired false alarm probability, so δ can be set in advance of obtaining any data. Furthermore, the threshold for detecting an anomaly $T = \delta\hat{s}$ and so the threshold decreases as $\hat{s}$ decreases, and smaller outliers will become detectable.

To determine detection probabilities, we must also have a model for the anomalies. Here, we use a Gaussian model (in the simulations we used a uniform random variable, and detection probabilities could be computed for other such models as needed). We assume that the anomalies are generated by a process Y with mean $\mu + a$ and variance b^2, and that the Y_j are uncorrelated with each other, and with the normal traffic process X_i. At times where there is no anomaly, the measured traffic is given by X_i. At time point j, where there is an anomaly present in the data the measured traffic is replaced by Y_j. Assume that K is large enough that the anomaly produces minimal bias in the MA estimate, the probability of detection of the anomaly is then given by

$$\text{probability of detection} = \text{prob}\left\{|\widehat{X}_j - Y_j| > \delta\sqrt{\hat{s}^2}\right\}. \tag{22}$$

Once again, due to the assumptions about Gaussianity of the traffic and anomaly process, the random variable $\widehat{X}_j - Y_j$ is Gaussian, this time with mean a and variance given by

$$\mathrm{Var}(\widehat{X}_j - Y_j) = \mathrm{Var}(\widehat{X}_j) + \mathrm{Var}(Y_j), \tag{23}$$

because the two random variables $\widehat{X}_j$ and Y_j are uncorrelated. The variance $\mathrm{Var}(\widehat{X}_j)$ is given in equation (12) and $\mathrm{Var}(Y_j) = b^2$ by definition, so the problem reduces to considering when the absolute value of a normally distributed random variable $N(a, b^2 + \mathrm{Var}(\widehat{X}_j))$ exceeds a threshold. The probability $P(T; a, \phi^2)$ that the absolute value of a Gaussian random variable $N(a, \phi^2)$ exceeds a threshold T is given by

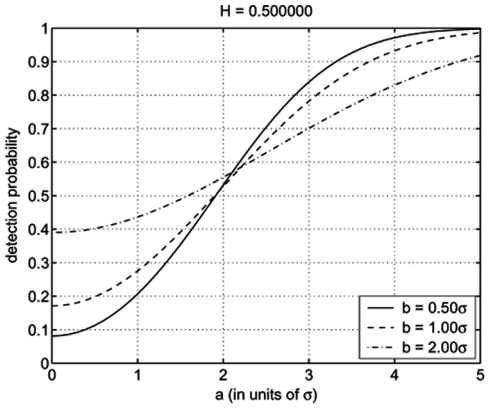
$$P(T; a, \phi^2) = 1 - \Phi\left(\frac{T - a}{\phi}\right) + \Phi\left(\frac{-T - a}{\phi}\right), \tag{24}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution, $T = \delta\hat{s}$, and the value of the variance of the difference is $\phi^2 = b^2 + \mathrm{Var}(\widehat{X}_j)$.

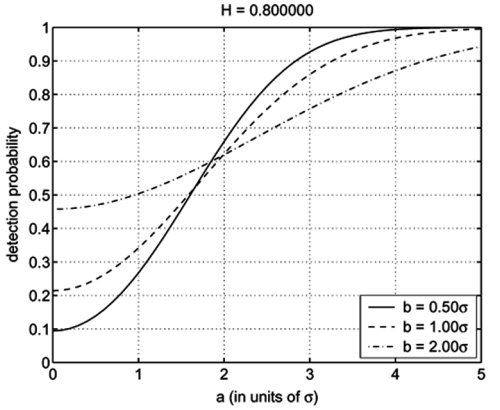
Notice that if we write b in units of σ, we can remove a factor of σ^2 from T and from ϕ, rescaling the problem in nondimensional units. Thus, in the following results, we scale a and b by σ, the standard deviation of the normal traffic process X. Figure 5 shows the detection probabilities for a fixed false-alarm probability (chosen by setting a fixed value of δ, in this case $\delta = 1.96$), and varying a, b, and H in the above computation. Although changing δ obviously changes the false-alarm probability and detection probabilities, the general trends observed here remain the same. The threshold T is fixed for a particular value of H, as previously described. We can see that, as the Hurst parameter increases, the mean difference decreases, and so the threshold required (for a fixed false alarm probability) decreases. The figures also show that as a increase (i.e., the mean value of the anomalies is further from the mean value of the normal traffic) the detection probability increases. With respect to b, a larger value is better for small a, because the wider variance will result in more anomalies outside the threshold. However, for large a (a noticeably larger than the threshold) we are better off with smaller b, because this better isolates the anomalies away from the normal mode of operation.

Figure 5 also shows that the detection probability increases with H, but this is hard to see in these plots, so we investigate it further in Figure 6, which shows the same results. However, this time, the plots highlight the difference with respect to H. Note that these plots show fixed false-alarm rate, so the curves clearly show an improved detection probability for higher Hurst parameter.

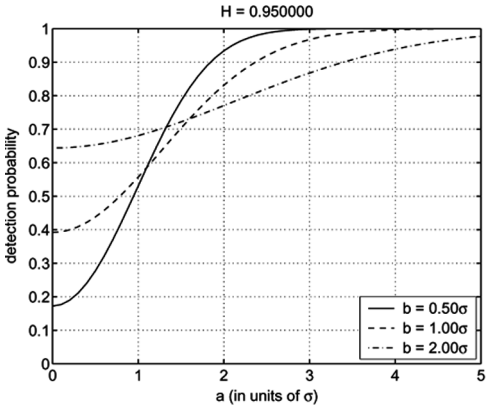
Another way to view the results is shown in Figure 7, this time as a continuous function of the Hurst parameter. This illustrates that the



(a)

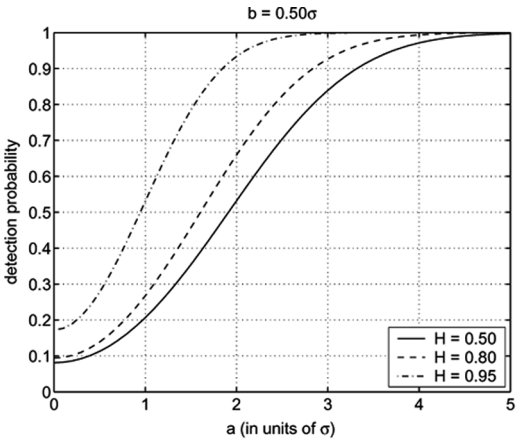


(b)

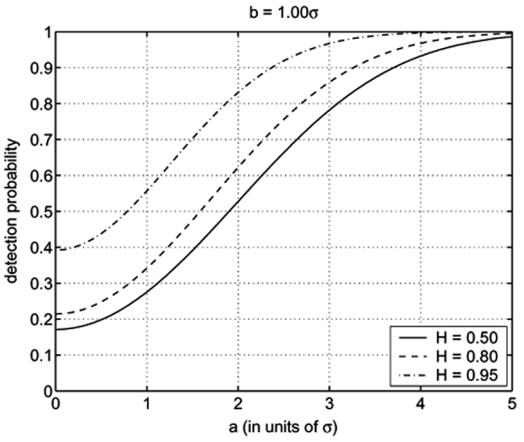


(c)

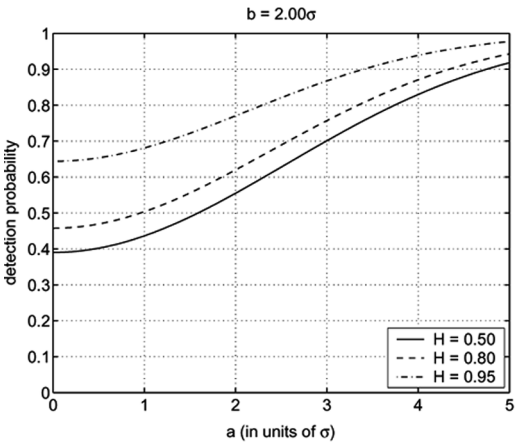
FIGURE 5 The detection probabilities for a fixed false alarm rate, and varying a, b, and H.



(a)



(b)



(c)

FIGURE 6 The detection probabilities for a fixed false alarm rate, and varying a, b, and H.

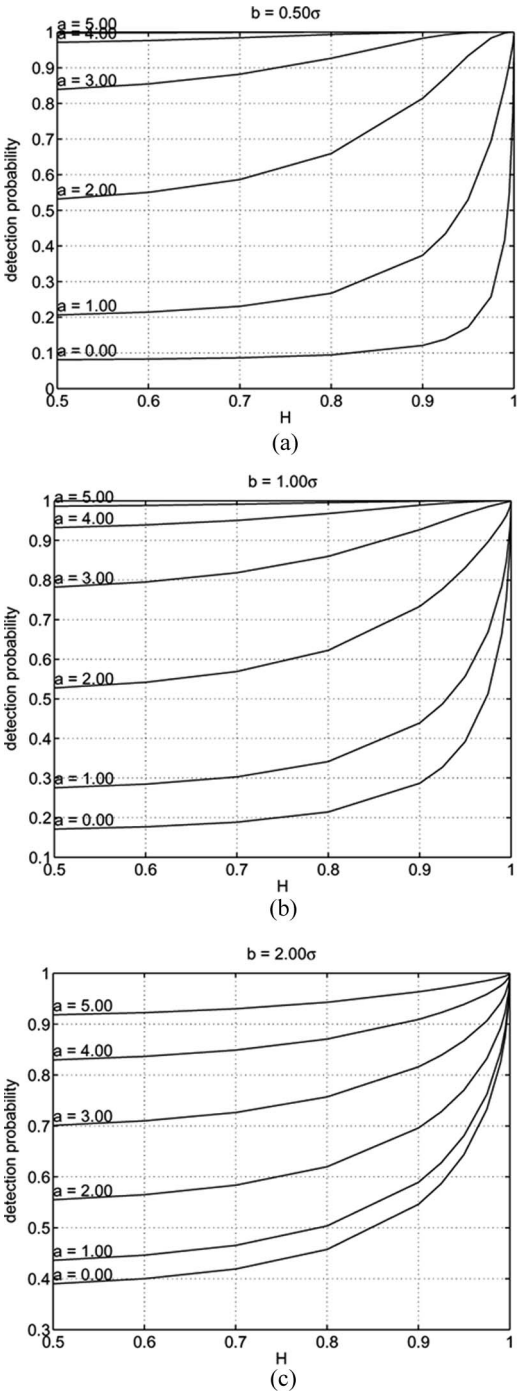


FIGURE 7 The detection probabilities for a fixed false alarm rate, and varying a, b, and H.

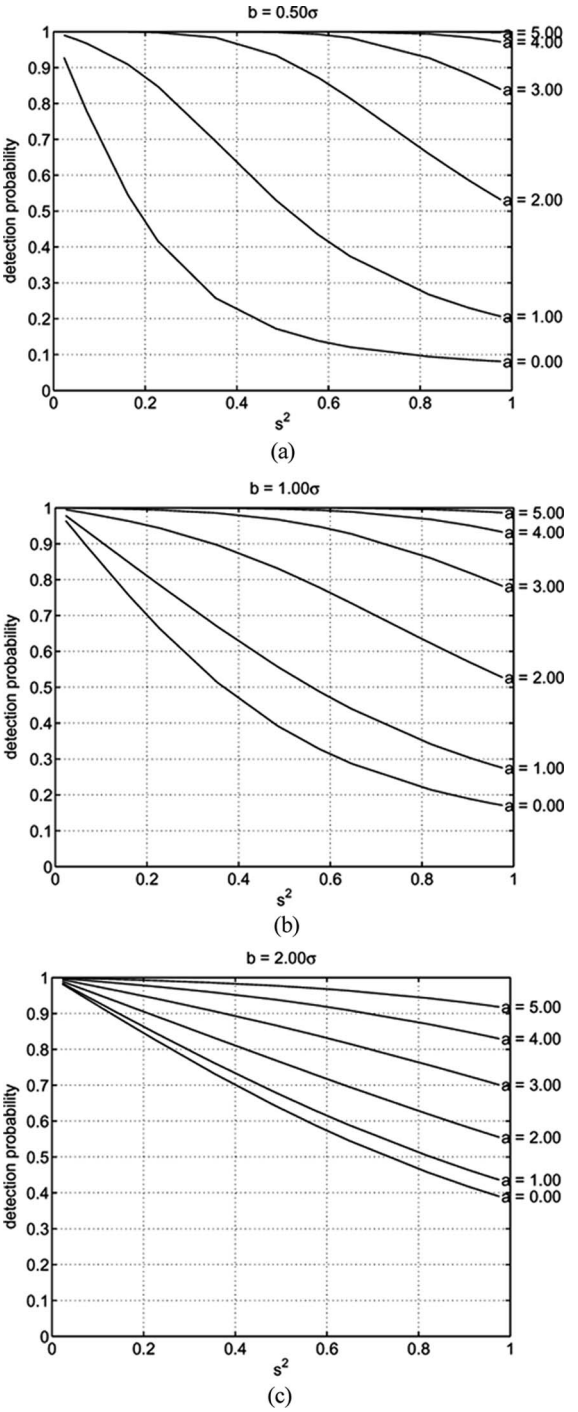


FIGURE 8 The detection probabilities for a fixed false alarm rate, and varying a, b, and $\hat{s}^2$.

magnitude of the effect is different for different parameters a, b, and H. For instance, for small a, the impact only occurs for larger H, or larger b. However, for moderate values of a, there is a noticeable effect even for $H = 0.8$. Furthermore, notice that for all a and b the probability of detection tends to one for large H near one. In the case $a = 0$, $b = 1$, the marginal distribution of the anomalies is identical to that of the normal traffic. Hence, in this case in particular, the only way of detecting the anomalies is by them not matching the correlations in the normal data. It is therefore somewhat remarkable that these correlations are such that (for any values) we could detect the anomalies with high probability, even where they have exactly the same marginal distribution as the normal data.

Figure 8 is a modification of Figure 7, where instead of plotting H on the x-axis, we plot the value of $\hat{s}^2$ that results from a particular value of H. This is not simply another illustration of the same phenomena. The previous sets of graphs have all concerned FGN, with a particular Hurst parameter. We could derive a value for $\hat{s}^2$ from Theorem 5.2.1 for any Gaussian process, for instance, an auto-regressive integrated moving average (ARIMA) process, or a fractional ARIMA process. The value of $\hat{s}^2$ depends on the correlation structure of the process, but this structure can be specified however we desire, for example as in the SRD AR(1) process described in the previous section. Figure 8 shows the detection probability we could expect, given that correlation structure. For smaller $\hat{s}^2$, the detection probability increases, with the size of the increase depending on a and b.

The aforementioned results are for Gaussian processes, but the only special feature of Gaussian processes applied here is that they are entirely modeled by first and second-order moments. We could equally present the above analysis for any known distribution functions. Given a model of traffic and anomalies one could therefore set both the false-alarm and the detection probability to suit a particular application. However, while it is quite possible for us to build a reasonable model of traffic, it is intrinsically hard to do so for anomalies. Anomalies include those things we cannot anticipate, and therefore are hard to model. Hence, in most applications, it makes sense to set the desired false-alarm probability, and then test whether an acceptable detection probability is achieved.

6. FILTER LENGTH

While the above results are useful in determining the detection probabilities for various scenarios, the results are of minimal use in designing an anomaly detection algorithm, for the simple reason that we have no control over the parameters (a, b, H). One parameter that we do control, δ, is fixed by the desired false-alarm probability. So the only parameter available for tuning to improve the detection probability is the

filter length K, or its width M. In this section we investigate the impact of changing this parameter.

In the previous section, we assumed that the filters were long enough that the single anomaly measurement (within a MA window) had negligible impact on the results. However, for filters shorter than around $M = 10$, this is clearly not the case. Anomalies have two impacts on the above results:

1. a bias of the mean of $\widehat{X}_i$ towards the anomaly,
2. an increase in the value of $\hat{s}^2$.

The first distortion is likely to be much more severe than the second, as we use the entire data set in evaluation of $\hat{s}^2$, but only a finite window of data in evaluating $\widehat{X}_i$, and so the anomaly introduces a larger bias to this measurement.

One method for dealing with these distortions is an iterative approach to anomaly detection. For instance, we could initially apply the anomaly detection, and then eliminate any anomalies in a second round of anomaly detection. By eliminating the larger, more obvious anomalies when computing statistics for the second round of anomaly detection, we reduce the biases introduced into $\widehat{X}_i$ and $\hat{s}^2$, and thereby achieve results closer to the theoretical results. However, such an approach requires two passes over the data, which is often undesirable.

Hence, in evaluating the impact of filter length, we should consider the additional biases in estimates $\widehat{X}_i$. For anomalies consisting of a single point, drawn (as before) from a Gaussian distribution, we can model $\widehat{X}_i$ as having $K - 1$ values drawn from the original distribution, and 1 value drawn from the anomaly distribution. The expected value will therefore be

$$E[\widehat{X}_i] = \frac{K-1}{K} E[X_i] + \frac{1}{K} E[Y_j] = \mu + \frac{a}{K}. \tag{25}$$

Notice that the mean estimate is pulled towards the anomaly by an amount determined by how far the anomalies lie from the normal process. Given this result, the separation between the mean estimates, and the anomalous values is no longer a, but really $A = E[Y_i] - E[\widehat{X}_i] = (1 - 1/K)a$, and the new probability of detection is $P(T; A, \phi)$. Figure 9(a) shows this probability of detection (for $a = 4$, $b = 1$, and $\delta = 1.96$) over a range of values of H and M. The case $H = 0.5$ follows conventional intuition that a longer filter is better (i.e., has a higher probability of detection). However, when $H > 0.6$, we notice that the detection probability curves have a distinct maximum for small values of M. This shows that the optimal window length for LRD traffic is small, unlike that for a SRD process. Figure 9(b) shows the optimal value over a range of values of H (typical values for traffic are between 0.75 and 0.9).

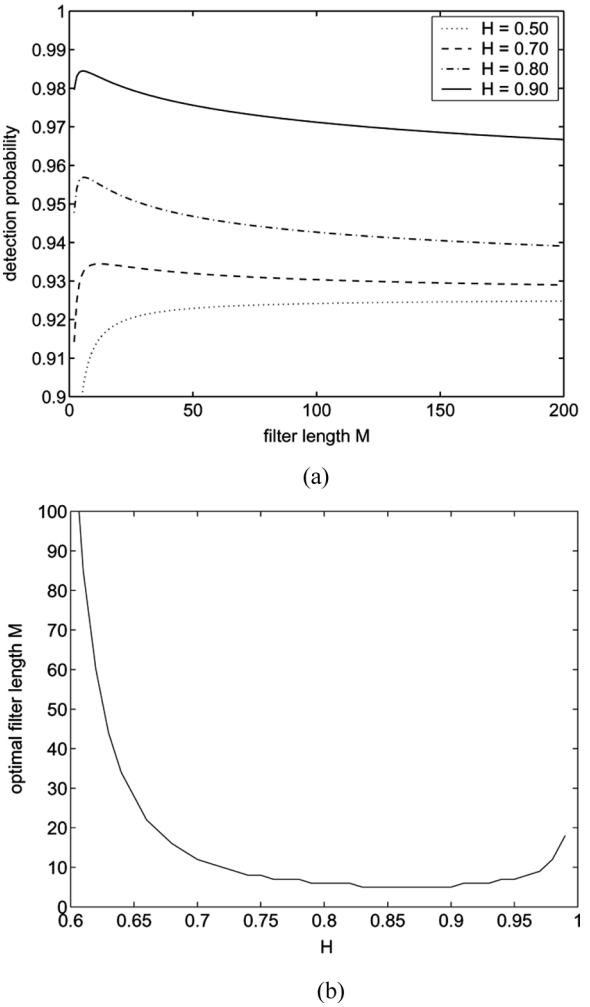


FIGURE 9 The impact of filter length M ($a = 4$, $b = 1$, $\delta = 1.96$).

The aforementioned results reflect a tradeoff between variance and bias in an estimator. Using shorter filters results in a larger variance for the estimator. However, the bias in the LRD case, towards the data points counters this variance increase resulting in more accurate detection. The cut-off between the LRD and SRD cases is not simply $H > 0.5$ because there are also complicating factors in the bias introduced by the anomaly itself, and so we observe the effect for values of H above around 0.6.

The important conclusion is that, where typically we might have chosen the longest filter possible given the constraints of nonstationarity and computational complexity, now we can simply choose a short filter, in the confidence that this will produce (near) optimal results. Typical Hurst parameter values fall in the range $[0.75, 0.90]$, for which the optimal values

of M are between 5–8, resulting in filters of length $K = 11$ and 17. Given that we may not know in advance the exact value of H, it appears from examining the curves in Figure 9(a) that is better to err conservatively on the side of a longer filter (as the curves drop away less precipitously), and so these results suggest that filters of $K = 17$ taps would be ideal for anomaly detection in this context. This is an interesting length—for SNMP data sampled at 5-minute intervals this filter would be 85 minutes long. This is within the bounds over which typical Internet data can be assumed to be stationary. It is certainly within the range where we consider that the traffic can be well approximated by a linear process. Hence, we should not need to build more complicated seasonal models as required by the Holt-Winters algorithm and others.

7. CONCLUSION

This article has demonstrated a somewhat counterintuitive result; namely, that LRD helps anomaly detection. It is counterintuitive, because in most of the literature on LRD, it is seen as a detrimental property, whereas here it is positive. However, we demonstrate this effect in simulations, and use theory to explain it. The practical conclusion of the work is that we show that shorter MA filters (used in anomaly detection) will perform better for LRD traffic. Shorter filters make anomaly detection algorithms easier to design and implement, and more robust to nonstationarity in the data.

The intuitive rationale for the result—that LRD processes are in some sense smoother than SRD process, and therefore anomalies "stick out" more—would apply to most approaches to anomaly detection. Hence, we expect the effect described here to apply to all anomaly detection methods, though demonstrating this is so is left for future work. A further extension of the work is to note that the aforementioned results might equally be applied to fine grained packet counts (obtained from a packet header trace) on a time-scale of milliseconds, or to measurements of performance, for instance, as might be collected by a series of active performance probes[1,13]. In the detection of anomalies in backbone traffic (as measured by SNMP measurements), the Gaussianity assumptions appear to be a reasonable approximation[17], but in other contexts, the particular form of the marginal distribution of data and anomalies might vary[2]. It is reasonable to expect that the qualitative features previously noted would be observed and quantitative results could be derived given suitable models.

Note also, that in the anomaly detection algorithms above, no explicit account is taken of the correlations of the process. We do not attempt to exploit the correlations explicitly to infer anomalies. The results arise purely for standard anomaly detection applied to a correlated time-series.

It may well be possible to do better by taking explicit care of these anomalies, but we leave this for future work.

ACKNOWLEDGMENT

I would like to thank the anonymous reviewers for many helpful comments, and acknowledge the support of the ARC through discovery grant DP0665427.

REFERENCES

1. Almes, G.; Kalidindi, S.; Zekauskas, M. A one-way delay metric for IPPM. *IETF. IP Performance Metrics, Request for Comments* **1999**, 2679.
2. Andren, J.; Hilding, M.; Veitch, D. Understanding end-to-end Internet traffic dynamics. In *IEEE GLOBECOM 98*; Sydney, Australia, 1998.
3. Barford, P.; Kline, J.; Plonka, D.; Ron, A. A signal analysis of network traffic anomalies. In *ACM SIGCOMM Internet Measurement Workshop*; Marseilles, France, November 2002.
4. Beran, J. *Statistics for Long-Memory Processes*; Chapman and Hall, New York, 1994.
5. Brutag, J.D. Aberrant behavior detection and control in time series for network monitoring. In *Proceedings of the 14th Systems Administration Conference (LISA 2000)*; New Orleans, LA, December 2000.
6. Erramilli, A.; Narayan, O.; Willinger, W. Experimental queueing analysis with long-range dependent packet traffic. IEEE/ACM Transactions on Networking **1996**, *4*, 209–223.
7. Grossglauser, M.; Bolot, J.-C. On the relevance of long-range dependence in network traffic. IEEE/ACM Transactions on Networking **1999**, *7*, 629–640.
8. Krishnamurthy, B.; Sen, S.; Zhang, Y.; Chen, Y. Sketch-based change detection: Methods, evaluation, and applications. In *ACM SIGCOMM Internet Measurement Conference*; Miami, FL, October 2003.
9. Leland, W.E.; Taqqu, M.S.; Willinger, W.; Wilson, D.V. On the self-similar nature of Ethernet traffic. IEEE/ACM Transactions on Networking **1994**, *2*, 1–15.
10. Mauro, D.R.; Schmidt, K.J. *Essential SNMP*. O'Reilly & Associates, Inc.: USA, 2001.
11. Norros, I. A storage model with self-similar input. Queueing Systems **1994**, *16*, 387—396.
12. Paxson, V.; Floyd, S. Wide-area traffic: The failure of Poisson modeling. IEEE/ACM Transactions on Networking **1995**, *3*, 226–244. Available at http://www.aciri.org/floyd/papers.html.
13. Paxson, V.; Mahdavi, J.; Adams, A.; Mathis, M. An architecture for large-scale internet measurement. IEEE Communications Magazine **1998**, 48–54.
14. Ringberg, H.; Roughan, M.; Rexford, J. The need for simulation in evaluating anomaly detectors. ACM SIGCOMM Computer Communication Review **2008**, *38*, 55—59.
15. Roughan, M.; Gottlieb, J. Large-scale measurement and modeling of backbone Internet traffic. In *SPIE ITCom*; Boston, 2002.
16. Roughan, M.; Greenberg, A.; Kalmanek, C.; Rumsewicz, M.; Yates, J.; Zhang, Y. Experience in measuring internet backbone traffic variability: Models, metrics, measurements and meaning. In *Providing Quality of Service in Heterogeneous Environments*; Charzinski, P.-G.J., Lehnert, R. Eds., Elsevier Science: Amsterdam, 2003; 221–230.
17. Roughan, M.; Griffin, T.; Mao, M.; Greenberg, A.; Freeman, B. Combining routing and traffic data for detection of IP forwarding anomalies. In *ACM SIGMETRICS* 2004; 416–417.
18. Roughan, M.; Griffin, T.; Mao, M.; Greenberg, A.; Freeman, B. IP forwarding anomalies and improving their detection using multiple data sources. In *ACM SIGCOMM Workshop on Network Troubleshooting*; Portland, OR, September 2004; 307–312.
19. Roughan, M.; Veitch, D. Measuring long-range dependence under changing traffic conditions. In *IEEE INFOCOM'99*; IEEE Computer Society Press: Los Alamitos, CA, New York, March 1999.
20. Samorodnitsky, G.; Taqqu, M.S. *Stable Non-Gaussian Random Processes*; Chapman and Hall: USA, 1994.
21. Veitch, D.; Abry, P. A wavelet based joint estimator of the parameters of long-range dependence. IEEE Transactions on Information Theory Special Issue on "Multiscale Statistical Signal Analysis and Its Applications" **1999**, *45*, 878–897.