Anomaly detection in communication networks using wavelets

V.Alarcon-Aquino and J.A.Barria

Abstract: An algorithm is proposed for network anomaly detection based on the undecimated discrete wavelet transform and Bayesian analysis. The proposed algorithm checks the wavelet coefficients across resolution levels, and locates smooth and abrupt changes in variance and frequency in the given time series, by using the wavelet coefficients at these levels. The unknown variance of the wavelet coefficients is considered as a stochastic nuisance parameter. Marginalisation is then used to remove this nuisance parameter by using three different priors: flat, Jeffreys' and the inverse Wishart distribution (scalar case). The different versions of the proposed algorithm are evaluated using synthetic data, and compared with autoregressive models and thresholding techniques. The proposed algorithm is applied to monitor events in a Dial Internet Protocol service. The results show that the proposed algorithm is able to identify the presence of abnormal network behaviours in advance of reported network anomalies.

1 Introduction

Early network anomaly detection has become critical to service providers in their commitment to maintain a given service level agreement. Hence, there has been much interest in studying adaptive event detection schemes with application to communication networks. Studies on TCP/IP network anomaly detection include a sequential generalised likelihood ratio test using auto-regressive (AR) models [1–3] and constant threshold schemes [4, 5]. Anomaly detection in Ethernet segments has also been studied [6], where detection is achieved via a fault feature vector of known faults. An approach has been proposed [7] that uses adaptive thresholds for proactive network/service anomaly detection in transaction-oriented wide area networks. A review of different reactive fault detection schemes for alarm correlation has been presented elsewhere [8].

It is well known that to truly detect anomalies, systems should consider the time-varying nature of the data, and the detection thresholding techniques should be able to adapt to the changing environment [6, 7, 9]. Episodes of abnormal network behaviours are, in most cases, reflected in statistical departures from the normal pattern [1–3, 7, 9]. Abnormal network behaviours can then be detected by observing the statistical behaviours of target network metrics. They can be traced by correlating events among the different network metrics being monitored.

Some approaches [1–4, 7] are most suitable if the data contain contributions at fixed resolution or scale in time and/or frequency. Unfortunately, data from almost all practical network processes are multi-scale in nature, due to events occurring at different points in time and frequency [10]. The work reported in this paper investigates the viability and usability of discrete wavelet transforms in this area. Owing to the inherent multi-scale nature of wavelet transforms, a wavelet-based analysis seems more appropriate for data containing events whose behaviour changes over time and frequency. Furthermore, since wavelets are able to adjust their scale to the nature of the signal features, subtle changes (e.g. in variance, frequency or both) can be detected at different resolution levels. Wavelet transforms have found applications in areas such as signal denoising; signal and image compression; signal estimation; partial differential equations; seismic and geophysical signals; and biomedical signals (see e.g. [11, 12]). However, to the best of our knowledge, there has so far been no reported work on network anomaly detection using wavelet transforms. The proposed algorithm checks the wavelet coefficients across resolution levels, and locates smooth and abrupt changes in variance and frequency in the given time series by using the wavelet coefficients at these levels. This method has the advantage of adapting locally to the features of the signal. By contrast, standard network anomaly detectors [1–4, 7] analyse with a fixed scale.

In the proposed wavelet-based algorithm, we consider the unknown variances of the wavelet coefficients as stochastic nuisance parameters. Marginalisation is used to eliminate these nuisance parameters from the wavelet domain using three different priors: flat, Jeffreys' and the inverse Wishart distribution. (Note that marginalisation is well known in estimation theory and has been applied to detect additive abrupt state vector changes in a linear state-space model, for example [14].) A survey of prior distributions is found elsewhere [13]. Fig. 1 illustrates the components of the proposed wavelet-based network anomaly detector.

2 Undecimated discrete wavelet transform

The wavelet transform of $f(t) \in L^2(R)$ involves the computation of the inner products of the signal and a family of

IEE Proceedings online no. 20010659

DOI: 10.1049/ip-com:20010659

Paper first received 8th March and in revised form 10th August 2001

The authors are with the Imperial College of Science, Technology and Medicine, Department of Electrical and Electronic Engineering, Exhibition Road, London SW7 2BT, UK

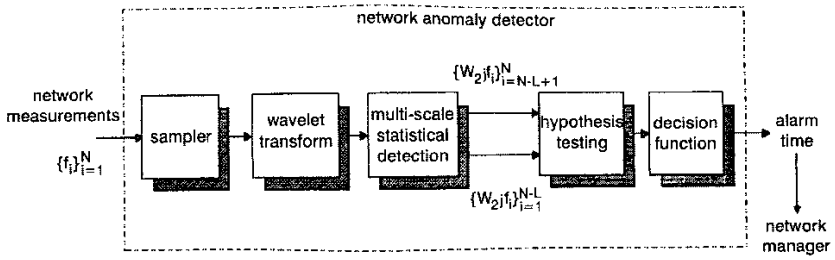

Fig. 1 *Architecture of proposed wavelet-based network anomaly detector*
$(2^j)_{1\le j\le J}$ = dyadic sequence, J = number of scales, $W_{2^j}f_i$ = wavelet coefficients, L = length of sliding test window

wavelets; L^2 denotes the Hilbert space of square integrable one-dimensional functions and R is the set of real numbers. A family of wavelets $\psi_{u,s}(t)$, $u, s \in R$ is obtained by translation, u, and dilation, s, (also known as scale) operations of the mother wavelet $\psi(t)$ [15].

To construct the undecimated discrete wavelet transform (UDWT), the scale s is discretised but not the translation parameter u. The scale is sampled along the dyadic sequence $(2^j)_{j\in Z}$, where Z is the set of integer numbers. The wavelet transform of $f(t)$ at the scale $s = (2^j)_{1\le j\le J}$ and at the position t has been defined [16] by the convolution product $W_{2^j}f(t) = f * \psi_{2^j}(t)$; J represents the number of scales or resolutions levels and $\psi_{2^j}(t) = (1/2^j)\psi(t/2^j)$ is the dilation of the basic wavelet by a factor of 2^j. Let S_{2^j} denote the smoothing operator defined by $S_{2^j}f(t) = f * \phi_{2^j}(t)$, where $\phi_{2^j}(t) = (1/2^j)\phi(t/2^j)$. The larger the scale 2^j, the more details of $f(t)$ are removed by S_{2^j}. This means that, at each scale 2^j, the UDWT of a discrete signal $(S_{2^0}f_i)_{i\in Z} = f_i$ decomposes $S_{2^j}f_i$ into $S_{2^{j+1}}f_i$ and $W_{2^{j+1}}f_i$:

$$W_{2^{j+1}}f_i = \sum_{k\in Z} g_k S_{2^j} f(i - k2^j)$$

$$\text{and } S_{2^{j+1}}f_i = \sum_{k\in Z} h_k S_{2^j} f(i - k2^j) \qquad (1)$$

where g_k and h_k are the coefficients of the wavelet ψ and scaling ϕ functions, respectively. Note that the conventional discrete wavelet transform (CDWT) can detect abrupt changes in a time series; however, it can introduce ambiguities in the time domain due to the decimation process that needs to be applied at the output. In contrast to the CDWT, the UDWT is translation-invariant in the sense that it preserves regularity information at each point in time for each scale, and it may be computed for an arbitrary length time series. This translation-invariant property allows alignment of events in a multi-resolution analysis with respect to the original time series. Further details of these wavelet transforms and a comparison between these transforms to locate transients events are found elsewhere [15–18].

2.1 Scale choice

Since the scale choice depends on the wavelet itself, the number of scales J is chosen according to the overall energy displayed at each scale. For example, Fig. 2 shows the frequency response magnitude of wavelet coefficients $W_{2^j}f_i$ using a quadratic spline wavelet. (The properties and coefficients of this wavelet can be found elsewhere [16].) From Fig. 2, it can be seen that the scales $j = 1, 2$ contain the high-frequency components of the signal, whereas the scales $j = 3, 4$ contain the low-frequency components of the signal. Therefore, scales $j = 1, 2$ are selected to carry out the work reported here since they contain most of the signal energy.

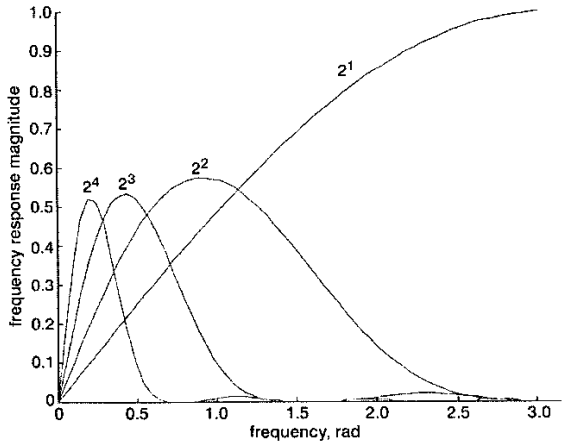

Fig. 2 *Frequency response magnitude of wavelet coefficient $W_{2^j}f_i$ using quadratic spline wavelet at scales 2^1 to 2^4*

3 Multi-scale statistical detection algorithm

3.1 Problem statement and assumptions

Let the time series of network measurements f_i be modelled by $f_i = \theta_i + e_i$, $e_i \sim N(0, \sigma_i^2)$, for $i = 1, ..., N$, where θ_i represents the parameters of the signal model and σ_i^2 denotes the unknown changing variance. Thus, the likelihood for data $f^N = f_1, f_2, ..., f_N$, given the parameters θ_i and σ_i^2, is denoted by $p(f^N \mid \theta_i, \sigma_i^2)$.

To determine whether a change has occurred at time t_0, the proposed algorithm uses a sliding window approach. This approach considers two windows; a test window and a reference window, from which two models are derived. These two models are then used for the detection algorithm (see Section 3.3) to perform the sequential decision process. In each time interval or window, the wavelet coefficients are regarded as zero-mean Gaussian stationary process $\{W_{2^j}f_i\}_{i=1}^N \sim N(0, \tau_j^2)$, for $1 \le j \le J$, where τ_j^2 represents the unknown variance of wavelet coefficients at each scale. The test window based on data $W_{2^j}f_{N-L+1}, ..., W_{2^j}f_N$ of length L is compared to a growing reference window based on all previous data $W_{2^j}f_1, W_{2^j}f_2, ..., W_{2^j}f_{N-L}$, or larger L, to determine whether both models are generated by the same or different distributions. Further details on sliding window approaches can be found in [14, 19] and references therein.

Note that, by using eqn. 1, the model for wavelet coefficients $\{W_{2^j}f_i\}_{i=1}^N \sim N(0, \tau_j^2)$ has zero-mean because $E\{W_{2^j}f_i\} = \Sigma_k g_k E\{S_{2^0}f(i-k)\} = \mu_f \Sigma_k g_k = 0$, where μ_f is the mean of network measurements. By using the fact that a wavelet filter must sum to zero [17, 18], i.e. $\Sigma_k g_k = 0$, the result follows. The wavelet coefficients $W_{2^j}f_i$ were evaluated using the auto-correlation function (ACF) of the different network metrics described in Section 5. Exploring the time-lagged properties of the wavelet coefficients, the correlograms showed no significant time-serial dependencies in all

network metrics being monitored. Fig. 3 shows the correlogram of only one of the network metrics. All other network metrics considered have a similar behaviour. Hence, for the study reported in this paper, these wavelet coefficients are assumed to be independently distributed random variables. Consequently, the likelihood can be computed as a product of the likelihoods before and after change. This means that the likelihood of the change point having taken place at $t_0 = N - L$ (where L is the length of a sliding test window) is given by

$$p\left(W_{2^j}f_1, W_{2^j}f_2, \ldots, W_{2^j}f_N | t_0, \tau_{j,1}^2, \tau_{j,2}^2\right) = \prod_{i=1}^{t_0} p\left(W_{2^j}f_i | \tau_{j,1}^2\right) \prod_{i=t_0+1}^{N} p\left(W_{2^j}f_i | \tau_{j,2}^2\right), \quad \text{for } 1 \le j \le J \quad (2)$$

Note that, since multi-scale decomposition of a signal is equivalent to band-pass filtering, and modifications in the process will mainly lead to variance changes after wavelet decomposition, the approach studied in this paper is free of model selection parameters. Hence, the problem is reduced to the estimation of unknown variances. The unknown variances $\tau_{j,1}^2$ for $i \le t_0$ and $\tau_{j,2}^2$ for $t_0 < i \le N$ are considered nuisance parameters and are removed in the wavelet domain, after integrating with respect to a prior distribution. The effect of the priors at different resolution levels J is addressed in Section 4.

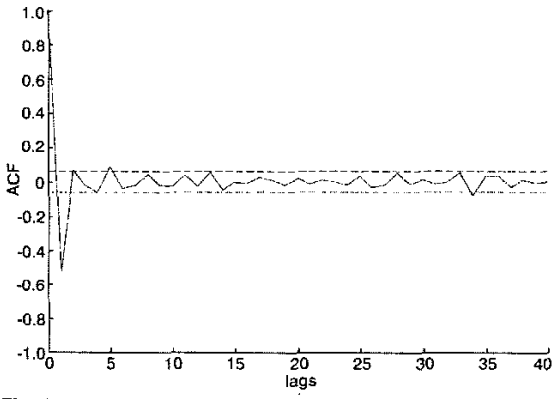

Fig. 3 *Auto-correlation of wavelet coefficients at scale j = 1 of Web_Latency metric*
– – – 95% confidence limits of independent and identically distributed (IID) process
Each lag corresponds to sampling period (see Section 5 for detailed definition of network metrics)

3.2 Choice of priors

Three different priors are analysed to determine the sensitivity of the proposed algorithm: the flat prior, Jeffreys' prior and the inverse Wishart distribution.

The flat or non-informative prior $p(\tau_j^2) = \kappa$, where κ is a constant, has been widely suggested. The problem with non-informative priors is that they often have improper density functions, i.e. their integral is not finite. However, their use is motivated by the fact that, in many cases, the posterior distribution is still proper [13, 14, 20].

Another improper prior can be obtained by applying Jeffreys' rule, which allows us to find prior distributions that are invariant under estimation. In the normal model case, Jeffreys' prior is $p(\tau_j^2) = \kappa/\tau_j^2$, $0 < \tau_j^2 < \infty$. Further details on Jeffreys' priors can be found elsewhere [13, 20].

Prior distributions with density functions similar to the likelihood function have also been suggested. When the likelihood function belongs to an exponential family of probability distributions, an acceptable criterion is to choose the prior to be conjugated to the likelihood function so that the posterior distribution belongs to the exponential family. Since the conjugate prior may have the kernel similar to the likelihood, it follows that the conjugate prior for the normal model is the inverse gamma distribution, or the inverse Wishart distribution in the scalar case [13, 14, 20]. The inverse Wishart distribution $Wi^{-1}(v, S)$ is given by

$$p\left(\tau_j^2\right) = \frac{S^{v/2}}{2^{v/2}\Gamma(v/2)} \left(\tau_j^2\right)^{-(v+2)/2} e^{-S/2\tau_j^2} \quad (3)$$

where Γ is the gamma function defined by $\Gamma(v) = \int_0^\infty x^{v-1}e^{-x}dx$, $v > 0$. This distribution has mean $E(\tau_j^2) = S/(v - 2)$ and variance $Var(\tau_j^2) = 2S^2/[(v - 4)(v - 2)^2]$.

3.3 Proposed algorithm in wavelet domain

As the core of our derivations are the so-called 'nuisance parameters', these nuisance parameters can be estimated or marginalised. The usual likelihood-ratio testing procedure involves computing the maximum likelihood estimates of the unknown nuisance parameters. In contrast, the concept of marginalisation is to assign a prior distribution to the unknown nuisance parameter and eliminate it from the analysis [14, 20]. This means that the likelihoods are marginal, in the sense that they are obtained after integrating out the nuisance parameters. The unknown change points are then estimated by comparing the posterior probabilities computed using Bayes' theorem [20]. The posterior probabilities associated with the hypotheses can be written as

$$p(H_q|W_{2^j}f_i) \propto (W_{2^j}f_i|H_q)p(H_q), \quad i = 1, \ldots, N \quad q = 0, 1, \quad \text{for } 1 \le j \le J \quad (4)$$

where $\propto$ denotes a relationship of proportionality and $W_{2^j}f_i$ represents the wavelet coefficients. The posterior probability associated with H_0 is obtained by using eqn. 4 and after integrating with respect to a prior distribution of the nuisance parameter:

$$p(H_0|W_{2^j}f_i) \propto p(H_0) \int_0^\infty p(W_{2^j}f_i|t_0, \tau_j^2)p(\tau_j^2)d\tau_j^2, \quad i = 1, \ldots, N, \quad \text{for } 1 \le j \le J \quad (5)$$

where τ_j^2 represents the unknown variance of wavelet coefficients, t_0 is the unknown change point and $p(\tau_j^2)$ is the prior to be considered. The prior probabilities associated with the hypotheses are $p(H_0) = \pi$ for $\pi \in (0, 1)$ and $p(H_1) = 1 - \pi$, where $p(H_0) + p(H_1) = 1$. Therefore, the prior probability of having a change point $p(H_q)$ can be incorporated into the wavelet-based statistical detection scheme.

The hypothesis to be tested is H_0 : $\text{var}\{W_{2^j}f_1\} = \ldots = \text{var}\{W_{2^j}f_N\}$, and the alternative hypothesis is H_1 : $\text{var}\{W_{2^j}f_1\} = \ldots = \text{var}\{W_{2^j}f_{t_0}\} \neq \text{var}\{W_{2^j}f_{t_0+1}\} = \ldots = \text{var}\{W_{2^j}f_N\}$ for $1 \le j \le J$. Since modifications in the process will mainly lead to variance changes after wavelet decomposition, the change points in the time series are then estimated from the posterior probabilities using the decision function $\delta = [\log p(H_0 \mid W_{2^j}f_i)/\log p(H_1 \mid W_{2^j}f_i)] > 1$, which provides the basis for choosing between H_0 and H_1.

To reduce the mathematical expressions, the following notations are introduced:

$$A_1 = \frac{1}{2}\sum_{i=1}^{N-L}(W_{2^j}f_i)^2, \; A_2 = \frac{1}{2}\sum_{i=N-L+1}^{N}(W_{2^j}f_i)^2,$$

$$A = A_1 + A_2 = \frac{1}{2}\sum_{i=1}^{N}(W_{2^j}f_i)^2$$

Note that A_1, A_2 and A contain all the wavelet coefficients. To obtain the posterior probability associated with the hypothesis H_0, consider the inverse Wishart distribution as prior, $p(\tau_j^2) \sim Wi^{-1}(v, S)$ given by eqn. 3, on the variance of wavelet coefficients:

$$p(H_0|W_{2^j} f_i) = p(H_0) \int_0^\infty (2\pi\tau_j^2)^{-N/2} e^{-\frac{A}{\tau_j^2}} \times p(\tau_j^2) d\tau_j^2, \quad \text{for } 1 \le j \le J \quad (6)$$

$$p(H_0|W_{2^j} f_i) = \frac{p(H_0)(2\pi)^{-N/2} S^{v/2}}{2^{v/2}\Gamma(v/2)} \times \int_0^\infty (\tau_j^2)^{-N/2} e^{-\frac{A}{\tau_j^2}} (\tau_j^2)^{-\frac{v+2}{2}} e^{-\frac{S}{2\tau_j^2}} d\tau_j^2, \quad \text{for } 1 \le j \le J \quad (7)$$

The integral is solved by using

$$\int_0^\infty x^{-(v+2)/2} e^{-\frac{S}{2x}} dx = \left[2^{v/2}\Gamma(v/2)\right]/S^{v/2} \quad (8)$$

to give

$$p(H_0|W_{2^j} f_i) = \frac{p(H_0) S^{v/2}(2\pi)^{-N/2}\Gamma\left((N+v)/2\right)}{2^{v/2}\Gamma(v/2)(A+S/2)^{(N+v)/2}}, \quad \text{for } 1 \le j \le J \quad (9)$$

(Note that this result is easily obtained by using the change of variable rule [20]; by setting $\lambda = 1/x$, the integral becomes $\int_0^\infty \lambda^{(v/2)-1} e^{-S\lambda/2} d\lambda = \Gamma(v/2)/(S/2)^{v/2}$, where Γ denotes the gamma function.) The second hypothesis verifies whether a change has occurred on the variance of wavelet coefficients. Recall that the unknown changing variances are $\tau_{j,1}^2$ for $i \le t_0$ and $\tau_{j,2}^2$ for $t_0 < i \le N$. Based on the assumptions in Section 3.1 and using eqn. 2, the data can be split into two integrals:

$$p(H_1|W_{2^j} f_i) \propto p(H_1) \int_{\tau_{j,1}^2} p\left(W_{2^j} f_1, \ldots, W_{2^j} f_{N-L} | t_0, \tau_{j,1}^2\right) p(\tau_{j,1}^2) d\tau_{j,1}^2 \times \int_{\tau_{j,2}^2} p\left(W_{2^j} f_{N-L+1}, \ldots, W_{2^j} f_N | t_0, \tau_{j,2}^2\right) p(\tau_{j,2}^2) d\tau_{j,2}^2, \quad \text{for } 1 \le j \le J \quad (10)$$

where $t_0 = N - L$ is the unknown change point. Reconsidering the inverse Wishart distribution as prior, $p(\tau_j^2) \sim Wi^{-1}(v, S)$, the integrals can be solved in the same way as eqn. 6, to obtain

$$p(H_1|W_{2^j} f_i) = \frac{p(H_1) S^v (2\pi)^{-N/2}\Gamma\left((N-L+v)/2\right) \times \Gamma\left((L+v)/2\right)(A_2 + S/2)^{-(L+v)/2}}{2^v \Gamma(v/2)\Gamma(v/2)(A_1 + S/2)^{(N-L+v)/2}}, \quad \text{for } 1 \le j \le J \quad (11)$$

where v and S are the hyperparameters of the inverse Wishart distribution (see eqn. 3) used for τ_j^2.

Now let us consider Jeffreys' prior $p(\tau_j^2) = 1/\tau_j^2$ on the variance of wavelet coefficients for the hypothesis H_0:

$$p(H_0|W_{2^j} f_i) = p(H_0) \int_0^\infty (2\pi\tau_j^2)^{-N/2} e^{-\frac{A}{\tau_j^2}} \times p(\tau_j^2) d\tau_j^2, \quad \text{for } 1 \le j \le J \quad (12)$$

The posterior probability associated with H_0 is derived as in eqn. 6, where the integrand is rewritten into the density function of the inverse Wishart distribution [14]:

$$p(H_0|W_{2^j} f_i) = \frac{p(H_0)(2\pi)^{-N/2}\Gamma(N/2)}{A^{N/2}} \times \int_0^\infty \frac{(2A)^{N/2} e^{-\frac{A}{\tau_j^2}}}{(\tau_j^2)^{(N+2)/2} 2^{N/2}\Gamma(N/2)} d\tau_j^2, \quad \text{for } 1 \le j \le J \quad (13)$$

The integrand in eqn. 13 is recognised as the inverse Wishart distribution, which integrates to 1. Therefore

$$p(H_0|W_{2^j} f_i) = \frac{p(H_0)(2\pi)^{-N/2}\Gamma(N/2)}{A^{N/2}}, \quad \text{for } 1 \le j \le J \quad (14)$$

The second hypothesis under Jeffreys' prior on the variance of wavelet coefficients can be derived using eqn. 10. The integrals can be solved in the same way as eqn. 12, to obtain

$$p(H_1|W_{2^j} f_i) = \frac{p(H_1)(2\pi)^{-N/2}\Gamma\left((N-L)/2\right)\Gamma(L/2)}{A_1^{(N-L)/2} A_2^{L/2}}, \quad \text{for } 1 \le j \le J \quad (15)$$

Finally, the posterior probabilities using a flat prior $p(\tau_j^2) = 1$, on the variance of wavelet coefficients for both hypotheses, can be obtained in the same way as eqn. 12. Solving as in eqn. 6 gives $v = N - 2$ and $S = 2A$. Thus, we obtain

$$p(H_0|W_{2^j} f_i) = \frac{p(H_0)(2\pi)^{-N/2}\Gamma\left((N-2)/2\right)}{A^{(N-2)/2}} \times \int_0^\infty \frac{(2A)^{(N-2)/2} e^{-\frac{A}{\tau_j^2}}}{(\tau_j^2)^{N/2} 2^{(N-2)/2}\Gamma\left((N-2)/2\right)} d\tau_j^2, \quad \text{for } 1 \le j \le J \quad (16)$$

Since the integrand is recognised as the inverse Wishart distribution, we obtain

$$p(H_0|W_{2^j} f_i) = \frac{p(H_0)(2\pi)^{-N/2}\Gamma\left((N-2)/2\right)}{A^{(N-2)/2}}, \quad \text{for } 1 \le j \le J \quad (17)$$

The second hypothesis under a flat prior, on the variance of wavelet coefficients, can be derived using eqn. 10, to obtain

$$p(H_1|W_{2^j} f_i) = \frac{p(H_1)(2\pi)^{-N/2}\Gamma\left((N-L-2)/2\right)\Gamma\left((L-2)/2\right)}{A_1^{(N-L-2)/2} A_2^{(L-2)/2}}, \quad \text{for } 1 \le j \le J \quad (18)$$

Note that, since the unknown changing variance has been marginalised, eqns. 11, 15 and 18 are just a function of the change point $t_0 = N - L$. To eliminate the gamma functions Γ from eqns. 9, 11, 14, 15, 17 and 18, Stirling's approximation formula can be used $\Gamma(v+1) \approx \sqrt{(2\pi v)}(v/e)^v$ [20].

4 Simulation results

In this Section, simulated data are generated by switching auto-regressive (AR) filters with time-invariant parameters, which are driven by a Gaussian white noise source [19]. A second-order model is chosen to allow a small variation in the AR parameters. This set of data series simulates differ-

ent kinds of changes, such as abrupt and smooth changes in the AR parameters. The boundaries of each segment are indicated by the values of the parameters, and the boundary positions are obtained by switching the outputs of these filters. A segment is then characterised by a set of AR parameters that remain fixed for a certain time interval.

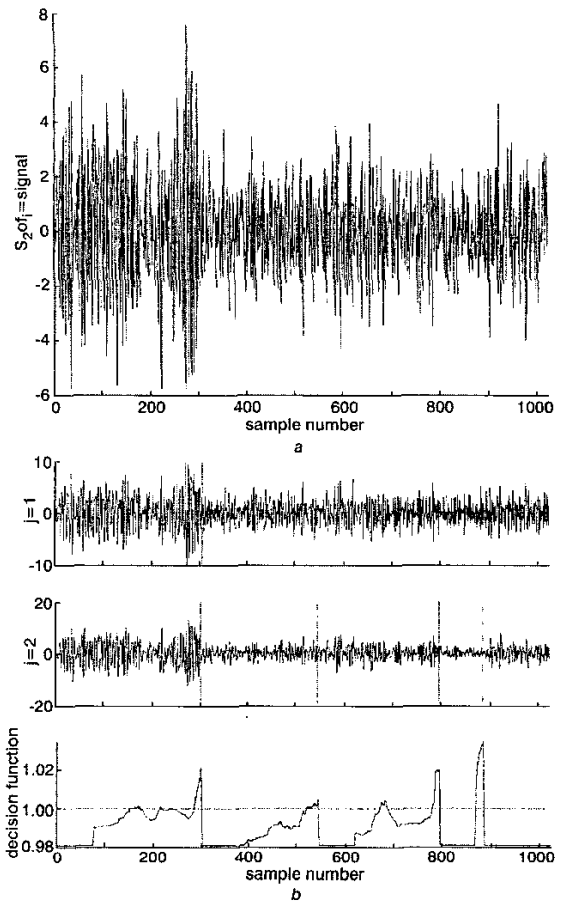

Fig. 4 *Realisation of AR(2) process and wavelet decomposition*
a Realisation of AR(2) process with parameter vector changes from (–0.9, 0.9) to (–0.6, 0.6) at $N = 300$ and from (–0.6, 0.6) to (–0.67, 0.67) at $N = 500$. It also includes a small ramp from $N = 800$ to $N = 899$
b Wavelet decomposition at scales $j = 1$ (306) and $j = 2$ (301 545 795 885) and behaviour of decision function at scale $j = 2$
Vertical lines indicate estimated change points; $\pi = 0.00098$, $L = 75$, Wishart prior

Fig. 4*a* illustrates a realisation of an AR (2) process. The AR vector parameter changes from (–0.9, 0.9) to (–0.6, 0.6) at $N = 300$, and from (–0.6, 0.6) to (–0.67, 0.67) at $N = 500$. A ramp behaviour from $N = 800$ to $N = 899$ is also included, which simulates a return to normal operation conditions after a smooth change in the signal.

In this example, the inverse Wishart distribution is used as prior and the quadratic spline wavelet is used for the wavelet decomposition. Fig. 4*b* shows that the proposed algorithm is able to detect and locate the boundary of each segment. This means that the algorithm detects the abrupt changes at $N = 306$ with scale $j = 1$ and at $N = 301$ with $j = 2$. Smooth changes are also detected at $N = 545795885$ with scale $j = 2$, where the vertical lines indicate the estimated change points at each scale.

The decision function $\delta > 1$ for $j = 2$ (also shown in Fig. 4*b*) illustrates the behaviour of the posterior probabilities. The change points are estimated once the decision function reaches a value greater than 1 for a certain number of consecutive samples, e.g. there is a persistency in the alarms prior to the fault condition. The sliding test window of length $L = 75$ is chosen for all reported tests (see Section 4.1). In this example, a non-informative prior probability $p(H_q)$ is used and computed by $\pi = 1/1024$. Recall that $p(H_0) = \pi$ and $p(H_1) = 1 - \pi$.

Table 1 illustrates the estimated change points, which are estimated by comparing the posterior probabilities computed using Bayes' theorem (see Section 3.3) of the proposed wavelet-based network anomaly detector and the generalised likelihood ratio (GLR) test using AR models [1–3]. Different priors and wavelets are considered: a quadratic spline wavelet, the least-asymmetric (LA) compactly supported wavelets, Daubechies wavelets and the Haar wavelet [17]. All the assessed priors produce similar responses. Table 1 shows that the best performance is achieved with the quadratic spline and the LA wavelets under different priors settings. This is to be expected, since the quadratic spline wavelet has linear phase and the LA wavelet has almost linear phase [17]. This characteristic allows alignment of events with respect to the original time series after wavelet decomposition [18]. On the other hand, Daubechies orthonormal wavelets with 4 and 6 filter coefficients (i.e. $D(4)$ and $D(6)$) do not have linear phase, and so there is a misalignment between the original time series and the wavelet coefficients. The Haar wavelet does not give

Table 1: Estimated change points for proposed wavelet-based approach and GLR test using AR models

Approaches			Estimated change points		
	Wavelet	Scale	Wishart prior	Flat prior	Jeffreys' prior
Proposed wavelet-based approach	Spline	$j = 1$	306	306	310
		$j = 2$	301,545,795,885	303,559,796,871	304, 581,798,873
	LA(8)	$j = 1$	–	–	–
		$j = 2$	311,566,807,882	238,313,566,805,880	311,566,805,880
	D(4)	$j = 1$	318	315	315
		$j = 2$	304,804,879	305,804,879	307,803,878
	D(6)	$j = 1$	–	–	330
		$j = 2$	311,808,883	309,806,881	311,806,881
	Haar	$j = 1$	312	310	310
		$j = 2$	303,800,875	302,800,875	303,800,875
GLR test using AR models [1–3]	Model order = 2 Threshold $h = 3.0$ $L = 75$			355	

good results for approximated smooth changes mainly because it has only one vanishing moment [15, 17]. Note that the GLR test using AR models is unable to identify the subtle change at $N = 500$ or the ramp behaviour from $N = 800$ to $N = 899$ (see Table 1).

4.1 Choice of length L of sliding test window

In the proposed algorithm, the choice of the sliding test window length L will affect the probability of detection, the probability of false alarm, and therefore the delay for detection. Using the previous simulation set-up, different values of sliding test window of length L are assessed. The inverse Wishart distribution is used as prior, and the quadratic spline wavelet is used for the wavelet decomposition. The GLR test using AR models uses a threshold $h = 3.0$, model order 2, and sliding window $L = 75$. These two approaches are tested using Monte Carlo simulations. 100 realisations of the AR (2) process, described previously, are evaluated. In this analysis, true alarms are considered within the interval of ±25 samples around the true change points.

Table 2 shows the detection rate (DR) and the false alarm rate (FAR) of the assessed approaches. (The DR is the number of correct matches divided by the total number of known change points, and the FAR is the number of false alarms divided by the total number of observations [6].) According to the results in Table 2, a suitable sliding test window length is $L \approx 0.75M$, where M denotes the length of the shortest segment to be detected. The proposed approach shows a higher DR than the GLR test using AR models used previously [1–3]. This is because the GLR test using AR models requires a long time interval to estimate the parameters describing the process, with the unavoidable long delay between the estimated values and true values. Other performance comparisons have been presented [11] where it is also shown that multi-scale approaches outperform auto-regressive approaches.

Table 2: Performance comparison between wavelet-based approach at scale $j = 2$ and GLR test using AR models after 100 Monte Carlo simulations

Locations of change points		300	500	800	900	
	STW	DR	DR	DR	DR	FAR
Proposed wavelet-based approach	$L = 50$	0.600	0.350	0.790	0.560	0.002
	$L = 75$	0.770	0.370	0.900	0.890	0.001
	$L = 100$	0.520	0.260	0.650	0.090	0.001
	$L = 125$	0.430	0.210	0.500	0.000	0.001
GLR test using AR models [1–3]	$L = 75$	0.180	0.150	0.090	0.050	0.001

DR = detection rate; FAR = false alarm rate; STW = sliding test window

From the simulations presented here, it is evident that the adaptive threshold technique [7] and the static thresholds proposed previously [4, 5] may not be able to detect subtle changes in variance and frequency of the given signal. However, these approaches, as well as the GLR test using AR models, are evaluated below.

5 Application to Dial IP network

In this Section, we consider real-world network data collected every 10 min over a period of six months from the BT (British Telecommunications) Ignite's Dial IP service (a wholesale service offered to ISPs (Internet service providers) and OLOs (other local operators)). Fig. 5*a* shows a generic architecture of an ISP network. This network is designed to provide a complete Internet service package for use by ISPs and other customers requiring dial access to the Internet and to a host site containing the customer's application service [21, 22]. To test the proposed wavelet-based algorithm, we selected measurements from August 1999 to January 2000 that appeared to represent abnormalities in the expected traffic pattern. The abnormal periods were confirmed with the IOC's (Internet Operations Centre's) log. (Note that BT QoS (quality of service) cannot be inferred from the results reported in this paper.) The explanation of the network metrics considered in this Section can be found elsewhere [23]. The set of network metrics involves a call establishment phase (i.e. the *Connect_Time* and the *Log_Time*), connection and data transfer (i.e. the Domain Name Server (DNS) *Lookup_Time*, the *Web_Latency* and the *Data_Time* (time to download)).

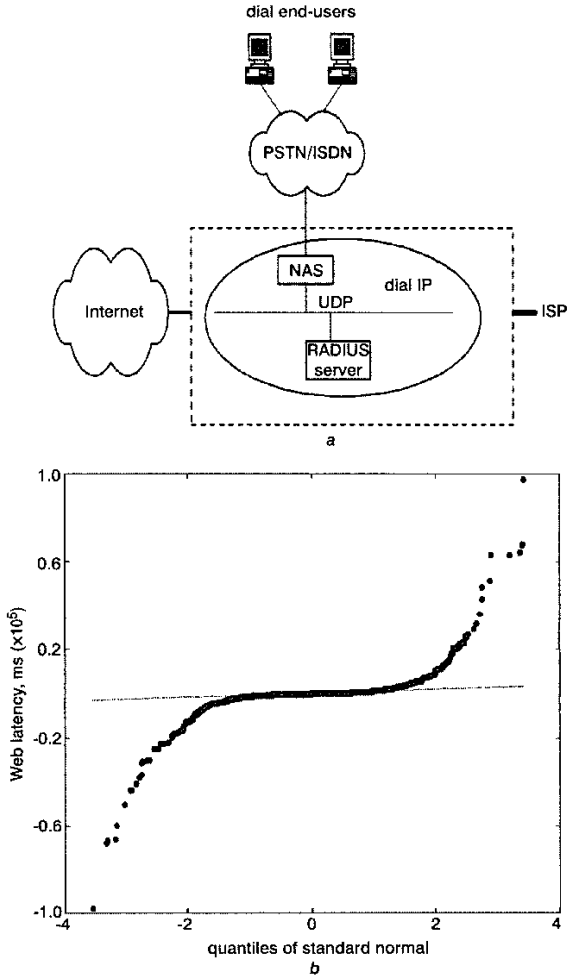

Fig. 5 *Dial IP network*
a Outline of monitored network indicating Dial IP network service within a typical ISP environment
UDP = user datagram protocol
PSTN = public switched telephone network
ISDN = integrated service digital network
NAS = network access server
ISP = Internet service provider
RADIUS = remote access dial in user service (RFC 2138/2139)
b Quantile-quantile plot of the *Web_Latency* metric after wavelet decomposition

As deviations from the normal pattern are recognised as the presence of network anomalies [1–3, 7, 9], the first step is to generate a template of normal behaviour. This

template then serves as a reference for detecting network anomalies from the normal behaviour of the current observations. The assumption of normal distribution is evaluated by using the marginal distribution of the time series after wavelet decomposition. The quantile distribution of the target network metrics is compared to the quantiles of the normal distribution using a quantile-quantile plot. Even though Fig. 5*b* shows that the distribution of the *Web_Latency* metric has a longer tail than the standard normal, we use the standard distribution since we are interested in the first and second moments only.

Table 3 shows three network anomalies in the Dial IP service under analysis. The Data Set No.1 anomalies were classified as an enhanced interior gateway routing protocol (EIGRP) storm on the network, which did not cause any noticeable QoS degradation. The anomaly in Data Set No.2 was classified as an unconfirmed anomaly (i.e. no IOC log was recorded).

Table 3: Network anomalies

Data set no.	No. of faults	Reported
1	2	20:15 on 27 August 1999 22:28 on 31 August 1999
2	1	04 November 1999

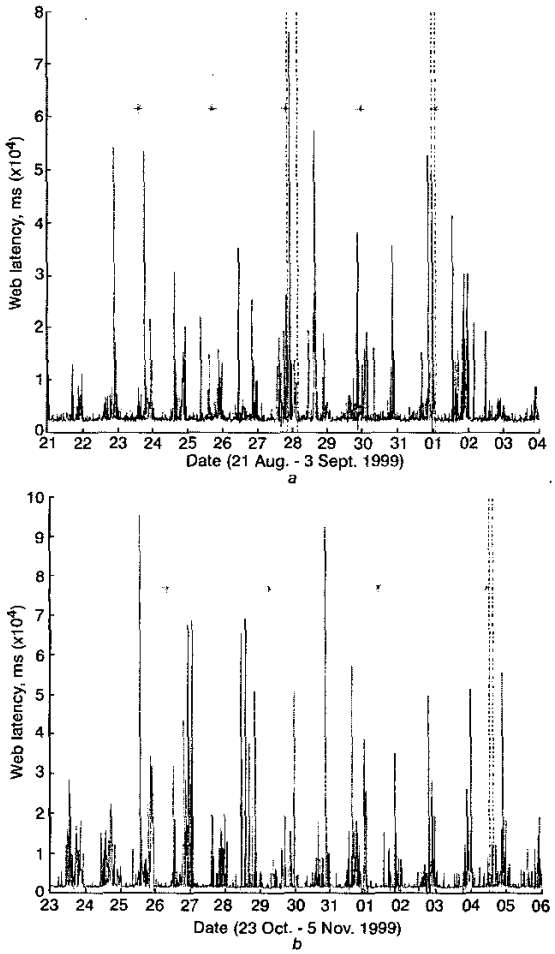

Fig. 6 *Network measurements of Web_Latency metric*
Web_Latency metric involves time (ms) taken for first packet to be returned from target web page (once connected)
* alarms
— · abnormal period

Fig. 6*a* shows the results of the Data Set No.1, using the inverse Wishart distribution as prior and the quadratic spline wavelet for the wavelet decomposition. Fig. 6*a* illustrates the *Web_Latency* metric alarms used to correlate the received IOC alarm logs. These alarms were generated at scales $j = 1$ and $j = 2$. Note that, by monitoring the *Log_Time* metric (a network metric representing the number of seconds taken to log into the system once connected), the proposed algorithm raised an alarm at 17:45 on 27 August and at 20:45 on 31 August. This is an example of how the algorithm is able to detect abnormal network behaviours in advance. (Taking into account the possible transitions between the chosen network metrics described elsewhere [23], a change observed in a particular network metric may propagate into another network metric.)

Fig. 6*b* shows the Data Set No.2. Although there was no confirmed IOC log record related to this data set, the algorithm was able to detect some abnormal network behaviours. The abnormal period (–·– in Fig. 6*b*) was also identified by other monitored network metrics. To increase the algorithm's sensitivity, the sliding test window of length L was tuned to track network short duration changes. Scales $j = 1$ and $j = 2$ were used to identify the abnormal behaviours.

Table 4 shows a summary of the faults detected, along with the performance measures: DR and FAR, as used previously [6]. A false alarm is normally declared when there is no related IOC log record. In the case of Data Set No.2, there was no confirmed IOC log record, and the abnormal condition was traced as correlated events among the different network metrics being monitored. An alarm is considered to be a true alarm if this is within the interval of 60min before and 25min after the network anomaly [1, 3]. For all the algorithms presented in Table 4 (i.e. static thresholds [4, 5], the GLR test using AR models [1–3] and adaptive threshold schemes [7]), only the results with the best performance are reported after tuning up all their parameters.

Table 4: Summary of faults detected

Approaches	Data set no.	No. of faults	Proactive alarms in minutes	DR	FAR
Wavelet-based network anomaly detector	1	2	50 17◇	1	0.00155
	2	1	50	1	0.00172
GLR test using AR models [1–3]	1	2	55 —	0.5	0.00155
	2	1	—	0	0.00345
Adaptive thresholds [7]	1	2	25◇ —	0.5	0.01810
	2	1	60	1	0.01727
Static thresholds [4, 5]	1	2	— —	0	0.01652
	2	1	—	0	0.02072

◇ detected after anomaly, — undetected anomaly
DR = detection rate; FAR = false alarm rate

Table 4 shows that the best performance is achieved by the proposed wavelet-based network anomaly detector. For example, regarding the FAR, the wavelet-based approach

shows the lowest FAR, and it is able to detect all network anomalies using neither thresholds nor AR modelling. In contrast, static thresholds [4, 5] cannot detect the network anomalies, and the GLR test using AR models [1–3] is only able to identify one of the three network anomalies. The adaptive threshold approach proposed elsewhere [7] is able to identify two of the three network anomalies. However, the design parameters, which should take into account the time-varying nature of the data, are difficult to tune (e.g. the magnitude of the event). Moreover, even with adaptive thresholding techniques, subtle behaviour changes in variance and frequency in the network metric being monitored may not be detected. Finally, note that static [4, 5] and adaptive thresholding [7] techniques show the highest FAR compared with the approach proposed here.

6 Conclusions

In this paper, we have presented an algorithm for detecting network anomalies, based on the undecimated discrete wavelet transform and Bayesian analysis. Since the analysis is carried out at different resolution levels, the algorithm is able to detect and locate subtle changes in variance and frequency in the given time series, by using the wavelet coefficients at these levels. Furthermore, this algorithm requires neither AR modelling nor thresholds to detect these changes. The unknown variance of the wavelet coefficients was considered as a stochastic nuisance parameter. Marginalisation was used to eliminate this nuisance parameter using informative and non-informative priors. Based on the reported tests, the choice of prior on the posterior probabilities has minimal effect on the estimated change points. These results also show that the best performance was obtained with the quadratic spline wavelet. This is to be expected because the quadratic spline wavelet with compact support has linear phase, and there is therefore an alignment between the original time series and the wavelet coefficients.

The proposed algorithm was also applied to monitor network measurements from the BT Ignite's Dial IP service. The results of the proposed algorithm show an improvement over adaptive thresholding techniques and AR models, and the ability to generate early warnings with a low false alarm rate. A suitable application of the proposed wavelet-based network anomaly detector algorithm is in the implementation of early warning procedures, to inform an Internet operation centre before network anomalous behaviour causes a loss of the service or degradation in QoS. At present, the wavelet-based framework is being extended to other types of network.

7 Acknowledgments

The authors would like to thank Ian Thurlow (BT) for providing the data analysed in Section 5, and Graham Walker (BT Ignite, IP Service Platforms Manager) for allowing these data sets to be used in this paper. They would also like to thank the referees for their constructive remarks. V. Alarcon-Aquino gratefully acknowledges the financial support from the National Council for Science and Technology (CONACYT), and the institutional support from the University of the Americas - Puebla, Mexico.

8 References

1 THOTTAN, M., and JI, C.: 'Statistical detection of enterprise network problems', *J. Netw. Syst. Manage.*, 1999, **7**, (1), pp. 27–45
2 HOOD, C.S., and JI, C.: 'Beyond thresholds: an alternative method for extracting information from network measurements'. Proceedings of IEEE Globecom Conference, 1997, Vol. 1, pp. 487–491
3 THOTTAN, M., and JI, C.: 'Proactive anomaly detection using distributed intelligent agents', *IEEE Netw.*, 1998, pp. 21–27
4 WARD, A., GLYNN, P., and RICHARDSON, K.: 'Internet service performance failure detection'. Presented at Workshop on *Internet server performance*, Wisconsin, 23 June 1998
5 MADRUGA, E.L., and TAROUCO, L.M.R.: 'Fault management tools for a co-operative and decentralised network operations environment', *IEEE J. Sel. Areas Commun.*, 1994, **12**, (6), pp. 1121–1130
6 FEATHER, F.F., and MAXION, R.A.: 'Fault detection in an ethernet network using anomaly signature matching'. Proceedings of ACM SIGCOMM, 1993, Vol. 23, pp. 279–288
7 HO, L.L., CAVUTO, D.J., PAPAVASSILIOU, S., and ZAWADZKI, A.G.: 'Adaptive and automated detection of service anomalies in transaction-oriented WAN's: network analysis, algorithms, implementation, and deployment', *IEEE J. Sel. Areas Commun.*, 2000, **18**, (5), pp. 744–757
8 LAZAR, A.A., WANG, W., and DENG, R.H.: 'Models and algorithms for network fault detection and identification: a review'. Proceedings of International Conference on *Communications systems*, ICCS' 92, Singapore, 1992, pp. 999–1003
9 MAXION, R.A., and FEATHER, F.E.: 'A case study of ethernet anomalies in a distributed computing environment', *IEEE Trans. Reliab.*, 1990, **39**, (4), pp. 433–443
10 BAKSHI, B.R.: 'Multiscale analysis and modelling using wavelets', *J. Chemometrics*, 1999, **13**, (3), pp. 415–434
11 KHALIL, M., and DUCHENE, J.: 'Detection and classification of multiple events in piecewise stationary signal: comparison between autoregressive and multi-scale approaches', *Signal Process.*, 1999, **75**, pp. 239–251
12 KOBAYASHI, M.: 'Wavelet analysis: application in industry'. IBM Research, Tokyo Research Laboratory, 1999
13 KASS, R.E., and WASSERMAN, L.: 'The selection of prior distribution by formal rules', *J. Am. Stat. Assoc.*, 1997, **96**, pp. 1343–1370
14 GUSTAFSSON, F.: 'The marginalised likelihood ratio test for detecting abrupt changes', *IEEE Trans. Autom. Control*, 1996, **41**, (1), pp. 66–78
15 MALLAT, S.: 'A wavelet tour of signal processing' (Academic Press 1999, 2nd edn.)
16 MALLAT, S., and ZHONG, S.: 'Characterisation of signals from multi-scale edges', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1992, **14**, (7), pp. 710–732
17 DAUBECHIES, I.: 'Ten lectures on wavelets' (SIAM, New York, 1992)
18 PERCIVAL, D.B., and MOFJELD, H.O.: 'Analysis of subtidal coastal sea level fluctuations using wavelets', *J. Am. Stat. Assoc.*, 1997, **92**, (439), pp. 868–880
19 APPEL, U., and BRANDT, A.V.: 'Adaptive sequential segmentation of piecewise stationary time series', *Inf. Sci.*, 1983, **29**, pp. 27–56
20 BERNARDO, J.M., and SMITH, A.F.M.: 'Bayesian theory' (Wiley, New York, 1994)
21 Supplier's information note (SIN 321): 'BTnet Dial IP service description'. Issue 3, British Telecommunications plc, London, UK, 2000 URL http://www.sinet.bt.com/
22 CISCO Systems Inc., URL http://www.cisco.com/warp/public/707/#radius
23 Visual Internet Benchmark, URL http://www.visualnetworks.com/products/prod_inbench_internet.html