2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

# Multi-Scale Analysis of Long Range Dependent Traffic for Anomaly Detection in Wireless Sensor Networks

Shanshan Zheng and John S. Baras

*Abstract*— **Anomaly detection is important for the correct functioning of wireless sensor networks. Recent studies have shown that node mobility along with spatial correlation of the monitored phenomenon in sensor networks can lead to observation data that have long range dependency, which could significantly increase the difficulty of anomaly detection. In this paper, we develop an anomaly detection scheme based on multi-scale analysis of the long range dependent traffic to address this challenge. In this proposed detection scheme, discrete wavelet transform is used to approximately de-correlate the traffic data and capture data characteristics in different time scales. The remaining dependencies are then captured by a multi-level hidden Markov model in the wavelet domain. To estimate the model parameters, we propose an online discounting Expectation Maximization (EM) algorithm, which also tracks variations of the estimated models over time. Network anomalies are detected as abrupt changes in the tracked model variation scores. We evaluate our detection scheme numerically using typical long range dependent time series.**

## I. Introduction

A wireless sensor network consists of a set of spatially scattered sensors that can be used to monitor and protect critical infrastructure assets, such as power grids, automated railroad control, water and gas distribution, etc. However, due to the unattended operating environment of sensor networks, it could be easy for attackers to compromise sensors and conduct malicious behaviors. Anomaly detection is thus critical to ensure the effective functioning of sensor networks. An anomaly detection procedure usually consists of two steps: first, collect network measurements and model the normal traffic as a reference; second, apply a decision rule to detect whether current network traffic deviates from the reference.Traditional anomaly detection methods usually assume that the network measurements are either independent or short range dependent. However, recent studies have shown that node mobility along with spatial correlation of the monitored phenomenon in sensor networks can lead to Long Range Dependent (LRD) traffic [1], which could lead to high false alarms using traditional methods.

In this paper, we develop an anomaly detection scheme based on multi-scale analysis of the long range dependent traffic. Discrete Wavelet Transform (DWT) is used to approximately de-correlate autocorrelated stochastic processes. Most of the literature work on using DWT for anomaly detection use the first order or second order statistics (mean or variance) of the wavelet coefficients for anomaly detection, e.g., detect changes in the mean or variance of the wavelet coefficients over a moving window [2], [3]. In contrast, we build a probabilistic model for the wavelet coefficients through a multi-level hidden Markov model (HMM), in the expect to capture the remaining dependency among the transformed data thus better model the network traffic and improve detection accuracy. We design a forward-backward decomposition scheme and an online discounting Expectation Maximization (EM) algorithm to estimate model parameters. The online EM algorithm can also track the structure changes of the estimated HMMs by evaluating a model variation score using the symmetric relative entropy between the current estimated model and a previous estimated model. Network anomalies are then detected as abrupt changes in the tracked model variation scores .

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under award number 013641-001 for the Multi-Scale Systems Center (MuSyC), through the FRCP of SRC and DARPA, US Air Force Office of Scientific Research MURI award FA9550-09-1-0538, and Army Research Office MURI award W911-NF-0710287

Shanshan Zheng and John S. Baras are with the Institute for Systems Research and the Department of Electrical and Computer Engineering, University of Maryland, College Park, {sszheng, baras}@umd.edu

## II. Related Work

Network anomaly detection is an important problem and has received numerous research efforts. It involves modeling of the normal traffic as a reference, and computing the statistical distance between the analyzed traffic and the reference. The modeling of the normal traffic can be based on various statistical characteristics of the data. For example, Lakhina et al. [4] proposed to use Principal Component Analysis to identify an orthogonal basis along which the network measurements exhibit the highest variance. The principal components with high variance model the normal behavior of a network, whereas the remaining components of small variance are used to identify and classify anomalies. Spectral density [5] and covariance [6], have also been used for modeling normal network traffic.

Besides these methods, wavelet transform is another popular technique used for modeling network traffic, especially for the LRD traffic. Abry et al. [7] proposed a wavelet-based tool for analyzing LRD time series and a related semi-parametric estimator for estimating LRD parameters. Barford et al. [2] assume that the low frequency band signal of a wavelet transform represents the normal traffic pattern. They then normalize both medium and high frequency band signals to compute a weighted sum. An alarm is raised if the variance of the combined signal exceeds a pre-selected threshold. The key feature of these wavelet-based methods lies in the fact that wavelet transform can turn the long range dependency among the data samples into a short memory structure among the wavelet coefficients [7]. In our work, we build a wavelet-domain multi-level hidden Markov model for the LRD network traffic. The merit of our method is the model's mathematical tractability and its capability of capturing data dependency.

To measure the deviation of the analyzed traffic from the reference model, various statistical distances can be used, including simple threshold, mean quadratic distances [8], and entropy[9]. Entropy is a measure of the uncertainty of a probability distribution. It can be used to compare certain qualitative differences of probability distributions. In our detection scheme, we apply the symmetric relative entropy as a distance measure. The online EM algorithm can efficiently compute the symmetric relative entropy between the current estimated HMM model and a previous estimated one. An anomaly is then detected as abrupt changes in the symmetric relative entropy measurements.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

## III. Long Range Dependent Traffic in Wireless Sensor Networks

A time series $\{x(t)\}_{t=1}^{N}$ is considered to be long range dependent if its autocorrelation function $\rho(k)$ decays at a rate slower than an exponential decay. Typically, $\rho(k)$ asymptotically behaves as $ck^{2H-2}$ for $0.5 < H < 1$, where $c > 0$ is a constant and $H$ is the Hurst parameter. The intensity of LRD is expressed as the speed of the decay for the autocorrelation function and is measured by the Hurst parameter, i.e., as $H \to 1$, the dependence among data becomes stronger. It can be shown that $\sum_{k=1}^{\infty} \rho(k) = \infty$. Intuitively, LRD implies that the process has infinite memory, i.e., individually small high-lag correlations have an important cumulative effect. This is contrast to the conventional Short Range Dependent (SRD) process which are characterized by an exponential decay of the autocorrelations resulting in a summable autocorrelation function. LRD is an important property for traffic modeling as it is likely to be responsible for the decrease in both the network performance and the quality of service [8] .

A wireless sensor network operates on the IEEE 802.15.4 standard. It has been shown recently [1] that the traffic generated from a single mobile node in the wireless sensor network can be represented by an ON/OFF process $X(t)$, where the probability density function of the ON period $\tau_a$ can be approximated by a truncated Pareto distribution [1]

$$f_{\tau_a}(x) = \frac{\gamma_a x^{-(\gamma_a+1)}}{t_{min}^{-\gamma_a} - t_{max}^{-\gamma_a}},$$

with $t_{min}(t_{max})$ denoting the minimum (maximum) ON time and $\gamma_a$ denoting the tail index. The value of $\gamma_a$ depends on the variability of node mobility pattern and the spatial correlation of the monitored phenomena by the network. Using this traffic model for wireless sensor networks, we analyze the dynamic behaviors of network traffic measurements, such as the packet round trip time, the number of received packets per second, etc., by conducting experiments using Network simulator 2 (NS-2). It is found that the LRD property do exit in the simulated data traces, with typical Hurst parameters between 0.8 and 0.9. Incorporating LRD precisely in the design of anomaly detection schemes for wireless sensor networks is critical.

## IV. Wavelet domain hidden Markov model for long range dependent traffic

Wavelet transforms have been popular for analyzing autocorrelated measurements due to their capability to compress multi-scale features and approximately de-correlate the autocorrelated stochastic processes. We build a Hidden Markov Model (HMM) for the wavelet transform coefficients of the network traffic. The basic idea for transform domain model is that a linear invertible transform can often 'restructure' an signal, generating transform coefficients whose structure is simpler to model.

### A. Wavelet domain hidden Markov model

In wavelet transform, the measurements $x(t), t = 1, \ldots, N$ are decomposed into multiple scales by a weighted sum of a certain orthonormal basis functions,

$$x(t) = \sum_{k=1}^{N} a_{L,k}\phi_{L,k}(t) + \sum_{m=1}^{L}\sum_{k} d_{m,k}\psi_{m,k}(t),$$

where $\phi_{j,k}$, $\psi_{j,k}$ are the orthonormal basis, $a_{L,k}$, $d_{m,k}$ are the approximation and detail coefficients. The approximation coefficients $a_{L,k}$ provide the general shape of the signal, while the detail coefficients $d_{m,k}$ from different scales provide different levels of details for the signal content, with $d_{1,k}$ providing the finest details and $d_{L,k}$ providing the coarsest details. In our work, we apply the Discrete Wavelet Transform (DWT) to the network traffic. A discrete wavelet transform is a wavelet transform for which the basis functions are discretely sampled. DWT can be explained using a pair of quadrature mirror filters, which includes a high pass filter $h[n]$ and a low pass filter $g[n]$. Efficient methods have been developed for decomposing a signal using a family of wavelet basis functions based on convolution with the corresponding quadrature mirror filters. A 2-level discrete wavelet transform using the corresponding quadrature mirror filters is illustrated in Fig. 1.

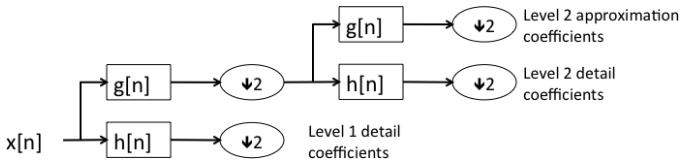


Fig. 1: 2-level discrete wavelet transform

However, the wavelet transform cannot completely de-correlate real-world signals, i.e., a residual dependency structure always remains among the wavelet coefficients. We use a hidden Markov model to capture the remaining dependency. It is based on two properties of the wavelet transform observed in literature [10], [11]: first is the *Clustering* property, which means that if a particular wavelet coefficient is large/small, the adjacent coefficients are very likely to also be large/small; second is the *Persistence* property, which means that large/small values of wavelet coefficients tend to propagate across scales.

For a signal $x(t)$ that is decomposed into $L$ scales, with wavelet coefficients $d_{j,k},\ k = 1, \ldots, n_j$ and $j = 1, \ldots, L$, we assume that each wavelet coefficient is associated with a hidden state $s_{j,k}$. We then use a hidden Markov model to characterize the wavelet coefficients through the factorization

$$\begin{aligned} &P(\{d_{1,i}, s_{1,i}\}_{i=1}^{n_1}, \ldots, \{d_{L,i}, s_{L,i}\}_{i=1}^{n_L}) \\ &= p(s_{L,1}) \prod_{j=2}^{n_L} p(s_{L,j}|s_{L,j-1}) \prod_{i=1}^{L-1} p(s_{i,1}|s_{i+1,1}) \\ &\cdot \prod_{i=1}^{L-1}\prod_{j=2}^{n_i} p(s_{i,j}|s_{i,j-1}, s_{i+1,\lceil j/2 \rceil}) \prod_{i=1}^{L}\prod_{j=1}^{n_i} p(d_{i,j}|s_{i,j}). \quad (1) \end{aligned}$$

This factorization involves three main conditional independence assumptions: first, conditioned on the states at the previous coarser scale $i+1$, the states at the scale $i$ form a first order Markov chain; second, conditioned on the corresponding state at the previous coarser scale $i+1$, i.e., $s_{i+1,\lceil j/2 \rceil}$, and the previous state at the same scale, i.e., $s_{i,j-1}$, the state $s_{i,j}$ is independent of all states in coarser scales; third, the wavelet coefficients are independent of everything else given their hidden states. The three independence assumptions are critical for deriving the inference algorithms for this wavelet domain HMM. Fig. 2 illustrate a hidden Markov model for a 3-level wavelet decomposition.

### B. Estimating model parameters using an Expectation-Maximization (EM) algorithm

Denote the set of the wavelet coefficients and their hidden states by $\mathcal{D} = \{\{d_{L,i}\}_{i=1}^{n_L}, \ldots, \{d_{1,i}\}_{i=1}^{n_1}\}$ and $\mathcal{S} = \{\{s_{L,i}\}_{i=1}^{n_L}, \ldots, \{s_{1,i}\}_{i=1}^{n_1}\}$ respectively, where $n_i$ is the number of wavelet coefficients in the $i^{th}$ scale. The parameters of the HMM include the following three probabilities: first is the initial probability for the state $s_{L,1}$, i.e., $\pi_k =$

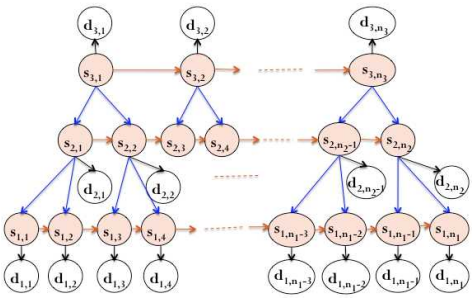

Fig. 2: HMM for 3-level wavelet decomposition

$P(s_{L,1} = k)$; second is the two types of state transition probabilities, i.e.,

$$\pi^{i,1}_{k_1|k_2} = P(s_{i,1} = k_1 | s_{i+1,1} = k_2) \text{ for } i < L,$$
$$\pi^{i}_{k_1|k_2,k_3} = P(s_{i,j} = k_1 | s_{i,j-1} = k_2, s_{i+1,\lceil j/2 \rceil} = k_3);$$

and third is the conditional probability of the wavelet coefficients given their hidden states at the $i^{th}$ scale, i.e., $P(d_{i,j}|s_{i,j} = k)$, which can be modeled by a mixture Gaussian distribution. For simplicity and presentation clearity, we use a single Gaussian distribution to catpure $P(d_{i,j}|s_{i,j} = k)$, i.e, $P(d_{i,j}|s_{i,j} = k) \sim \mathcal{N}(\mu^i_k, \sigma^i_k)$, where $\mu^i_k$ and $\sigma^i_k$ are the mean and the standard deviation for the state $k$ in the $i^{th}$ scale. The extension to mixture Gaussian distribution is straightforward. These model parameters, denoted by $\boldsymbol{\theta} = \{\pi_k, \pi^{i,1}_{k_1|k_2}, \pi^i_{k_1|k_2,k_3}, \mu^i_k, \sigma^i_k\}$, can be estimated from the real data using the maximum likelihood criterion. Due to the intractability of direct maximization of the likelihood function, we apply an Expectation Maximization (EM) algorithm to estimate the parameters.

The EM algorithm provides an maximum likelihood estimation of model parameters by iteratively applying an E-step and a M-step. In the E-step, the expected value of the log likelihood function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\mathcal{S}|\mathcal{D},\boldsymbol{\theta}^{(t)}}[\log P_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{D})]$ is computed. Then in the M-step, the parameter that maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is computed, i.e., $\boldsymbol{\theta}^{(t+1)} = arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. To implement the two steps, we define the following posterior probabilities,

$$\begin{aligned}
\gamma^{i,j}_k &= P(s_{i,j} = k|\mathcal{D}), \\
\gamma^{i,1}_{k_1,k_2} &= P(s_{i,1} = k_1, s_{i+1,1} = k_2|\mathcal{D}), \text{ for } i < L \\
\gamma^{i,j}_{k_1,k_2,k_3} &= P(s_{i,j} = k_1, s_{i,j-1} = k_2, s_{i,\lceil j/2 \rceil} = k_3|\mathcal{D}).
\end{aligned}$$

According to equation (1), maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ using Lagrange method leads to the following estimation of $\boldsymbol{\theta}$,

$$\pi_k = \gamma^{L,1}_k,$$
$$\pi^{i,1}_{k_1|k_2} = \frac{\gamma^{i,1}_{k_1,k_2}}{\sum_{l\in\mathcal{K}} \gamma^{i,1}_{l,k_2}}, \ \pi^i_{k_1|k_2,k_3} = \frac{\sum_{j=2}^{n_i} \gamma^{i,j}_{k_1,k_2,k_3}}{\sum_{l\in\mathcal{K}} \sum_{j=2}^{n_i} \gamma^{i,j}_{l,k_2,k_3}},$$
$$\mu^i_k = \frac{\sum_{j=1}^{n_i} \gamma^{i,j}_k d_{i,j}}{\sum_{j=1}^{n_i} \gamma^{i,j}_k}, \ (\sigma^i_k)^2 = \frac{\sum_{j=1}^{n_i} \gamma^{i,j}_k (d_{i,j} - \mu^i_k)^2}{\sum_{j=1}^{n_i} \gamma^{i,j}_k},$$

where $\mathcal{K}$ represents the domain of the hidden states. The computation of the posterior probabilities is a little more involved. Using a brute force computation by direct marginalization will take $O(N \cdot |\mathcal{K}|^N)$ operations, where $N$ is the length of the input signal. However, by exploiting the sparse factorization in equation (1) and manipulating the distributive property of '+' and '×', we are able to design an forward-backward decomposition algorithm to compute these posterior probabilities with computational complexity $O(N \cdot |\mathcal{K}|^{L+1})$, where $L$ is the wavelet decomposition level and much smaller than $N$.

### C. Forward-backward decomposition

Our algorithm extends the classical forward-backward decomposition algorithm for a one-level hidden Markov model to our multi-level case. The key point is to only maintain $L$ appropriate hidden states in both the forward and backward variables for computational efficiency.

*1) Forward decomposition:* Let

$$\begin{aligned}
\mathcal{S}_{i,j} &= \{s_{L,\lceil 2^{i-L} j \rceil}, \ldots, s_{i+1,\lceil 2^{-1} j \rceil}, s_{i,j}, \\
&\qquad s_{i-1,2(j-1)}, \ldots, s_{1,2^{i-1}(j-1)}\}, \text{ and} \\
\mathcal{D}_{i,j} &= \{d_{L,k\le\lceil 2^{i-L} j \rceil}, \ldots, d_{i+1,k\le\lceil 2^{-1} j \rceil}, d_{i,k\le j}, \\
&\qquad d_{i-1,k\le 2(j-1)}, \ldots, d_{1,k\le 2^{i-1}(j-1)}\},
\end{aligned}$$

we define the forward variable to be $\alpha_{i,j} = P(\mathcal{S}_{i,j}, \mathcal{D}_{i,j})$. Denote $[\alpha_{1,2^{h-1}j}] = f(h, \alpha_{h,j})$ for $h, j \in Z^+$ to be a dynamic programming algorithm with input parameters $(h, \alpha_{h,j})$ and output parameter $\alpha_{1,2^{h-1}j}$. The pseudo-code for computing the forward variables using dynamic programming is shown in Table I. Its correctness can be proved using the three conditional independence assumptions in our HMM. For simplicity and presentation clarity, in Table I we assume that the input data length $N$ is an order of 2, and denote the conditional probability $P(d_{i,j}|s_{i,j})$ by $g_1(d_{i,j})$, and $P(s_{i,j}|s_{i,j-1}, s_{i+1,\lceil j/2 \rceil})$ by $g_2(s_{i,j})$.

TABLE I: Computing forward variables

Initialization: $\alpha_{L,1} = P(s_{L,1}, d_{L,1})$
For $k_L = 1$ to $2^{-L}N$
  $\alpha_{1,2^{L-1}k_L} = f(L, \alpha_{L,k_L})$
  $\alpha_{L,k_L+1} = g_1(d_{L,k_L+1}) \sum_{s_{L,k_L}} [g_2(s_{L,k_L+1}) \cdot \alpha_{1,2^{L-1}k_L}]$
end

function $[\alpha_{1,2^{h-1}j}] = f(h, \alpha_{h,j})$
If $h == 2$,
  $\alpha_{1,2j-1} = g_1(d_{1,2j-1}) \sum_{s_{1,2j-2}} [g_2(s_{1,2j-1}) \cdot \alpha_{2,j}]$
  $\alpha_{1,2j} = g_1(d_{1,2j}|s_{1,2j}) \sum_{s_{1,2j-1}} [g_2(s_{1,2j}) \cdot \alpha_{1,2j-1}]$
else
  $\alpha_{h-1,2j-1} = g_1(d_{h-1,2j-1}) \sum_{s_{h-1,2j-2}} [g_2(s_{h-1,2j-1}) \cdot \alpha_{h,j}]$
  $\alpha_{1,2^{h-2}(2j-1)} = f(h-1, \alpha_{h-1,2j-1})$
  $\alpha_{h-1,2j} = g_1(d_{h-1,2j}) \sum_{s_{h-1,2j-1}} [g_2(s_{h-1,2j}) \cdot \alpha_{1,2^{h-2}(2j-1)}]$
  $\alpha_{1,2^{h-2}(2j)} = f(h-1, \alpha_{h-1,2j})$
End

There is one implementation issue for the algorithm in Table I, namely, the numerical under- or over-flow of $\alpha_{i,j}$ as $P(\mathcal{S}_{i,j}, \mathcal{D}_{i,j})$ becomes smaller and smaller with the increasing number of observations. Therefore, it is necessary to scale the forward variables by positive real numbers to keep the numeric values within reasonable bounds. One solution is to use a scaled version $\bar{\alpha}_{i,j} = \frac{\alpha_{i,j}}{c_{i,j}}$, where $c_{i,j} = \sum_{\mathcal{S}_{i,j}} \alpha_{i,j}$. In this way, $\bar{\alpha}_{i,j}$ represents the probability $P(\mathcal{S}_{i,j}|\mathcal{D}_{i,j})$ and $c_{i,j}$ represents the probability $P(d_{i,j}|\mathcal{D}_{i,j}\backslash d_{i,j})$. It is straightforward to prove that both $c_{i,j}$ and $\bar{\alpha}_{i,j}$ do not depend on the number of observations. The algorithm for computing $(\bar{\alpha}_{i,j}, c_{i,j})$ can be obtained by adding a normalization step after each update of $\alpha_{i,j}$ for the algorithm in Table I.

*2) Backward decomposition:* Let $\mathcal{D}^c_{i,j} = \mathcal{D} - \mathcal{D}_{i,j}$, we define the backward variable to be $\beta_{i,j} = P(\mathcal{D}^c_{i,j}|\mathcal{S}_{i,j})$, which can be computed using a similar dynamic programming algorithm as the one in Table I. To avoid the numerical under- or over-flow problem, instead of computing $\beta_{i,j}$, we compute a scaled version $\bar{\beta}_{i,j}$ as is shown in Table II. The scaled backward variable represents the probability $\frac{P(\mathcal{D}^c_{i,j}|\mathcal{S}_{i,j})}{P(\mathcal{D}^c_{i,j}|\mathcal{D}_{i,j})}$ . The correctness of the algorithm in Table II can be verified using the three conditional independence assumptions in our HMM. Due to space limit, we omit the proof here.

TABLE II: Computing scaled backward variables

Initialization: $\bar{\beta}_{1,N/2} = 1$
For $k_L = 2^{-L}N$ to 1
$\bar{\beta}_{L,k_L} = f(L, \bar{\beta}_{1,2^{L-1}k_L})$
$\bar{\beta}_{1,2^{L-1}(k_L-1)} = \sum_{s_{L,k_L}} \frac{g_1(d_{L,k_L})g_2(s_{L,k_L})\bar{\beta}_{L,k_L}}{c_{L,k_L}}$
end

function $[\bar{\beta}_{h,j}] = f(h, \bar{\beta}_{1,2^{h-1}j})$
If $h == 2$,
$\bar{\beta}_{1,2j-1} = \sum_{s_{1,2j}} \frac{g_1(d_{1,2j})g_2(s_{1,2j})\cdot\bar{\beta}_{1,2j}}{c_{1,2j}}$
$\bar{\beta}_{2,j} = \sum_{s_{1,2j-1}} \frac{g_1(d_{1,2j-1})g_2(s_{1,2j-1})\cdot\bar{\beta}_{1,2j-1}}{c_{1,2j-1}}$
else
$\bar{\beta}_{h-1,2j} = f(h-1, \bar{\beta}_{1,2^{h-2}2j})$
$\bar{\beta}_{1,2^{h-2}(2j-1)} = \sum_{s_{h-1,2j}} \frac{g_1(d_{h-1,2j})g_2(s_{h-1,2j})\cdot\bar{\beta}_{h-1,2j}}{c_{h-1,2j}}$
$\bar{\beta}_{h-1,2j-1} = f(h-1, \bar{\beta}_{1,2^{h-2}(2j-1)})$
$\bar{\beta}_{h,j} = \sum_{s_{h-1,2j-1}} \frac{g_1(d_{h-1,2j-1})g_2(s_{h-1,2j-1})\cdot\bar{\beta}_{h-1,2j-1}}{c_{h-1,2j-1}}$
End

*3) Computing posterior probabilities:* Since $\bar{\alpha}_{i,j} = P(\mathcal{S}_{i,j}|\mathcal{D}_{i,j})$ and $\bar{\beta}_{i,j} = \frac{P(\mathcal{D}^c_{i,j}|\mathcal{S}_{i,j})}{P(\mathcal{D}^c_{i,j}|\mathcal{D}_{i,j})}$, we have $\bar{\alpha}_{i,j} \cdot \bar{\beta}_{i,j} = P(\mathcal{S}_{i,j}|\mathcal{D})$ according to the Markovian property of our HMM. Then the posterior probability $\gamma(\cdot)$ can be computed as

$$\begin{aligned}
\gamma^{i,j}_k &= \sum \bar{\alpha}_{i,j} \cdot \bar{\beta}_{i,j}, \\
\gamma^{i,1}_{k_1,k_2} &= \sum \bar{\alpha}_{i,1} \cdot \bar{\beta}_{i,1}, \\
\gamma^{L,j}_{k_1,k_2} &= \sum \bar{\alpha}_{1,2^{L-1}(j-1)} \cdot \bar{\beta}_{L,j} \cdot \frac{g_1(d_{L,j})g_2(s_{L,j})}{c_{L,j}}, \\
\gamma^{i,j}_{k_1,k_2,k_3} &= \begin{cases} \sum \bar{\alpha}_{1,2^{i-1}(j-1)}\bar{\beta}_{i,j} \cdot \frac{g_1(d_{i,j})\cdot g_2(s_{i,j})}{c_{i,j}}, \text{if j is even,} \\ \sum \bar{\alpha}_{i+1,\lceil j/2 \rceil}\bar{\beta}_{i,j} \cdot \frac{g_1(d_{i,j})\cdot g_2(s_{i,j})}{c_{i,j}}, \text{if j is odd.} \end{cases}
\end{aligned}$$

Without confusion, we omit the variables under $\sum$ for the above equations. The correctness of these equations can be derived according to the three conditional independence assumptions in our HMM.

## V. Anomaly detection by tracking HMM model variations

A first thought on the anomaly detection problem is to treat the anomalies as abrupt changes in the HMM modeled data and then apply change-point detection methods to detect these abrupt changes. However, it is found that directly applying change-point detection methods to the HMM modeled data is computationally expensive. We designed here a lightweight anomaly detection scheme based on detecting the structure changes of the estimated HMM. An important requirement for anomaly detection is to make the decision making process online. Therefore, we first develop an online EM algorithm for HMM model estimation.

### A. An online discounting EM algorithm

Before we present the online EM algorithm, we first introduce the so called *limiting EM algorithm* [12]. Let $\mathbf{x}$ denote the hidden states and $\mathbf{y}$ denote the observations. If the joint probability distribution $p_\theta(\mathbf{x}, \mathbf{y})$ belongs to an exponential family such that

$$p_\theta(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}, \mathbf{y}) \exp(\langle \phi(\theta), ss(\mathbf{x}, \mathbf{y}) \rangle - A(\theta))$$

where $\langle \cdot \rangle$ denotes the scalar product, $ss(\mathbf{x}, \mathbf{y})$ is the sufficient statistics and $A(\theta)$ is some log-partition function. If the equation $\langle \nabla_\theta \phi(\theta), ss \rangle - \nabla_\theta A(\theta) = 0$ has a unique solution, denoted by $\theta = \bar{\theta}(ss)$, then the limiting EM algorithm obeys the simple recursion

$$ss_{k+1} = E_{\theta^*}[E_{\bar{\theta}(ss_k)}[ss(\mathbf{x}, \mathbf{y})|\mathbf{y}]],$$

where $\theta^*$ represents the true model parameter . Since $E_{\theta^*}[E_\theta[ss(\mathbf{x}, \mathbf{y})|\mathbf{y}]]$ may be estimated consistently from the observations by $\frac{1}{n}\sum_{t=1}^n E_\theta[ss(x_t, y_t)|y_t]$ , an online EM algorithm can be obtained by using the conventional stochastic approximation procedure

$$\hat{ss}_{k+1} = \gamma_{k+1} E_{\bar{\theta}(\hat{ss}_k)}[ss(x_{k+1}, y_{k+1})|y_{k+1}] + (1-\gamma_{k+1})\hat{ss}_k,$$

where $\gamma_k$ is a time discounting factor. The estimation of model parameters can then be derived from the sufficient statistics $\hat{ss}$. It is proved [12] that under suitable assumptions, this online EM algorithm is an asymptotically efficient estimator of the model parameter $\theta^*$.

It is not difficult to see that the joint probability distribution $P(\mathcal{S}, \mathcal{D})$ for our HMM model satisfies the above mentioned conditions for applying the limiting EM algorithm. For each wavelet coefficient $d_{i,j}$, we have the following sufficient statistics for computing the HMM model parameters,

$$\begin{aligned}
\tau^{i,j}_{l,k} &= \textstyle\sum_{m=1}^{n^{i,j}_l} P(s_{l,m} = k, \mathcal{S}_{i,j}|\mathcal{D}_{i,j}), \\
\hat{\tau}^{i,j}_{l,k} &= \textstyle\sum_{m=1}^{n^{i,j}_l} P(s_{l,m} = k, \mathcal{S}_{i,j}|\mathcal{D}_{i,j}) \cdot d_{l,m}, \\
\bar{\tau}^{i,j}_{l,k} &= \textstyle\sum_{m=1}^{n^{i,j}_l} P(s_{l,m} = k, \mathcal{S}_{i,j}|\mathcal{D}_{i,j}) \cdot d^2_{l,m}, \\
\tau^{i,j}_{l,k_1,k_2,k_3} &= \textstyle\sum_{m=2}^{n^{i,j}_l} P(s_{l,m} = k_1, s_{l,m-1} = k_2, \\
&\qquad s_{l+1,\lceil m/2 \rceil} = k_3, \mathcal{S}_{i,j}|\mathcal{D}_{i,j}),
\end{aligned}$$

where $l \in \{1, \ldots, L\}$ is the scale index, $k \in \mathcal{K}$ is the hidden state index, and $n^{i,j}_l$ represents the number of observed wavelet coefficients in scale $l$ after $d_{i,j}$ arrives, i.e.,

$$n^{i,j}_l = \begin{cases} \lceil 2^{i-l} j \rceil & \text{if } l \geq i, \\ 2^{l-i} j & \text{if } l < i. \end{cases}$$

It is straightforward to prove that the HMM model parameters $\{\pi^l_{k_1|k_2,k_3}, \mu^l_k, \sigma^l_k\}$ can be updated using $\{\tau^{i,j}_{l,k}, \hat{\tau}^{i,j}_{l,k}, \bar{\tau}^{i,j}_{l,k} \tau^{i,j}_{l,k_1,k_2,k_3}\}$ as follows,

$$\pi^l_{k_1|k_2,k_3} = \frac{\sum_{\mathcal{S}_{i,j}} \tau^{i,j}_{l,k_1,k_2,k_3}}{\sum_{k_1}\sum_{\mathcal{S}_{i,j}} \tau^{i,j}_{l,k_1,k_2,k_3}}, \quad (2)$$

$$\mu^l_k = \frac{\sum_k \sum_{\mathcal{S}_{i,j}} \hat{\tau}^{i,j}_{l,k}}{\sum_k \sum_{\mathcal{S}_{i,j}} \tau^{i,j}_{l,k}}, \quad (3)$$

$$(\sigma^l_k)^2 = \frac{\sum_k \sum_{\mathcal{S}_{i,j}} \bar{\tau}^{i,j}_{l,k}}{\sum_k \sum_{\mathcal{S}_{i,j}} \tau^{i,j}_{l,k}} - (\frac{\sum_k \sum_{\mathcal{S}_{i,j}} \hat{\tau}^{i,j}_{l,k}}{\sum_k \sum_{\mathcal{S}_{i,j}} \tau^{i,j}_{l,k}})^2. \quad (4)$$

Note that the HMM parameter $\pi^k$ and $\pi^{i,1}_{k_1|k_2}$ can be updated using the sufficient statistics $\tilde{\tau}^{i,j}_{L,k} = P(s_{L,1} = k, \mathcal{S}_{i,j}|\mathcal{D}_{i,j})$ and $\tilde{\tau}^{i,j}_{l,1} = P(s_{l,1} = k_1, s_{l+1,1} = k_2, \mathcal{S}_{i,j}|\mathcal{D}_{i,j})$. But we omit

the related discussions here, as the computations are similar.

The next step is to design recursive (online) updates of the sufficient statistics $\tau_{l,k}^{i,j}$, $\hat{\tau}_{l,k}^{i,j}$, $\bar{\tau}_{l,k}^{i,j}$ and $\tau_{l,k_1,k_2,k_3}^{i,j}$. According to the Markovian property of our HMM, the online updates of the sufficient statistics can be achieved by following a similar dynamic programming procedure as the one for computing the scaled forward variables $\bar{\alpha}_{i,j}$. Recall that $\bar{\alpha}_{i,j}$ is computed by adding a normalization step after $\alpha_{i,j}$ is computed in the algorithm in Table I. The sufficient statistics are updated once $\bar{\alpha}_{i,j}$ is updated. For illustration purpose, we only show here how to update the sufficient statistics when $\alpha_{h,2j-1}$ in Table I is computed. Updates for the other cases are similar. For $l \in \{1, \ldots, L\}$, let $\gamma_{h-1,2j-1}$ be a time discounting factor, and $\delta_{h-1}^{l}$ be the Dirac Delta function such that

$$\delta_{h-1}^{l} = \begin{cases} 1 & \text{if } l = h-1, \\ 0 & \text{if } l \neq h-1. \end{cases}$$

Define

$$r_{h-1,2j-1}^{l} = \delta_{h-1}^{l} \cdot \gamma_{h-1,2j-1} \cdot \bar{\alpha}_{h-1,2j-1},$$
$$t_{h-1,2j-1}^{l} = (1 - \delta_{h-1}^{l}\gamma_{h-1,2j-1}) g_1(d_{h-1,2j-1})/c_{h,j},$$
$$q_{h-1,2j-1}^{l} = \delta_{h-1}^{l}\gamma_{h-1,2j-1} \frac{g_1(d_{h-1,2j-1}) g_2(s_{h-1,2j-1}) \bar{\alpha}_{h,j}}{c_{h-1,2j-1}}.$$

We then have the following equations for updating the sufficient statistics,

$$\tau_{l,k}^{h-1,2j-1} = r_{h-1,2j-1}^{l} + t_{h-1,2j-1}^{l} \sum_{s_{h-1,2j-2}} g_2(s_{h-1,2j-1}) \tau_{l,k}^{h,j}, \quad (5)$$

$$\hat{\tau}_{l,k}^{h-1,2j-1} = r_{h-1,2j-1}^{l} d_{h-1,2j-1} + t_{h-1,2j-1}^{l} \sum_{s_{h-1,2j-2}} g_2(s_{h-1,2j-1}) \hat{\tau}_{l,k}^{h,j}, \quad (6)$$

$$\bar{\tau}_{l,k}^{h-1,2j-1} = r_{h-1,2j-1}^{l} d_{h-1,2j-1}^{2} + t_{h-1,2j-1}^{l} \sum_{s_{h-1,2j-2}} g_2(s_{h-1,2j-1}) \bar{\tau}_{l,k}^{h,j}, \quad (7)$$

$$\tau_{l,k_1,k_2,k_3}^{h-1,2j-1} = q_{h-1,2j-1}^{l} + t_{h-1,2j-1}^{l} \sum_{s_{h-1,2j-2}} g_2(s_{h-1,2j-1}) \cdot \tau_{l,k_1,k_2,k_3}^{h,j}. \quad (8)$$

The correctness of these updates can be proved using the three conditional independence assumptions in our HMM. Due to space limit, we omit the proof here.

### B. Change-point detection on model variations

To measure the structure changes of the estimated HMM models over time, we use the concept of the symmetric relative entropy to define a model variation score. Denote the model at time $t-1$ and $t$ by $P_{t-1}$ and $P_t$ respectively, then the model variation score is defined to be

$$v_t = \lim_{n\to\infty} \frac{1}{n} D(P_t || P_{t-1}) + \lim_{n\to\infty} \frac{1}{n} D(P_{t-1} || P_t),$$

where $D(p||q)$ represents the relative entropy of distribution $p$ to $q$, and $n$ represents the length of the input data. It is natural to let $n \to \infty$ as we can then compare the two models under the stationary states in the limit of $n \to \infty$.

It can be proved that

$$\lim_{n\to\infty} \frac{1}{n} D(P_t||P_{t-1}) = \sum_{i=1}^{L} \frac{1}{2^i} D((\pi_{k_1|k_2,k_3}^{i})^t || (\pi_{k_1|k_2,k_3}^{i})^{t-1})$$
$$+ \sum_{i=1}^{L} \frac{1}{2^i} \sum_{k} (\pi_k^i)^t D(\mathcal{N}((\mu_k^i)^t, (\sigma_k^i)^t) || \mathcal{N}((\mu_k^i)^{t-1}, (\sigma_k^i)^{t-1})),$$

where $\pi_k^i = P(s_{i,j} = k)$. Therefore, besides the probability distributions $\pi_{k_1|k_2,k_3}^{i}$ and $\mathcal{N}(\mu_k^i, \sigma_k^i)$ provided by the online EM algorithm, the computation of $\lim_{n\to\infty} \frac{1}{n} D(P_t||P_{t-1})$ also involves the computation of the probability distributions $\pi_{k_1,k_2,k_3}^{i} = P(s_{i,j} = k_1, s_{i,j-1} = k_2, s_{i+1,\lceil j/2 \rceil} = k_3)$ and $\pi_k^i$. The estimation of $\pi_k^i$ and $\pi_{k_1,k_2,k_3}^{i}$ can be obtained from the sufficient statistics $\tau_{i,k}^{l,m}$ and $\tau_{i,k_1,k_2,k_3}^{l,m}$ as follows,

$$\pi_k^i = \sum_{\mathcal{S}_{l,m}} \tau_{i,k}^{l,m}, \quad \pi_{k_1,k_2,k_3}^{i} = \sum_{\mathcal{S}_{l,m}} \tau_{i,k_1,k_2,k_3}^{l,m}.$$

The other relative entropy $\lim_{n\to\infty} \frac{1}{n} D(P_{t-1}||P_t)$ can be computed similarly. We can see that the symmetric model variation score actually captures two types of changes. The first is the changes in the transition probabilities of the hidden states while the second is the changes in the generation pattern of the observed data from a fixed state. By using the symmetric relative entropy as a distance measure between two HMM models, it is expected that not only the changes of the data generation pattern will be detected, but also the changes in the hidden states can also be detected.

## VI. Evaluations

We evaluate the performance of our anomaly detection algorithm numerically. The algorithm performances are evaluated using the detection latency and the Receiver Operating Characteristic (ROC) curve, which is a plot of the detection rate versus the false alarm rate at different threshold values. The selection of the wavelet basis used in our anomaly detection scheme is based on a balance between its *time localization* and *frequency localization* characteristics [2]. In our experiments, we found that for the Daubechies family wavelets, the D2 (Haar wavelets) and D4 wavelets can give us reasonably good performance. Hence we use the Haar wavelets for all the experiments in this paper.

In the numerical experiments, two well known LRD time series including the Fractional Gaussian Noise (FGN) and the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model, are used for generating data. We then inject two types of model variations into the time series as anomalies. First is the mean level shift, i.e., a step function with a constant amplitude will be imposed on the original signal. Second is to vary model parameters for the data generation process, including the standard deviation and the Hurst parameter for FGN and ARFIMA. The duration for the normal state and the anomaly state is generated from exponential distribution with different mean values.

We first discuss the detection performance on mean level shifts in the synthetic LRD time series. Fig. 3 shows one representative example. The top figure illustrates the time series, which is generated from an ARFIMA model with Hurst parameter $0.9$ and length $2^{15}$. The standard deviation for the generated data sequence is set to be $1$. The mean level shift occurs at the first quarter of the time series and ends at the middle with intensity $0.75$, which is less than the standard deviation. The bottom two figures show the corresponding model variation scores computed by our online EM algorithm with 5-level and 4-level wavelet decomposition respectively. The $x$ axis represents the scaled time due to wavelet transform and the $y$ axis represents the model variation score.

From Fig. 3, we can see that the visual inspection of the injected mean level shift from the time series directly can be difficult. However, in the tracked model variation scores, there are abrupt changes at the two time locations where the mean level shift starts and ends. These two abrupt changes suggest where the injection starts and ends. Especially when

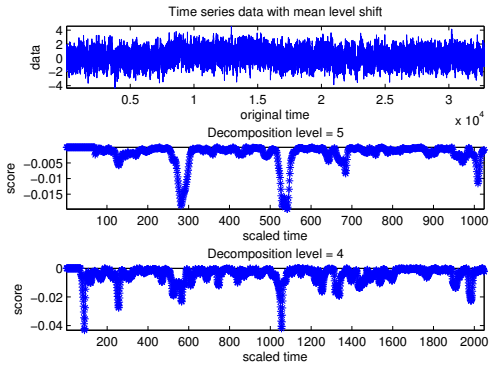


Fig. 3: Effects of different decomposition levels

the wavelet decomposition level is 5, these are the only 2 abrupt changes that exist. When the wavelet decomposition level is 4, there are some false alarms, therefore the decomposition level of the wavelet transform is an important parameter for the anomaly detection scheme. A higher level decomposition can give higher detection rate and smaller false alarm rate. However, since an L-level decomposition has a $2^L$ time aggregation scale, i.e., it transforms the data samples within a $2^L$ time window to the wavelet domain so the wavelet coefficients within this window is time-indistinguishable, a higher level decomposition would often give longer detection latency than that of a lower level decomposition. In our experiments, we found that a 5-level wavelet decomposition can give a reasonable good balance between detection accuracy and latency.

The intensity of the injected mean level shift also affects the detection performance. Fig. 4 shows the ROC curve and detection latency for the injected mean shift with different intensities. Each curve is obtained over 1000 simulation traces with the 5-level wavelet decomposition. As is expected, for higher injected mean level shifts, the detection becomes much easier, in terms of lower false alarms, higher detection rates, and smaller detection latency.

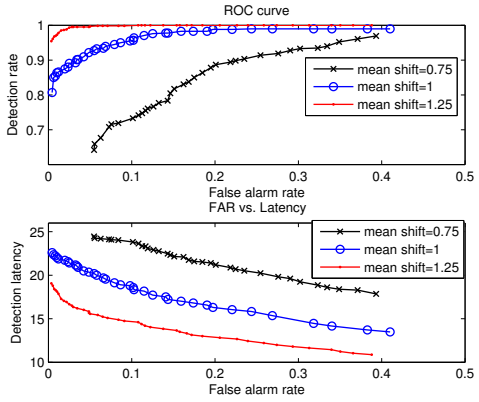


Fig. 4: Effects of different injected intensities

For the second type of anomalies, i.e., the one that varies model parameters for the data generation process, similar good performances are observed for our detection method. When the variation becomes larger, the detection becomes easier. Due to space limit, we omit the results here.

For performance comparison, we implement a baseline method adapted from [2], in which only the mean and variance of the wavelet coefficients is used for anomaly detection. It is observed that our method can always beat this baseline method. For example, Fig. 5 shows the ROC curves and detection latency for our method and the baseline methods on the detection of Hurst parameter changing from 0.9 to 0.7.

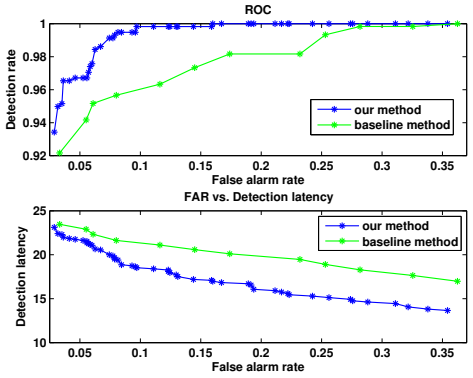


Fig. 5: Comparison of our method and the baseline method

## VII. Conclusions

In this paper, we studied the anomaly detection problem for LRD traffic in wireless sensor networks. We proposed a wavelet-domain hidden Markov model for capturing the properties of network traffic. The wavelet transform is able to turn the long range dependency that exists among the sample data into a short memory structure among its wavelet coefficients. The HMM in the wavelet-domain is used to further capture the remaining dependency among the wavelet coefficients, thus model the traffic variability more accurately. Network anomalies are then detected as abrupt changes in the tracked HMM model structures. We evaluate the performance of our algorithm numerically using typical LRD time series. In the future work, we plan to study the optimization of model parameters for the wavelet domain HMM model, in order to achieve better performance.

## References

[1] P. Wang and I. F. Akyildiz, "Spatial correlation and mobility aware traffic modeling for wireless sensor networks," in *Proceedings of Globecom*, 2009.

[2] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *ACM SIGCOMM Internet Measurement Workshop*, 2002.

[3] P. Zuraniewski and D. Rincon, "Wavelet transforms and chenge-point detection algorithms for tracking network traffic fractality," in *Proceedings of 2nd Conference on Next Generation Internet Design and Engineering*, 2006.

[4] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using trafc feature distributions," in *Proceedings of ACM SIGCOMM*, 2005.

[5] A. Hussain03, J. Heidemann, and C. Papadopoulos, "A framework for classifying denial of service attacks," in *Proceedings of ACM SIGCOMM*, 2003.

[6] S. Jin and D. Yeung, "A covariance analysis model for DDoS attack detection," in *Proceedings of IEEE ICC*, 2004.

[7] P. Abry and D. Veitch, "Wavelet analysis of long range dependent traffic," *IEEE Transactions of Information Theory*, 1998.

[8] L. Rabiner, "Non-Gaussian and long memory statistical characterisations for Internet traffic with anomalies," *IEEE Transactions on Dependable and Secure Computing*, 2007.

[9] Y. Gu, A. McCallum, and D. Towsley, "Detecting anomalies in network traffic using maximum entropy estimation," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, 2005.

[10] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Transactions on signal processing*, 1998.

[11] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.

[12] O. Cappe, "Online EM algorithm for hidden Markov models," 2009.