Statistical Anomaly Detection with Sensor Networks

IOANNIS CH. PASCHALIDIS
and
YIN CHEN
Boston University

We seek to detect statistically significant temporal or spatial changes in either the underlying process the sensor network is monitoring or in the network operation itself. These changes may point to faults, adversarial threats, misbehavior, or other anomalies that require intervention. To that end, we introduce a new statistical anomaly detection framework that uses Markov models to characterize the "normal" behavior of the sensor network. We develop a series of Markov models, including tree-indexed Markov chains which can model its spatial structure. For each model, an anomaly-free probability law is estimated from past traces. We leverage large deviations techniques to develop optimal anomaly detection rules for each corresponding Markov model, assessing whether its most recent empirical measure is consistent with the anomaly-free probability law. A series of simulation results, some with real sensor data, validate the effectiveness of the proposed anomaly detection algorithms.

Categories and Subject Descriptors: C.2.0 [**Computer-Communication Networks**]: Security and protection; G.3 [**Probability and Statistics**]: Markov processes

General Terms: Algorithms; Security; Theory

Additional Key Words and Phrases: Sensor networks, anomaly detection, large deviations.

1. INTRODUCTION

Wireless Sensor NETworks (WSNETs) have found many applications in the large-scale, automated, and intelligent monitoring of a plethora of physical systems, including industrial plants and machinery, homes and buildings, agricultural fields, and wildlife habitats. New innovative applications have been introduced that enhance standard monitoring, e.g., see Paschalidis and Guo [2010] and Ray et al. [2006] for work on localization. Another important class of application arises in

Ioannis Ch. Paschalidis is with the Department of Electrical and Computer Engineering, the Division of Systems Engineering, and the Center for Information and Systems Engineering, Boston University, 8 St. Mary's St., Boston, MA 02215, e-mail: yannisp@bu.edu, url: http://ionia.bu.edu/.
Yin Chen is with the Center for Information and Systems Engineering, Boston University, e-mail: yinchen@bu.edu.

Research partially supported by the NSF under grant EFRI-0735974, by the DOE under grant DE-FG52-06NA27490, and by the ARO under grant W911NF-09-1-0492.
A preliminary version of this paper appeared in Paschalidis and Chen [2008].

monitoring urban environments, water supplies, and public infrastructure, for instance, for homeland defense purposes.

In all these settings, one critical question that arises is that of detecting abnormalities (or anomalies) in the various quantities that are being measured. In some instances, simple threshold rules (e.g., a variable, or its average over a certain time-window, exceeding a certain value) may suffice to that end. Often, though, deviations from normal behavior may be subtle and may result in changes in the spatial or temporal probability distribution of measurements without affecting first (or even second) moments. Such changes in environments are much harder to detect, yet their detection is important as they may be precursors (either in time or space) of "larger" and more dangerous abnormalities.

A related problem is that of detecting anomalies in the operation of the WSNET itself, for instance, changes in routing patterns, connectivity, and other operational quantities that may indicate malfunctioning of the network due to natural or adversarial causes. In WSNETs, the use of wireless communications leaves them extremely vulnerable to a wide range of adversarial attacks, exploits, viruses, and other information security vulnerabilities (Perrig et al. [2004]). These vulnerabilities are not only due to the broadcast nature of wireless but also due to the severe energy limitations (especially in WSNETs) that preclude the use of very sophisticated coding and other prevention mechanisms. Following a traditional computer systems approach, most of the literature has focused on protocol design as a way to address these network security concerns (see Perrig et al. [2002], Chan et al. [2003], and Zhu et al. [2006]). Some work that is closer in spirit to ours and considers the distribution of sensor data and the divergence between distributions is presented by Subramaniam et al. [2006], but the authors took a different approach by integrating kernel estimators as part of their algorithm design.

The work in this paper considers anomaly detection in general enough terms that accommodate both the monitoring and network security related problems outlined above. Indeed, we present examples of applying our methodology to both domains. In particular, we define the notion of a "state" associated with a node of the network and assume that its evolution is Markovian. We consider a series of models that can model the evolution of the state both in time and in space, the latter using the connectivity model of the network. We pay particular attention to trees which is a common connectivity structure for WSNETs.

The key idea underlying our work is to compare empirical measures of recent activity with anomaly-free probability laws estimated from past activity. Specifically, assuming that we know the transition probability matrix of the Markov chain of interest, which can be estimated from past anomaly-free observations, we study the large deviations of the empirical measure obtained from recent observations. Large deviations theory provides a powerful way of handling rare events and their associated probabilities and results related to our models have been developed in Dembo and Zeitouni [1998] and Dembo et al. [2005]. If the empirical measure takes very unlikely values this points to a statistical anomaly. Using hypothesis testing techniques, for each one of our Markov models we develop appropriate anomaly detection tests and establish their optimality in a generalized Neyman-Pearson sense. In contrast to most of the work in the literature, we treat anomaly detection in a rigorous statistical framework that allows us to control the false alarm rate. Related techniques, using i.i.d. sequences of observations and much simpler Markovian models than the ones we adopt here, have also been applied in monitoring internet traffic with considerable success (Paschalidis and Smaragdakis [2008]). The novelty of our approach lies in the fact that we do not assume a particular distribution for the collected data, and given long enough observation, legal changes in the network will be captured and contribute to our characterization of "normal" behavior. Further, the proposed method does not distinguish different types of anomalies as they are well addressed individually in the literature (e.g., Hu et al. [2003], Lazos et al. [2005]). We mainly target general detection of anomalies regardless of their types, which has the advantage of been able to detect novel/unforeseen anomalies.

The remainder of this paper is organized as follows. In Sec. 2, we consider a node-level Markov chain model of a WSNET, develop the necessary large deviations results, and devise an anomaly detection test. In Sec. 3 and 4, we adopt two different tree-indexed Markov models, develop large deviations results for their empirical measure, and formulate the corresponding anomaly detection tests. For each model we consider, we present simulation results in Sec. 5, where we also test our approach in an application monitoring water temperature in a pond using real data. Conclusions are in Sec. 6.

2. A NODE-LEVEL MARKOV CHAIN MODEL

We start with a node-level Markov chain to model the propagation of events of interest in the WSNET. For the purposes of this section we assume that the WSNET is connected and every node can send a message to every other node, potentially over multiple hops.

Let the WSNET have n nodes. Nodes can assume one of two states: 0 and 1. When node i observes an event of interest it switches from state 0 to state 1 and stays in that state for as long as the conditions that constitute the event persist.

Here are some motivating examples: a WSNET tracking an object as it moves through the coverage area of the network, or monitoring events generated by the nodes and relayed to a gateway. A related example concerns the routing of packets through the network: each node is in state 1 if it has packets to send (self-generated or received from another node) and in state 0 otherwise.

Let now the state of the WSNET be the vector of the states its nodes assume, i.e., a vector of length n with each element 0 or 1. To reduce the size of the state space, in case of a large network, we can partition the nodes into several groups, each with a clusterhead in charge of the state updates. Let $\Sigma = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_M\}$ be the state space, i.e., a set of vectors with length n and elements 0 or 1. Note that M can be much less than 2^n, due to the fact that a large portion of vectors will not occur in a real WSNET setup. We use $q_0(\mathbf{z}_j|\mathbf{z}_i)$ to denote the transition probability from WSNET state $\mathbf{z}_i$ to $\mathbf{z}_j$ and denote by $\mathbf{Q}_0$ the corresponding $M \times M$ transition probability matrix, namely, $\mathbf{Q}_0 = (q_0(\mathbf{z}_j|\mathbf{z}_i))_{i,j=1}^{M}$. Assume that $\mathbf{Q}_0$ is irreducible and aperiodic with a unique stationary distribution $\boldsymbol{\pi}_0 = (\pi_1^0, \ldots, \pi_M^0)$, where π_i^0 is the steady-state fraction of time the Markov chain is in state $\mathbf{z}_i$.

We are interested in detecting changes in the steady-state distribution of the Markov chain; as we explained in the introduction these may correspond to anomalies. For the routing example above these anomalies may indicate some change in the "typical" routing pattern of the network which may be caused by natural (e.g., some wireless link is down) or adversarial reasons (e.g., interference or some other attack to the network). The transition probability matrix $\mathbf{Q}_0$ can be easily estimated from a long sequence of past observations. Given a recent trace (i.e., sequence) of states $\mathbf{Y}_t = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t)$ the Markov chain visits we seek to detect whether this sequence has been generated from the law $\mathbf{Q}_0$ or from some other (unknown) law $\mathbf{Q}_1$. The problem at hand is a *composite hypothesis testing* problem as we seek to differentiate between a known law $\mathbf{Q}_0$ (hypothesis H_0) and an unknown law $\mathbf{Q}_1$ (hypothesis H_1). A decision test can be defined as follows.

Definition 1
A decision test $\mathscr{S}$ is a sequence of maps $\mathscr{S}^t : \Sigma^t \to \{0, 1\}$, with the interpretation that when $\mathbf{Y}_t = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ is observed, then H_0 is accepted (H_1 rejected) if $\mathscr{S}(\mathbf{Y}_t) = 0$, and H_1 is accepted (H_0 rejected) if $\mathscr{S}(\mathbf{Y}_t) = 1$.

The performance of a decision test $\mathscr{S}$ is characterized by the type I and type II, respectively, error probabilities

$$\alpha_t \triangleq \mathbf{P}_{\mathbf{Q}_0}[\mathscr{S}^t \text{ rejects } H_0], \quad \beta_t \triangleq \mathbf{P}_{\mathbf{Q}_1}[\mathscr{S}^t \text{ rejects } H_1],$$

where $\mathbf{P}_{\mathbf{Q}_i}$ denotes a probability evaluated under law $\mathbf{Q}_i$. Since we can not minimize both error probabilities at the same time, we consider the following optimality criterion known as the *generalized Neyman-Pearson criterion* (Hoeffding [1965]).

Definition 2
A test $\mathscr{S}$ is optimal (for a given $\eta > 0$) if, among all tests that satisfy

$$\limsup_{t\to\infty} \frac{1}{t} \log \alpha_t \leq -\eta, \tag{1}$$

the test $\mathscr{S}$ maximizes the asymptotic exponent of the type II error probability, i.e., uniformly over all possible possible laws $\mathbf{Q}_1$, $-\limsup_{t\to\infty} \frac{1}{t} \log \beta_t$ is maximal.

In the case where $\mathbf{y}_i$ are i.i.d., Hoeffding [1965] has proposed a simple test that compares the relative entropy between the empirical measure (or type) of $\mathbf{Y}_t$ and the anomaly-free law to a threshold η in order to decide between H_0 and H_1. Zeitouni et al. [1992] have shown that a natural generalization of Hoeffding's test to the case of general Markov sources is optimal according to the criterion in Definition 2.

Specifically, define $\mu_t^{\mathbf{Y}}(\mathbf{w}, \mathbf{v})$ as the empirical joint 2-step occurrence of the system states

$$\mu_t^{\mathbf{Y}}(\mathbf{w}, \mathbf{v}) = \frac{1}{t} \sum_{k=1}^{t} 1\{\mathbf{y}_{k-1} = \mathbf{w}, \mathbf{y}_k = \mathbf{v}\}, \quad \mathbf{w}, \mathbf{v} \in \Sigma$$

where $1\{\cdot\}$ denotes the indicator function. $\mu_t^{\mathbf{Y}}(\mathbf{w}, \mathbf{v})$ can be interpreted as the frequency that the pair $(\mathbf{w}, \mathbf{v})$ of consecutive states appears in the trace $\mathbf{Y}_t$. We will write $\boldsymbol{\mu}_t^{\mathbf{Y}}$ for the vector of all $\mu_t^{\mathbf{Y}}(\mathbf{w}, \mathbf{v})$ and we will use the same convention of denoting vectors with bold letters throughout the paper. The marginals of $\boldsymbol{\mu}_t^{\mathbf{Y}}$ are denoted by the vectors $\boldsymbol{\mu}^{\mathbf{Y}}_{L,t}$ and $\boldsymbol{\mu}^{\mathbf{Y}}_{R,t}$ with elements

$$\mu^{\mathbf{Y}}_{L,t}(\mathbf{w}) = \sum_{\mathbf{v}\in\Sigma} \mu^{\mathbf{Y}}_t(\mathbf{w},\mathbf{v}), \quad \mu^{\mathbf{Y}}_{R,t}(\mathbf{w}) = \sum_{\mathbf{v}\in\Sigma} \mu^{\mathbf{Y}}_t(\mathbf{v},\mathbf{w}).$$

Without loss of generality, we assume that they are identical (then $\boldsymbol{\mu}^{\mathbf{Y}}_t$ is called *shift invariant*). The empirical transition probability from state $\mathbf{w}$ to state $\mathbf{v}$ is defined as

$$\mu^{\mathbf{Y}}_t(\mathbf{v}|\mathbf{w}) = \frac{\mu^{\mathbf{Y}}_t(\mathbf{w},\mathbf{v})}{\mu^{\mathbf{Y}}_{L,t}(\mathbf{w})},$$

with the convention that $0/0$ equals 0.

Define $\boldsymbol{\mu}^{\mathbf{Y}}_{L,t} \otimes \mathbf{Q}_0$ as the vector with elements $\mu^{\mathbf{Y}}_{L,t}(\mathbf{w})q_0(\mathbf{v}|\mathbf{w})$, $\mathbf{w},\mathbf{v} \in \Sigma$, and consider the divergence (relative entropy) between $\boldsymbol{\mu}^{\mathbf{Y}}_t$ and $\boldsymbol{\mu}^{\mathbf{Y}}_{L,t} \otimes \mathbf{Q}_0$:

$$\begin{aligned} D(\boldsymbol{\mu}^{\mathbf{Y}}_t || \boldsymbol{\mu}^{\mathbf{Y}}_{L,t} \otimes \mathbf{Q}_0) &= \sum_{\mathbf{w},\mathbf{v}\in\Sigma} \mu^{\mathbf{Y}}_t(\mathbf{w},\mathbf{v}) \log \frac{\mu^{\mathbf{Y}}_t(\mathbf{w},\mathbf{v})}{\mu^{\mathbf{Y}}_{L,t}(\mathbf{w})q_0(\mathbf{v}|\mathbf{w})} \\ &= \sum_{\mathbf{w}\in\Sigma} \mu^{\mathbf{Y}}_{L,t}(\mathbf{w}) \sum_{\mathbf{v}\in\Sigma} \mu^{\mathbf{Y}}_t(\mathbf{v}|\mathbf{w}) \log \frac{\mu^{\mathbf{Y}}_t(\mathbf{v}|\mathbf{w})}{q_0(\mathbf{v}|\mathbf{w})} \\ &= \sum_{\mathbf{w}\in\Sigma} \mu^{\mathbf{Y}}_{L,t}(\mathbf{w}) D(\boldsymbol{\mu}^{\mathbf{Y}}_t(\cdot|\mathbf{w}) || \mathbf{q}_0(\cdot|\mathbf{w})). \end{aligned} \tag{2}$$

Zeitouni et al. [1992] establish the following result.

Theorem 2.1 (Zeitouni et al. [1992]) *The decision test*

$$\mathscr{S}^*_1(\mathbf{Y}_t) = \begin{cases} 0, & \textit{if } D(\boldsymbol{\mu}^{\mathbf{Y}}_t || \boldsymbol{\mu}^{\mathbf{Y}}_{L,t} \otimes \mathbf{Q}_0) < \eta, \\ 1, & \textit{otherwise,} \end{cases}$$

is optimal according to the generalized Neyman-Pearson criterion of Definition 2.

This theorem provides an optimal anomaly detection test for the node-level Markov chain model we considered in this section. The threshold η is user-defined and represents the user's tolerance for false alarms. In particular, the false alarm probability α_t is bounded above by $e^{-t\eta}$ for large enough t and the user can set $\eta = -(\log \epsilon)/t$ to ensure that the false alarm probability stays below ϵ. The proposed test is optimal in the sense that it maximizes the exponential (with t) decay rate of the misdetection probability β_t among all tests with a false alarm probability bounded above by ϵ.

3. A TREE-INDEXED MARKOV CHAIN MODEL: THE EDGE-WISE CASE

Next we consider a more general Markov model that accounts for the connectivity structure of the WSNET. In particular, we consider tree-indexed Markov chains since in many WSNET implementations (for example the popular TinyOS platform) the multihop network formed by the sensor nodes adopts a tree structure. The tree will be formed randomly according to an arbitrary probability law we will specify. Generalizing the model of the previous section, every node of the WSNET (i.e., the tree) can be in one out of a finite number of states. We will assume, though, that the state of the children of a node is selected conditional on the state of the parent according to some Markov chain indexed by the nodes of the tree.

A model of this type can model the propagation of events up or down the tree. In a monitoring application, suppose that a WSNET node i measures a vector $\mathbf{x}_i$ of quantities in its environment and passes information on $\mathbf{x}_i$ to all nodes j communicating with it. Consider a fixed time interval $[0, t]$ and define the state of a node depending on the average value of $\mathbf{x}_i$ in $[0, t]$ and the corresponding values $\mathbf{x}_j$ of nodes that communicate with i. As a result, the state of the children is influenced by the state of the parent and vice versa. Similarly, one could also model information flow in the network by defining the state of a node to be the average flow through the node during $[0, t]$. It is important to note that given the way we generate the random tree, each node keeps track of the types of its children, and the calculation of empirical measure can be done distributively following the leaf-root path and propagating the lower-level information recursively to the gateway.

As in Section 2, we are interested in identifying statistical anomalies (i.e., distributional changes) in the Markov model representing the WSNET. In the monitoring application such anomalies can point to changes in the underlying processes the WSNET is monitoring. Similarly, in the case when the state is defined based on the flow through the node, anomalies identify disruption in the routing of the WSNET and may correspond to attacks.

The technical development of an optimal anomaly detection test is based on large deviations results for Markov chains indexed by random trees derived by Dembo et al. [2005]. We consider trees conditioned to have exactly n nodes and then take the large deviations limit as $n \to \infty$.

3.1 Large deviations results: edge-wise case

We start with the simpler case where the number of children for each node of the tree is drawn independently from an arbitrary discrete probability distribution and only the state of the children depends on the state of the parent. In Section 4 we will consider the more general case where both the number of children and the corresponding states depend on the state of the parent.

The tree-indexed Markov chain studied in this section is generated as follows. Suppose that $T = (\rho, \mathscr{V}, \mathscr{E})$ is any finite tree with root ρ and sets of vertices (nodes) and edges denoted by $\mathscr{V}$ and $\mathscr{E}$, respectively. Each node of the tree is in a state selected from a finite set $\mathscr{X}$. We use $X(i)$ to denote the state of node i. Without loss of generality we can let $\mathscr{X} = \{1, \ldots, m\}$. We are given a discrete probability law $\boldsymbol{\nu}$ on $\mathscr{X}$ and a Markovian $m \times m$ transition probability matrix $\mathbf{Q}_0 = (q_0(b|a))_{a,b=1}^{m}$. We first construct the random tree starting from the root ρ and selecting the number of children $N(v)$ for each node $v \in \mathscr{V}$ independently of every other node and according to a discrete probability distribution $p(\cdot) = \mathbf{P}[N(v) = \cdot]$ such that $0 < p(0) < 1$. We then assign a state to each node by first drawing the state $X(\rho)$ of the root according to $\boldsymbol{\nu}$ and then selecting the state $X(v)$ of every node v conditional on the state of its parent by using the transition probability matrix $\mathbf{Q}_0$.

Consider now a finite instance of the random tree and a realization (sample path) X of the tree-indexed Markov chain. Define the empirical measure of X as the m^2-dimensional vector $\mathbf{L}_X$ with elements

$$L_X(a,b) = \frac{1}{|\mathscr{E}|} \sum_{(v_1,v_2)\in\mathscr{E}} 1\{X(v_1)=a, X(v_2)=b\}, \tag{3}$$

for each $a, b \in \mathscr{X}$, where (v_1, v_2) denotes an edge of the tree between parent v_1 and child v_2.

Dembo et al. [2005] establish a large deviations result for $\mathbf{L}_X$ for trees conditioned to have n nodes. The result assumes that the tree is *critical*, that is, the mean number of children of each node is 1. As we will see this assumption can be easily relaxed. In preparation for the result, for each probability law $\boldsymbol{\mu}$ on $\mathscr{X} \times \mathscr{X}$ (an m^2-dimensional vector) we let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ denote the two marginals so that

$$\mu_1(a) = \sum_{b=1}^{m} \mu(a,b), \quad \mu_2(a) = \sum_{b=1}^{m} \mu(b,a).$$

Let $I_p(\cdot)$ denote the convex dual of the log-moment generating function of the offspring law $p(\cdot)$, namely,

$$I_p(x) = \sup_{\lambda\in\mathbb{R}} \left\{ \lambda x - \log\left(\sum_{n=0}^{\infty} p(n) e^{\lambda n} \right) \right\}.$$

It is well known (Cramér's theorem, see Dembo and Zeitouni [1998]) that $I_p(\cdot)$ is the large deviations rate function associated with the law $p(\cdot)$. Finally, as in Section 2, define $\boldsymbol{\mu}_1 \otimes \mathbf{Q}_0$ as the vector with elements $\mu_1(a) q_0(b|a)$, $a, b = 1, \dots, m$, and let $<<$ denote pointwise strict inequality between vectors.

Theorem 3.1 (Dembo et al. [2005]) *Suppose that T is a tree with offspring law $p(\cdot)$ such that $0 < p(0) < 1 - p(1)$, $\sum_l lp(l) = 1$ and $l^{-1} \log p(l) \to -\infty$. Let X be a Markov chain indexed by T with an arbitrary initial distribution and an irreducible Markovian matrix $\mathbf{Q}_0$. Then, for $n \to \infty$, the empirical pair measure $\mathbf{L}_X$, conditioned on $\{|\mathscr{V}| = n\}$ satisfies a large deviation principle in the space of probability vectors on $\mathscr{X} \times \mathscr{X}$ with speed n and the convex, good rate function*

$$I(\boldsymbol{\mu}) = \begin{cases} D(\boldsymbol{\mu}||\boldsymbol{\mu}_1 \otimes \mathbf{Q}_0) + \sum_{a=1}^{m} & \mu_2(a) I_p\left(\frac{\mu_1(a)}{\mu_2(a)}\right) \\ & \textit{if } \boldsymbol{\mu}_1 << \boldsymbol{\mu}_2, \\ \infty & \textit{otherwise.} \end{cases} \tag{4}$$

In (4) $D(\cdot||\cdot)$ is the relative entropy between two probability vectors defined as in (2). Note that the first term in the rate function above characterizes large deviations of the assignment of states to the nodes of the tree while the second term is related to the structure of the tree.

The following lemma is useful in relaxing the assumption that the tree is critical.

Lemma 3.2 *Suppose that $\sum_l lp(l) \neq 1$. The distribution of T conditioned on $\{|\mathscr{V}| = n\}$ under the offspring law $p(\cdot)$ does not change when we we use the offspring law $p_\theta(l) = p(l)e^{\theta l} / (\sum_j p(j) e^{\theta j})$ for any value $\theta \in \mathbb{R}$.*

Proof. Fix $\theta \in \mathbb{R}$, $|\mathscr{V}| = n$, and assume we have K different tree structures $T_1(\rho_1, \mathscr{V}_1, \mathscr{E}_1), \ldots, T_K(\rho_K, \mathscr{V}_K, \mathscr{E}_K)$ with n nodes. Set $Z = \sum_j p(j)e^{\theta j}$ and let $\mathbf{P}_\theta[\cdot]$ and $\mathbf{P}[\cdot]$ denote probabilities under the laws $p_\theta(\cdot)$ and $p(\cdot)$, respectively. For any $i, j = 1, \ldots, K$ we have

$$\begin{aligned}\frac{\mathbf{P}_\theta[T_i]}{\mathbf{P}_\theta[T_j]} &= \frac{\prod_{v \in \mathscr{V}_i} p(N(v))e^{\theta N(v)}/Z^n}{\prod_{v \in \mathscr{V}_j} p(N(v))e^{\theta N(v)}/Z^n} \\ &= \frac{e^{\theta \sum_{v \in \mathscr{V}_i} N(v)} \prod_{v \in \mathscr{V}_i} p(N(v))}{e^{\theta \sum_{v \in \mathscr{V}_j} N(v)} \prod_{v \in \mathscr{V}_j} p(N(v))} \\ &= \frac{e^{\theta n} \prod_{v \in \mathscr{V}_i} p(N(v))}{e^{\theta n} \prod_{v \in \mathscr{V}_j} p(N(v))} = \frac{\mathbf{P}[T_i]}{\mathbf{P}[T_j]}.\end{aligned}$$

It follows that the conditional distribution on $\{|\mathscr{V}| = n\}$ is identical under either the original or the θ-twisted offspring law. □

We can use the above result to handle non-critical trees. In particular, we twist the offspring law by θ as described in the statement of Lemma 3.2. Note that with $0 < p(0) < 1 - p(1)$ there exists a unique θ^* such that $\sum_l l p_{\theta^*}(l) = 1$. Hence, Theorem 3.1 holds under $p_{\theta^*}(\cdot)$ which implies that the large deviations rate function for non-critical trees is given by (4) when we replace $I_p(\cdot)$ with $I_{p_{\theta^*}}(\cdot)$.

3.2 Anomaly detection test: edge-wise case

Next we will develop an anomaly detection test and show that it is optimal in a generalized Neyman-Pearson sense.

Given a long sequence of realizations X^k of a tree-index Markov chain defined on a tree T with n nodes we can approximate the offspring law $p(\cdot)$ and the transition probability matrix $\mathbf{Q}_0$ by taking the average frequencies of the corresponding samples. In particular, if $\mathbf{L}_{X^k}$ denotes the empirical measure of the kth realization (cf. (3)) then $(\frac{1}{k}\sum_{l=1}^{k} L_{X^l}(a, b))/(\frac{1}{k}\sum_{l=1}^{k}\sum_{b=1}^{m} L_{X^l}(a, b))$ converges to $q_0(b|a)$ with probability one (w.p.1) as $k \to \infty$. Alternatively, one can compute the frequencies on a single large tree as $n \to \infty$.

Assuming that we have (or have estimated) $p(\cdot)$ and $\mathbf{Q}_0$ we are interested in a test that determines whether a particular realization (sample) X is typical or not. That is, as in Sec. 2 we want to differentiate between $p(\cdot)$ and $\mathbf{Q}_0$ (hypothesis H_0) and any other unknown law (hypothesis H_1).

Let us denote by $\mathbf{L}_X^n$ the empirical measure of X derived as in (3), where the superscript n indicates that the tree has n nodes. Using similar terminology and notation as in Section 2 the following theorem provides the test and establishes its optimality.

Theorem 3.3 *The decision test $\mathscr{S}_2^{*,n}(X)$*

$$\mathscr{S}_2^{*,n}(X) = \begin{cases} 0, & \textit{if } I(\mathbf{L}_X^n) < \eta, \\ 1, & \textit{otherwise.} \end{cases}$$

is optimal according to the generalized Neyman-Pearson criterion.

Proof. We start by establishing that the type I error exponent is no smaller than η (cf. (1)). Define the set of "abnormal samples" according to this test as the set of realizations with empirical measures resulting in $\mathscr{S}_2^{*,n}(X)=1$, namely

$$\Lambda_n^* = \{X \mid I(\mathbf{L}_X^n) \geq \eta\}.$$

Letting $\mathbf{P}_0[\cdot]$ denote probabilities under $p(\cdot)$ and $\mathbf{Q}_0$ and using Theorem 3.1 we have

$$\begin{aligned}\alpha_n =& \mathbf{P}_0[\mathscr{S}_2^{*,n}(X)=1]\\ =& \mathbf{P}_0[\Lambda_n^*]\\ \leq& \sum_{\{\boldsymbol{\mu} \mid I(\boldsymbol{\mu})\geq\eta\}} \mathbf{P}_0[\mathbf{L}_X^n = \boldsymbol{\mu}]\\ \leq& n^{m^2} \max_{\{\boldsymbol{\mu} \mid I(\boldsymbol{\mu})\geq\eta\}} \mathbf{P}_0[\mathbf{L}_X^n = \boldsymbol{\mu}]\\ \leq& n^{m^2} \exp\{-n \min_{\{\boldsymbol{\mu} \mid I(\boldsymbol{\mu})\geq\eta\}} I(\boldsymbol{\mu}) + n\epsilon\}\\ \leq& n^{m^2} \exp\{-n\eta + n\epsilon\}. \end{aligned} \tag{5}$$

The first inequality above is due to the union bound. The 2nd inequality is due to the fact that the empirical measure $\mathbf{L}_X^n$ can take n^{m^2} values. The 3rd inequality above is due to Theorem 3.1 and holds for every $\epsilon > 0$ and large enough n. Eq. (5) implies

$$\limsup_{n\to\infty} \frac{1}{n}\alpha_n \leq -\eta,$$

which establishes (1).

Next consider any arbitrary decision rule $\mathscr{S}^n$ that satisfies (1) and let Λ_n be the "abnormal" sample set associated with it. Since we are dealing with a Markov chain the empirical measure is a sufficient statistic for constructing an optimal decision rule. Thus, without loss of generality, we can assume that membership in Λ_n only depends on the empirical measure. Since $\mathscr{S}^n$ satisfies (1), for all $\epsilon > 0$, large enough n, and any X with empirical measure $\mathbf{L}_X^n$ that belongs to Λ_n we have

$$2^{-n(\eta+\epsilon)} \geq \mathbf{P}_0[\Lambda_n] \geq \exp\{-nI(\mathbf{L}_X^n) - n\epsilon\}, \tag{6}$$

where the last inequality is due to Theorem 3.1. It follows that $X \in \Lambda_n^*$ as well, that is, $\Lambda_n \subseteq \Lambda_n^*$. We conclude that if $\mathbf{P}_1[\cdot]$ denotes probabilities under any other distributions $p_1(\cdot)$ and $\mathbf{Q}_1$ satisfying (1), then

$$\mathbf{P}_1[X \notin \Lambda_n^*] \leq \mathbf{P}_1[X \notin \Lambda_n],$$

which suggests that the test $\mathscr{S}_2^{*,n}(X)$ minimizes the type II error probability. □

4. TREE-INDEXED MARKOV CHAIN MODEL: THE LEVEL-WISE CASE

In this section we consider the most general case where both the number of children and their states depend on the state of the parent.

To motivate this model, consider an example where a WSNET is deployed hierarchically and each parent node is responsible for a certain geographical area while the children nodes are assigned to subareas. In a such a situation, if the parent enters specific states indicating “alert” it may call upon more (mobile) nodes to be deployed in its area. This suggests that the number and the state of the children may depend on the state of the parent.

4.1 Large deviations results: level-wise case

As in the previous Section we let $\mathscr{X} = \{1, \ldots, m\}$ denote the set of states for every node of the tree. We construct a random tree $T = (\rho, \mathscr{V}, \mathscr{E})$ as follows. We first select the state $X(\rho)$ of the root according to some probability law $\boldsymbol{\nu}$ on $\mathscr{X}$. The offspring of any node $v \in \mathscr{V}$ is characterized by an element of $\mathscr{X}^* = \cup_{n=0}^{\infty} \{n\} \times \mathscr{X}^n$. Specifically, for each node v we denote by

$$C(v) = (N(v), X_1(v), \ldots, X_{N(v)}(v)) \in \mathscr{X}^* \tag{7}$$

the number and the types of the children of v, ordered from left to right. For each node v with state $X(v) = a$, $C(v)$ is drawn independently of everything else but conditional on the state a according to a Markovian transition kernel $\mathbf{Q}_0$ from $\mathscr{X}$ to $\mathscr{X}^*$. We write

$$\mathbf{Q}_0\{(n, x_1, \ldots, x_n)|a\} = \mathbf{P}[(N, X_1, ..., X_N) = (n, x_1, \ldots, x_n)|a]$$

for the probability of having n children with states $x_1, \ldots, x_n$, respectively, conditional on the state a of the parent.

Consider now a realization X of a tree generated as described above. We define the empirical measure $\mathbf{M}_X$ of X as a measure on $\mathscr{X} \times \mathscr{X}^*$ so that

$$M_X(a, c) = \frac{1}{|\mathscr{V}|} \sum_{v \in \mathscr{V}} 1\{X(v) = a, C(v) = c\}. \tag{8}$$

Dembo et al. [2005] establish a large deviations result for $\mathbf{M}_X$ for trees conditioned on having n nodes. To state the result we now introduce some additional notation.

For every $c = (n, a_1, \ldots, a_n) \in \mathscr{X}^*$ and $a \in \mathscr{X}$, denote the multiplicity of the symbol a in c by

$$m(a, c) = \sum_{i=1}^{n} 1\{a_i = a\},$$

and define the matrix $\mathbf{A} \in \mathbb{R}^{m^2}$ with (nonnegative) elements

$$A(a, b) = \sum_{c \in \mathscr{X}^*} Q\{c|b\} m(a, c), \text{ for } a, b \in \mathscr{X}.$$

Namely, $A(a, b)$ is expected number of type a children of a type b node. Let $\mathscr{G}(\mathbf{A})$ denote the directed graph with m nodes associated with the matrix $\mathbf{A}$ so that there is a directed link from node a to node b if and only if $\mathbf{A}(a, b) > 0$. We will say that $\mathbf{A}$ is weakly irreducible if we can partition $\mathscr{X}$ into a recurrent and transient subset, denoted by $\mathscr{X}_r$ and $\mathscr{X}_t$, respectively, so that (i) for any node a of $\mathscr{G}(\mathbf{A})$ there is a directed path to any node b of $\mathscr{G}(\mathbf{A})$ if $b \in \mathscr{X}_r$, and (ii) there is no directed path from any node a of $\mathscr{G}(\mathbf{A})$ to a node $b \in \mathscr{X}_t$ of $\mathscr{G}(\mathbf{A})$ if either $a = b$ or $a \in \mathscr{X}_r$. We will call the tree-indexed Markov chain X weakly irreducible if the corresponding $\mathbf{A}$ is weakly irreducible and the number of transient children, given by $\sum_{a \in \mathscr{X}_t} m(a, c)$ is uniformly bounded under the law $\mathbf{Q}_0$. We will also say that X is critical if the largest eigenvalue of $\mathbf{A}$ (which is real and positive due to irreducibility) is equal to 1.

For every probability measure $\boldsymbol{\sigma}$ on $\mathscr{X} \times \mathscr{X}^*$, let $\boldsymbol{\sigma}_1$ the $\mathscr{X}$-marginal of $\boldsymbol{\sigma}$, i.e., $\sigma_1(a) = \sum_{c \in \mathscr{X}^*} \sigma(a,c)$. We call $\boldsymbol{\sigma}$ shift-invariant if

$$\sigma_1(a) = \sum_{(b,c) \in \mathscr{X} \times \mathscr{X}^*} m(a,c)\sigma(b,c), \; \forall a \in \mathscr{X}.$$

Using similar notation as in Sec. 3 we denote by $\boldsymbol{\sigma}_1 \otimes \mathbf{Q}_0$ the law specified by $(\boldsymbol{\sigma}_1 \otimes \mathbf{Q}_0)(a,c) = \sigma_1(a) Q_0(c|a)$ for all $a \in \mathscr{X}$ and $c \in \mathscr{X}^*$. The following theorem from Dembo et al. [2005] states the large deviations result for the empirical measure $\mathbf{M}_X$.

Theorem 4.1 (Dembo et al. [2005]) *Suppose that X is a weakly irreducible and critical tree-indexed Markov chain X with an offspring $\mathbf{Q}_0$ law whose exponential moments are all finite, conditioned to have exactly n vertices. Then, for $n \to \infty$, the empirical measure $\mathbf{M}_X$ satisfies a large deviation principle in the space of probability measures in $\mathscr{X} \times \mathscr{X}^*$ with speed n and the convex, good rate function*

$$J(\boldsymbol{\sigma}) = \begin{cases} D(\boldsymbol{\sigma} | \boldsymbol{\sigma}_1 \otimes \mathbf{Q}_0), & \textit{if } \boldsymbol{\sigma} \textit{ is shift-invariant,} \\ \infty, & \textit{otherwise.} \end{cases} \tag{9}$$

4.2 Anomaly detection test: level-wise case

Next we develop an anomaly detection test and show that it is optimal in a generalized Neyman-Pearson sense.

As we described in Sec. 3.2, given a long sequence of realizations of the tree-indexed Markov chain we can approximate the law $\mathbf{Q}_0$ by computing the corresponding frequencies. Assuming that we have (or have estimated) $\mathbf{Q}_0$ we are interested in a test that determines whether a particular realization (sample) X is typical or not. Let us denote by $\mathbf{M}_X^n$ the empirical measure of X derived as in (8), where the superscript n indicates that the tree has n nodes. Using similar terminology and notation as in Sec. 3.2 the following theorem provides the test and establishes its optimality.

Theorem 4.2 *The decision test $\mathscr{S}_3^{*,n}(X)$*

$$\mathscr{S}_3^{*,n}(X) = \begin{cases} 0, & \textit{if } J(\mathbf{M}_X^n) < \eta, \\ 1, & \textit{otherwise.} \end{cases}$$

is optimal according to the generalized Neyman-Pearson criterion.

Proof. We will follow the structure of the proof of Thm. 4.2. We first establish that the type I error exponent is no smaller than η. Define the set of "abnormal samples" according to this test as the set of realizations with empirical measures resulting in $\mathscr{S}_3^{*,n}(X) = 1$, namely

$$\Lambda_n^* = \{X \mid J(\mathbf{M}_X^n) \geq \eta\}.$$

Let $\mathbf{P}_0[\cdot]$ denote probabilities under $\mathbf{Q}_0$. Note that membership of a realization X in Λ_n^* depends only the empirical measure $\mathbf{M}_X^n$. It follows

$$\begin{aligned}\alpha_n =& \mathbf{P}_0[\mathscr{S}_3^{*,n}(X) = 1] \\ =& \mathbf{P}_0[\Lambda_n^*] \\ =& \mathbf{P}_0[\mathbf{M}_X^n \in \{\boldsymbol{\sigma} \mid J(\boldsymbol{\sigma}) \geq \eta\}].\end{aligned}$$

Using Thm. 4.1 and the above we obtain

$$\limsup_{n\to\infty} \frac{1}{n}\alpha_n \leq - \inf_{\{\boldsymbol{\sigma} \mid J(\boldsymbol{\sigma})\geq\eta\}} J(\boldsymbol{\sigma}) = -\eta,$$

which establishes (1).

Next consider any arbitrary decision rule $\mathscr{S}^n$ that satisfies (1) and let Λ_n be the “abnormal” sample set associated with it. As in the proof of Thm. 4.2 the empirical measure is a sufficient statistic for constructing an optimal decision rule. Thus, without loss of generality, we can assume that membership in Λ_n only depends on the empirical measure. Since $\mathscr{S}^n$ satisfies (1), for all $\epsilon > 0$, large enough n, and any X with empirical measure $\mathbf{M}_X^n$ that belongs to Λ_n we have

$$2^{-n(\eta+\epsilon)} \geq \mathbf{P}_0[\Lambda_n] \geq \exp\{-nJ(\mathbf{M}_X^n) - n\epsilon\},$$

where the last inequality is due to Theorem 4.1. It follows that $X \in \Lambda_n^*$, which implies $\Lambda_n \subseteq \Lambda_n^*$. Thus, if $\mathbf{P}_1[\cdot]$ denotes probabilities under any other law $\mathbf{Q}_1$ satisfying (1) we have

$$\mathbf{P}_1[X \notin \Lambda_n^*] \leq \mathbf{P}_1[X \notin \Lambda_n],$$

which suggests that the test $\mathscr{S}_3^{*,n}(X)$ minimizes the type II error probability. □

5. NUMERICAL RESULTS

In this section we validate the results we obtained for each of the three models we considered using simulation. We elected not to perform experiments using WSNET devices because our emphasis is on algorithm design and the purpose of the numerical results was to validate the detection algorithms; a task that can be accomplished using simulation. However, our simulation scenarios used standard routing protocol assumptions and mechanisms from the TinyOS operating system employed by many WSNET platforms.

5.1 Results for the node-level Markov model

Consider a small network with size $n = 10$ as shown in Fig. 1. Assume that packets are generated at each node according to identical independent Poisson processes with rate 0.15. Once a packet is generated it is routed to the root. We generate a malicious attack which interferes with a link in the network. As a result, the downstream node at the jammed link is forced to switch its parent, thus changing the routing pattern in the network.

We use a long trace (time window size = 5000) of observations before the attacks to estimate the anomaly-free law $\mathbf{Q}_0$ and then apply the test of Thm. 2.1. The detection window was set to $t = 100$ and the detection threshold to $\eta = 0.03$, which results in a type I error probability equal to $e^{-t\eta} = 0.05$. Fig. 2 shows the time it

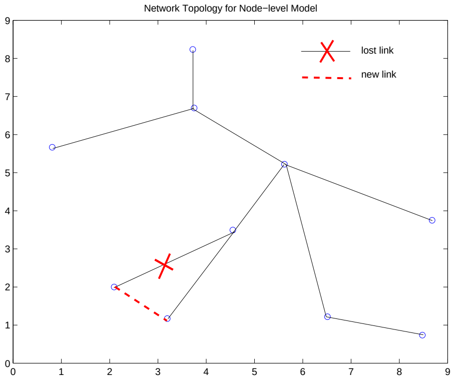

Fig. 1. Network topology for the node-level Markov chain model. A node is forced to switch parent due to the jammed link.

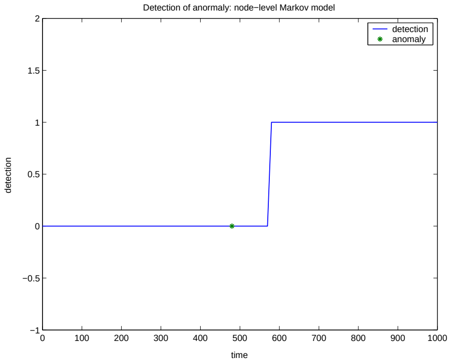

Fig. 2. Anomaly detection in the node-level Markov chain model.

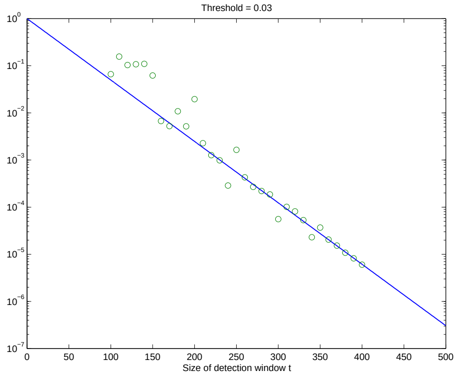

Fig. 3. Exponential decay rate of the type I error probability: node-level Markov chain model.

takes to detect each attack (100-200 time units), which we will refer to as response time. Fig. 3 verifies that the exponent of the type I error probability for both attacks fits closely with the theoretical value η.

5.2 Results for the tree-index model

5.2.1 *Edge-wise case.* Next we turn our attention to the edge-wise tree-indexed Markov model. The nodes of the tree monitor events in their environment and for each observed event they route a packet with the necessary information to the root of the tree. In all examples in this subsection and the following (Sec. 5.2.2) events at each node occur according to independent Poisson processes. Our objective is to detect changes in the event generation rates as described in Sec. 3. The offspring law $p(\cdot)$ is uniform in $\{0, 1, \ldots, 5\}$. We define the "state" of each node depending on the average packet flow per unit time through the node, including packets that

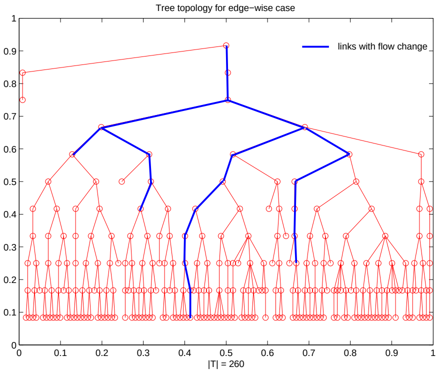

Fig. 4. Tree structure and packet flow change: edge-wise case.

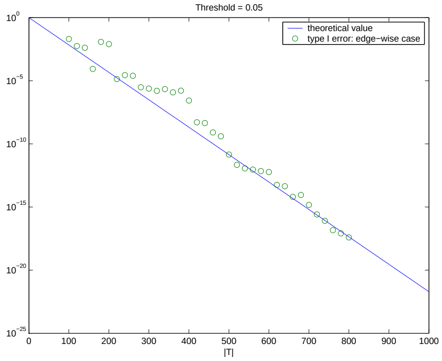

Fig. 5. Exponential decay of error probabilities: edge-wise case.

originate at the node. The average flow is mapped to 10 states. The transition probability matrix $\mathbf{Q}_0$ is selected so that for each a, $\mathbf{Q}_0(\cdot|a)$ is an appropriately truncated triangularly shaped mass function with mode at a and symmetrically diminishing mass as one moves away from a.

Consider the tree of Fig. 4. We selectively changed the event generation rate at a few nodes but this results in packet flow changes in all nodes that are in the corresponding paths to the root. Note that depending on how flow is mapped to states flow changes may not result in state transitions.

The calculation of the empirical measure $\mathbf{L}_X$ is done in a distributed way: each node keeps a vector (with dimension equal to the number of states) of counts of how many downstream nodes, including itself, are in a certain state. When a node changes its state, the corresponding value in this vector is changed and this update is propagated up the tree. We note that distributed computation is useful in implementing anomaly detection techniques of this type in large WSNETs. We applied the detection test of Thm. 3.3, where the anomaly-free laws $p(\cdot)$ and $\mathbf{Q}_0$ can be computed from observations before the anomaly is introduced. Fig. 6 (top) shows the response time of the test.

Type I error probabilities were calculated for tree sizes from 100 to 800. As we can see from Fig. 5, the theoretical exponent fits very well with the observed values for a network (tree) with size larger than 200.

Fig. 7 shows the average detection response time corresponding to different initial percentages of nodes with altered event generation rates. We observe a drastic drop in the response time when this percentage is approximately 15%. The choice of the threshold value η plays an important role in the response time as well. In this figure η is set to 0.05.

5.2.2 *Level-wise case.* Finally, and for the same setting as in Sec. 5.2.1, we consider the level-wise model and apply the detection test of Thm. 4.2. As before, the state of a node is defined depending on the average flow through the node. The offspring law $\mathbf{Q}_0$ was selected so (i) nodes with states corresponding to large flows are more likely to have more children, (ii) nodes with zero flow have no children,

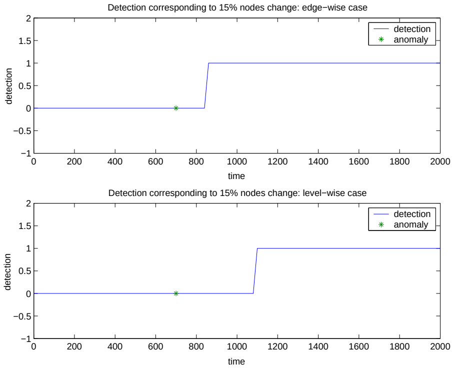

Fig. 6. Response times in the tree models.

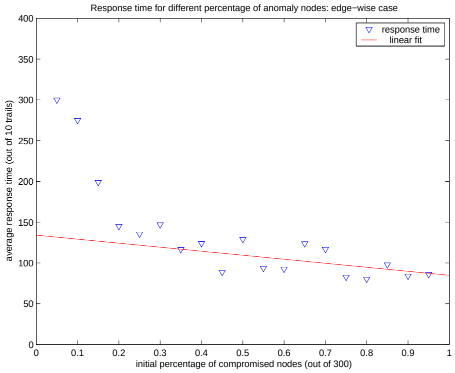

Fig. 7. Average response time as a function of the percentage of nodes with altered event generation rates: edge-wise case.

and (*iii*) the children have higher probability of being in a state close to the parent's state (in terms of average flow).

The detection response time is shown in Fig. 6 (bottom). The response time to detect an anomaly is now larger, due to the increased number of "types" when calculating empirical measures. We expect though the level-wise model to be sensitive to even subtler changes than the edge-wise model, detecting for instance changes that result in deviations in the "type" (cf. Eq. (7)) of children without significant changes in the overall fraction of nodes at a certain state. The type I error exponent reveals similar properties as in the edge-wise case, and is in line with Thm. 4.1.

5.2.3 *Application of the tree-indexed models to monitoring water temperature.* In this section we apply the edge-wise tree model to a simulated environment where the sensors are deployed to monitor water temperature in a pond area. Following the model of Sec. 3.1, a tree is generated with a predetermined probability distribution. The environment we are simulating has a depth of about 30 meters and the water temperature changes following a linear fluid model with gradient taken from Fig. 8. A random perturbation is added to the model to simulate temperature changes during the day. We use average data from the Walden Pond in Concord, Massachusetts (Friesz and Colman [2001]) as reference, and the simulated "rough surface" of underwater temperatures is shown in Fig. 9. A total of 600–1600 sensor nodes are deployed in an area of size 200m × 200m, and the size of the type set is 15. To generate an anomaly event, we randomly initiate type changes in about 12% of the sensor nodes.

To prove the effectiveness of our method, we compare it with the following average rule: nodes are divided into several groups (for example, 3 groups) based on their depth; average measurements (temperature data) are collected among the nodes in the same group and compared to a threshold (which is set based on the long-run anomaly-free average); if in any of these groups the deviation from the threshold is more than 15% then an anomaly is declared. We compare our proposed method (LD rule for short) with such average rules that use 3 groups, 6 groups, 10 groups, and 20 groups of sensors.

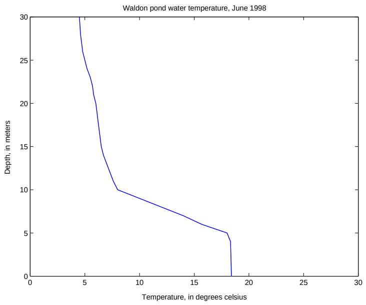

Fig. 8. Statistical data from the Walden Pond, June 1998.

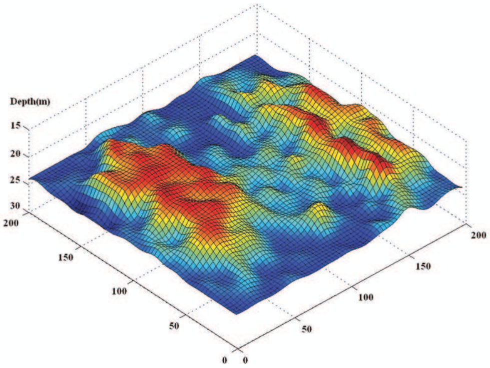

Fig. 9. Simulated rough surface of underwater temperatures.

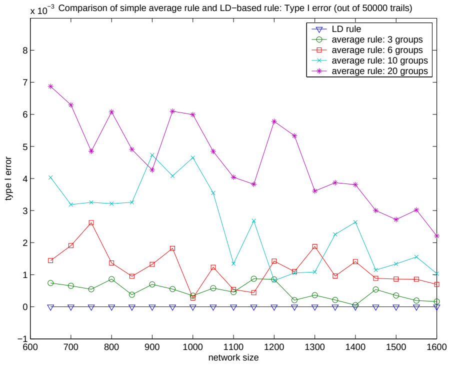

Fig. 10. Type I error - linear scale.

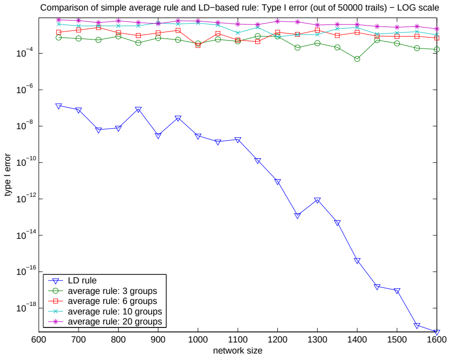

Fig. 11. Type I error - log scale.

Figs. 10, 11, 12, and 13 compare the type I and II errors in linear and log scales with the various average rules. Clearly the LD rule outperforms all average rules, with the performance gap increasing as the network size grows. In addition, Fig. 11 shows, as expected, an exponential decay of the type I error with the preset threshold η. The ROC curves shown in Fig. 14 were drawn for a network of size equal to 1200 nodes and more convincingly demonstrate that the LD rule outperformed all average rules, i.e., the LD rule has the smallest type II error probability for any fixed type I error probability.

6. CONCLUSIONS

We considered the problem of anomaly detection in wireless sensor networks. The framework we introduced is general enough to be applied to detecting statistically significant temporal or spatial changes in the environment the sensor network is monitoring but also in the typical packet routing patterns in the network. The latter type of disruptions may indicate naturally occurring phenomena (changes in wireless connectivity, e.g., due to new structures introduced to the environment) or

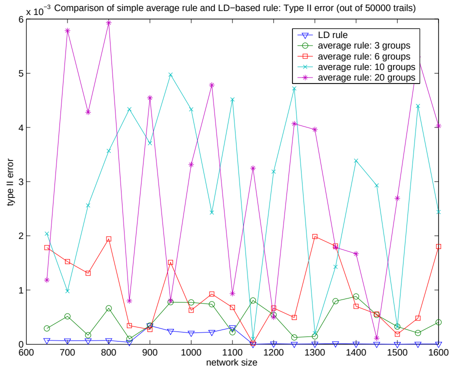

Fig. 12. Type II error - linear scale.

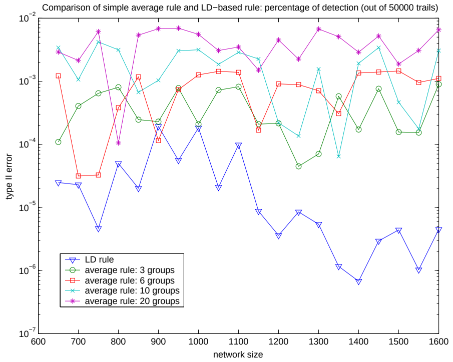

Fig. 13. Type II error - log scale.

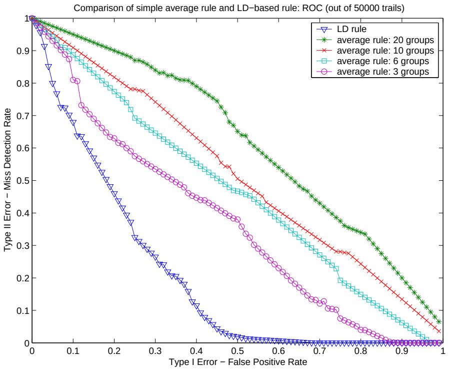

Fig. 14. Comparison of ROC curves.

adversarial attacks.

Our work proposed the use of a series of Markov models to characterize normal behavior. We considered, and analyzed, three (increasingly more detailed) models. Our first model captured temporal changes in the state of the network – defined as a vector of individual node states. The two remaining models were based on tree-indexed Markov chains and captured tree WSNET topologies. These latter models can detect spatial and temporal anomalies by defining the state of the tree nodes as a function of measurements over a certain time-window (e.g., an average).

In each case we developed a rigorous anomaly detection test based on large deviations results for the corresponding model. We showed that these anomaly detection tests are optimal in a decision-theoretic sense (asymptotic Neyman-Pearson optimality). Illustrative numerical results demonstrate that the techniques we developed can detect – within a reasonable amount of time – a broad variety of attacks, changes in the underlying process being monitored, and network failures.

REFERENCES

Chan, H., Perrig, A., and Song, D. 2003. Random key predistribution schemes for sensor networks. *IEEE Symposium on Security and Privacy*, 197–213.

Dembo, A., Mörters, P., and Sheffield, S. 2005. Large deviations of Markov chains indexed by random trees. *Ann. I. H. Poincaré PR-41*, 971–996.

Dembo, A. and Zeitouni, O. 1998. *Large Deviations Techniques and Applications*, 2nd ed. Springer-Verlag, NY.

Friesz, P. and Colman, J. 2001. Hydrology and trophic ecology of Walden Pond, Concord, Massachusetts: US Geological survey water-resources investigations report 01-4153. Tech. rep., US Geological Survey.

Hoeffding, W. 1965. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist. 36*, 369–401.

Hu, Y. C., Perrig, A., and Johnson, D. 2003. Packet leashes: a defense against wormhole attacks in wireless networks. In *INFOCOM*.

Lazos, L., Poovendran, R., Meadows, C., Syverson, P., and Chang, L. W. 2005. Preventing wormhole attacks on wireless ad hoc networks: A graph theoretic approach. In *Proceedings of the IEEE Wireless Communications and Networking Conference*. 1193–1199.

Paschalidis, I. and Smaragdakis, S. 2008. Spatio-temporal network anomaly detection by assessing deviations of empirical measures. *IEEE/ACM Trans. on Networking*.

Paschalidis, I. C. and Chen, Y. 2008. Anomaly detection in sensor networks based on large deviations of Markov chain models. In *Proceedings of the 47th IEEE Conference on Decision and Control*. Cancun, Mexico, 2338–2343.

Paschalidis, I. C. and Guo, D. 2010. Robust and distributed stochastic localization in sensor networks: Theory and experimental results. *ACM Trans. Sensor Networks*. in print.

Perrig, A., Stankovic, J., and Wagner, D. 2004. Security in wireless sensor networks. *Communications of the ACM 47,* 6, 53–57.

Perrig, A., Szewczyk, R., Wen, V., Culler, D., and Tygar, J. D. 2002. Spins: Security protocols for sensor networks. *Wireless Networks* 5, 521–534.

Ray, S., Lai, W., and Paschalidis, I. C. 2006. Statistical location detection with sensor networks. *Joint special issue IEEE/ACM Trans. Networking and IEEE Trans. Information Theory 52,* 6, 2670–2683.

Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D. 2006. Online outlier detection in sensor data using non-parametric models. In *Proceedings of the International Conference on Very Large Data Bases*. Seoul, Korea, 187–198.

Zeitouni, O., Ziv, J., and Merhav, N. 1992. When is the generalized likelihood ratio test optimal? *IEEE Trans. Inform. Theory 38,* 5, 1597–1602.

Zhu, S., Setia, S., and Jajodia, S. 2006. Leap+: Efficient security mechanisms for large-scale distributed sensor networks. *ACM Transactions on Sensor Networks 2,* 4, 500–528.