Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Anomaly Detection in Sensor Networks based on Large Deviations of Markov Chain Models*

Ioannis Ch. Paschalidis† Yin Chen‡

Abstract— **We introduce an anomaly detection framework for wireless sensor networks able to detect statistically significant temporal or spatial changes in either the underlying process the sensor network is monitoring or the network operation itself. We consider a series of Markov models to characterize the behavior of the sensor network, including tree-indexed Markov chains which can model its spatial structure. Large deviations techniques are used to compare the distribution of the Markov model estimated from past anomaly-free traces with its most recent empirical measure. We develop optimal decision rules for each corresponding Markov model to identify anomalies in recent activity. Simulation results validate the effectiveness of the proposed anomaly detection algorithms.**

I. Introduction

Wireless Sensor NETworks (WSNETs) are networks of small devices that communicate wirelessly and are used to monitor (and control) a physical system. WSNET nodes can be powered by batteries and have limited processing and storage abilities. Coupled with appropriate software (protocols) they can self-configure to form a network which collects data of interest but also can push commands to specific nodes that may control actuators. Emerging applications include industrial and building automation, homeland defense, asset and personnel tracking [1] and patient monitoring.

In monitoring applications one interesting question that arises is that of detecting abnormalities (or anomalies) in the various quantities that are being measured. Often, simple threshold rules (e.g., a variable exceeding a certain value) may suffice to that end. There exist however cases where deviations from normal behavior are subtle and result in changes in the spatial or temporal distribution of measurements. Such changes are much harder to detect, yet their detection is important as they may be precursors (either in time of space) of "bigger" abnormalities. Homeland defense (e.g., air quality monitoring) applications offer examples of this type of anomalies.

The security of the network itself is also an important consideration. Due the the use of wireless communications, and the power constrained limited capabilities of the nodes which precludes the use of sophisticated security measures, WSNETs are extremely vulnerable to a wide range of adversarial attacks, exploits, viruses, and other information security vulnerabilities [2]. Following a computer systems approach, most of the literature has only considered protocol design as a way to address these security concerns.

Our work considers anomaly detection in general enough terms that can accommodate both the monitoring and network security related problems outlined above. 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 technical tool we use is large deviations for Markov chains developed in [3], [4]. Large deviations theory provides a powerful way of handling rare events and their associated probabilities. 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. If the empirical measure takes very unlikely values this points to a statistical anomaly. Borrowing from hypothesis testing techniques, for each of the Markov models we consider we develop appropriate anomaly detection tests and establish that they are optimal in a generalized Neyman-Pearson sense. Related techniques have also been applied in [5] to detect anomalies in Internet traffic. We refer the interested reader to [5] for an extensive literature review on anomaly detection.

The rest of this paper is organized as follows. In Sec. II, we consider a simple Markov chain model of a multi-hop sensor network and develop the requisite anomaly detection test. In Sec. III and IV, we adopt a tree-indexed Markov model, survey large deviations results for its empirical measure, and develop the corresponding anomaly detection test. We present simulation results for each model in Sec. V, and draw conclusions in Sec. VI.

II. A simple Markov chain model

We start with a simple 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 via other nodes acting as relays.

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. We assume that only a single node can be in state 1 at any given point in time.

* Research partially supported by the NSF under grants DMI-0330171, ECS-0426453, CNS-0435312, EFRI-0735974, and by the DOE under grant DE-FG52-06NA27490.

† Corresponding author. Center for Information & Systems Eng., Dept. of Electrical and Computer Eng., and Division of Systems Eng., Boston University, 15 St. Mary's St., Brookline, MA 02446, e-mail: yannisp@bu.edu, url: http://ionia.bu.edu/.

‡ Center for Information & Systems Eng., Boston University, e-mail: yinchen@bu.edu.

To motivate the above setting, consider a WSNET tracking an object as it moves through the coverage area of the network. One situation that can be modeled in this way is the routing of packets through the network. Assume that the network handles very light traffic, so that only a single packet is present at any given point in time. Any node that receives the packet switches its state from 0 to 1 and switches back to 0 once it transmits the packet to another node. In applications where a WSNET is used to monitor events of interest (temperature, radioactivity, etc.), time between two consecutive events (measurements) is in general much longer than the time required for the relevant information to be received by the gateway.

Let now the state of the WSNET be the index of the node which is in state 1 and assume that state transitions satisfy the Markov property. We use $q_0(j|i)$ to denote the transition probability from WSNET state i to j and denote by $\mathbf{Q}_0$ the corresponding $n \times n$ transition probability matrix, namely, $\mathbf{Q}_0 = (q_0(j|i))_{i,j=1}^n$. Assume that $\mathbf{Q}_0$ is irreducible and aperiodic with a unique stationary distribution $\boldsymbol{\pi}_0 = (\pi_1^0, \ldots, \pi_n^0)$, where π_i^0 is equal to the steady-state fraction of time the Markov chain is in state 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 $\mathbf{Y}_t = (Y_1, Y_2, \ldots, Y_t)$ 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).

Letting $\Sigma = \{1, 2, \ldots, n\}$, 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 $\mathbf{Y}_t = (y_1, \ldots, y_t)$ 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 $\mathbf{Q}_i$. To decide whether a hypothesis test is optimal, the following generalized Neyman-Pearson criterion for finite alphabets was suggested by Hoeffding [6].

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 Y_i are i.i.d., Hoeffding [6] 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. [7] 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 $\mu_t^{\mathbf{Y}}(i, j)$ as the empirical joint 2-step occurrence of the system states

$$\mu_t^{\mathbf{Y}}(i, j) = \frac{1}{t} \sum_{k=1}^{t} 1(Y_{k-1} = i, Y_k = j), \quad i, j = 1, \ldots, n,$$

where $1\{\cdot\}$ denotes the indicator function. $\mu_t^{\mathbf{Y}}(i, j)$ can be interpreted as the fraction of time the system makes transitions from state i to state j. We will write $\boldsymbol{\mu}_t^{\mathbf{Y}}$ for the vector of all $\mu_t^{\mathbf{Y}}(i, j)$ and we will use the same convention of denoting vectors with bold letters for the rest of the paper. The marginals of $\boldsymbol{\mu}_t^{\mathbf{Y}}$ are denoted by the vectors $\boldsymbol{\mu}_{L,t}^{\mathbf{Y}}$ and $\boldsymbol{\mu}_{R,t}^{\mathbf{Y}}$ with elements

$$\mu_{L,t}^{\mathbf{Y}}(i) = \sum_{j=1}^{n} \mu_t^{\mathbf{Y}}(i, j), \quad \mu_{R,t}^{\mathbf{Y}}(i) = \sum_{j=1}^{n} \mu_t^{\mathbf{Y}}(j, i).$$

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

$$\mu_t^{\mathbf{Y}}(j|i) = \frac{\mu_t^{\mathbf{Y}}(i, j)}{\mu_{L,t}^{\mathbf{Y}}(i)},$$

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

Define $\boldsymbol{\mu}_{L,t}^{\mathbf{Y}} \otimes \mathbf{Q}_0$ as the vector with elements $\mu_{L,t}^{\mathbf{Y}}(i) q_0(j|i)$, $i, j = 1, \ldots, n$, and consider the divergence (relative entropy) between $\boldsymbol{\mu}_t^{\mathbf{Y}}$ and $\boldsymbol{\mu}_{L,t}^{\mathbf{Y}} \otimes \mathbf{Q}_0$:

$$\begin{aligned} D(\boldsymbol{\mu}_t^{\mathbf{Y}} || \boldsymbol{\mu}_{L,t}^{\mathbf{Y}} \otimes \mathbf{Q}_0) &= \sum_{i,j=1}^{n} \mu_t^{\mathbf{Y}}(i, j) \log \frac{\mu_t^{\mathbf{Y}}(i, j)}{\mu_{L,t}^{\mathbf{Y}}(i) q_0(j|i)} \\ &= \sum_{i=1}^{n} \mu_{L,t}^{\mathbf{Y}}(i) D(\boldsymbol{\mu}_t^{\mathbf{Y}}(\cdot|i) || \mathbf{q}_0(\cdot|i)). \end{aligned} \tag{2}$$

Zeitouni et al. [7] establish the following result.

Theorem II.1 ([7]) *The decision test*

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

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

This theorem provides an optimal anomaly detection test for the simple 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.

III. A TREE-INDEXED MARKOV CHAIN MODEL: EDGE-WISE

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 $[0, t]$. 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 distributively done following the leaf-root path and propagating the lower-level information recursively to the gateway.

As in Section II, 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 [4]. We will consider trees that are conditioned to have exactly n nodes and we will take the large deviations limit as $n \to \infty$.

A. 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 IV 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

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

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. [4] 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

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

We also 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 [3]) that $I_p(\cdot)$ is the large deviations rate function associated with the law $p(\cdot)$. Finally, as in Section II, define $\boldsymbol{\mu}_1 \otimes \mathbf{Q}_0$ as the vector with elements $\mu_1(a) q_0(b|a)$, $a, b = 1, \ldots, m$, and let $<<$ denote pointwise strict inequality between vectors.

Theorem III.1 ([4]) *Suppose that T is a tree with offspring law $p(\cdot)$ such that $0 < p(0) < 1 - p(1)$, $\sum_l l p(l) = 1$ 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*

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

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. We have omitted the proof for brevity.

Lemma III.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* $p_\theta(l) = p(l)/(\sum_j p(j)e^{\theta j})$ *for any value* $\theta \in \mathbb{R}$.

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 III.2. Note that with $0 < p(0) < 1-p(1)$ there exists a unique θ^* such that $\sum_l lp_{\theta^*}(l) = 1$. Hence, Theorem III.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 $I_{p_{\theta^*}}(\cdot)$.

B. 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. II 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 II the following theorem provides the test and establishes its optimality; the proof is omitted in the interest of space.

Theorem III.3 *The decision test* $\mathscr{S}_2^{*,n}(X)$

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

is optimal according to the generalized Neyman-Pearson criterion.

IV. Tree-indexed Markov chain model: level-wise

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.

A. 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

$$C(v) = (N(v), X_1(v), \ldots, X_{N(v)}(v)) \in \mathscr{X}^* \tag{5}$$

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

$$M_X(a,c) = \frac{1}{|\mathscr{V}|}\sum_{v \in \mathscr{V}} 1\{X(v) = a, C(v) = c\}. \tag{6}$$

Dembo et al. [4] establish a large deviations result for $\mathbf{M}_X$ for trees that are conditioned to have n nodes. To state the result we need to 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. III 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 [4] states the large deviations result for the empirical measure $\mathbf{M}_X$.

Theorem IV.1 ([4]) *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*

$$J(\boldsymbol{\sigma}) = \begin{cases} D(\boldsymbol{\sigma}|\boldsymbol{\sigma}_1 \otimes \mathbf{Q}_0), & \text{if } \boldsymbol{\sigma} \text{ is shift-invariant,} \\ \infty, & \text{otherwise.} \end{cases} \quad (7)$$

B. 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. III-B, 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 (6), where the superscript n indicates that the tree has n nodes. Using similar terminology and notation as in Sec. III-B the following theorem provides the test and establishes its optimality. The proof is omitted due to space limitations.

Theorem IV.2 *The decision test $\mathscr{S}_3^{*,n}(X)$*

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

is optimal according to the generalized Neyman-Pearson criterion.

V. Numerical results

In this section we present a host of numerical results that validate the anomaly detection tests we developed.

A. Results for the simple Markov model

Figures 1 considers the following disruption in the network routing we wish to detect: a compromised node has changed its parent due to a jammed link. As we described in Sec. II we model the routing by a simple Markov chain whose state is the index of the node possessing the data packet, assuming that we operate in the light traffic regime where only a single node has a packet at any given point in time. In this scenario we considered packets are generated at each node according to independent Poisson processes. Once a packet is generated it is routed to the root. The Poisson rates are small enough to ensure a single packet in the network at any point in time (during our simulation).

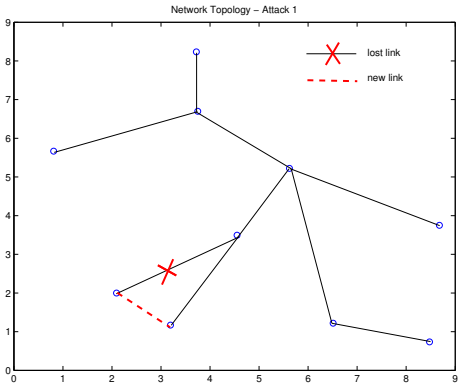

Fig. 1. A compromised node switches its parent.

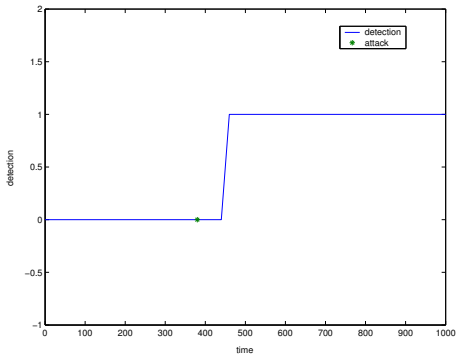

Fig. 2. Detection of attack in the simple Markov chain model.

We use a long trace of observations before the attacks to estimate the anomaly-free law $\mathbf{Q}_0$ and then apply the test of Thm. II.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 $e^{-t\eta} = 0.05$. Figure 2 shows the time it takes to detect each attack (about 100 time units), which we will refer to as response time. Simulation results verified that the exponent of the type I error probability fits closely with theoretical value η.

B. Results for the tree model: edge-wise case

Next we consider 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. V-C) 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. III. The offspring law $p(\cdot)$ is uniform in $\{0, 1, \ldots, 5\}$. We defined the "state" of each node depending on the average packet flow per unit time through the node, including packets that originate at the node. The average flow was mapped to 10 states. The transition probability matrix $\mathbf{Q}_0$ was selected so that for each a, $\mathbf{Q}_0(\cdot|a)$ is an appropriately truncated triangularly shaped

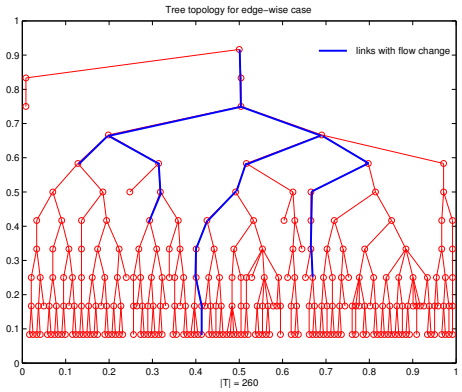

Fig. 3. Tree structure and packet flow change: edge-wise case.

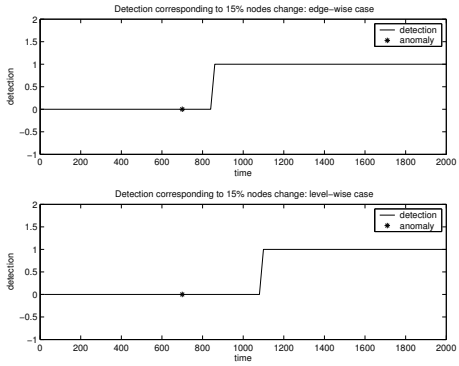

Fig. 4. Response times in the tree models.

mass function with mode at a and symmetrically diminishing mass as one moves away from a.

Consider the tree of Fig. 3. 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. III.3, where the anomaly-free laws $p(\cdot)$ and $\mathbf{Q}_0$ can be computed from observations before the anomaly is introduced. Type I error probabilities were calculated for tree sizes from 100 to 800. And the theoretical exponent fits very well with the observed values for a network (tree) with size larger than 200.

With η set to 0.05, we performed simulations corresponding to different initial percentages of nodes with altered event generation rates, and observed a drastic drop in the response time when this percentage is approximately 15%, which reasonably demonstrates our method is able to detect anomalies within a short time.

C. Results for the tree model: level-wise case

Finally, and for the same setting as in Sec. V-B, we consider the level-wise model and apply the detection test of Thm. IV.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, and (iii) the children have higher probability of being in a state close to the state of the parent (in terms of average flow).

The detection response time is shown in Figure 4. Comparing to the edge-wise case, 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. 5) of children without significant changes in the overall fraction of nodes at a certain state. The exponent of Type I error probabilities presents similar properties as the edge-wise case and is in line with Thm. IV.1.

VI. Conclusions

We considered the problem of anomaly detection in wireless sensor networks in a general enough framework to be able to detect statistically significant temporal or spatial changes in either the underlying process the sensor network is monitoring or the typical packet routing patterns in the network. The latter type of disruptions may indicate naturally occurring phenomena (changes in wireless connectivity) or adversarial attacks.

We proposed the use of Markov models to characterize normal behavior and analyzed three (increasingly more detailed) models. In each case we developed a rigorous anomaly detection test based on large deviations results for the corresponding model. 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

[1] I. C. Paschalidis and D. Guo, "Robust and distributed localization in sensor networks," in *Proceedings of the 46th IEEE Conference on Decision and Control*, New Orleans, Louisiana, December 2007, pp. 933–938.

[2] A. Perrig, J. Stankovic, and D. Wagner, "Security in wireless sensor networks," *Communications of the ACM*, vol. 47, no. 6, pp. 53–57, 2004.

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

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

[5] I. C. Paschalidis and G. Smaragdakis, "Spatio-temporal network anomaly detection by assessing deviations of empirical measures," *IEEE/ACM Trans. Networking*, in print.

[6] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–401, 1965.

[7] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, 1992.