This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

Statistical Anomaly Detection via Composite Hypothesis Testing for Markov Models*

Jing Zhang†, *Student Member, IEEE*, and Ioannis Ch. Paschalidis‡, *Fellow, IEEE*

***Abstract*—Under Markovian assumptions, we leverage a Central Limit Theorem (CLT) for the empirical measure in the test statistic of the composite hypothesis Hoeffding test so as to establish weak convergence results for the test statistic, and, thereby, derive a new estimator for the threshold needed by the test. We first show the advantages of our estimator over an existing estimator by conducting extensive numerical experiments. We find that our estimator controls better for false alarms while maintaining satisfactory detection probabilities. We then apply the Hoeffding test with our threshold estimator to detect anomalies in two distinct applications domains: one in communication networks and the other in transportation networks. The former application seeks to enhance cyber security and the latter aims at building smarter transportation systems in cities.**

***Index Terms*—Hoeffding test, weak convergence, false alarm rate, Markov chains, network anomaly detection, cyber security, non-typical traffic jams, smart cities.**

I. Introduction

For a given system, *Statistical Anomaly Detection (SAD)* involves learning from data the normal behavior of the system and identifying/reporting time instances corresponding to atypical system behavior. SAD has vast applications. For instance, motivated by the importance of enhancing cyber security, recent literature has seen applications in communication networks; see, e.g., [1], [2], [3], [4]. The behavior of the system is typically represented as a time series of real vectors and, in its most general version, anomaly detection is done through some *Composite Hypothesis Test (CHT)*.

Specifically, a CHT aims to test the hypothesis that a given sequence of observations is drawn from a known *Probability Law (PL)* (i.e., *probability distribution*) defined on a finite alphabet [5]. Among numerous such tests, the one proposed by Hoeffding [6] has been well known for decades. When implementing the Hoeffding test in the context of SAD, one must appropriately set a threshold η so as to ensure a low false alarm rate while maintaining a reasonably high detection rate. In the existing literature, this threshold is typically estimated by using Sanov's theorem [7] – a large deviations result. Note that such an estimator (let us denote it by η^{sv}) is valid only in the asymptotic sense. In practice, however, only a finite number of observations are available, and it can be observed in simulations that η^{sv} is not accurate enough, especially for relatively small sample sizes.

Our contributions in this paper include:

1) Under Markovian assumptions, we leverage a *Central Limit Theorem (CLT)* for a selected empirical measure related to the test statistic of the Hoeffding test, so as to establish weak convergence results for the test statistic, and derive a threshold estimator η^{wc} therefrom, thus, extending the work of [5] which tackles the problem under independent and identically distributed (i.i.d.) assumptions.
2) We propose algorithms to calculate the threshold estimator η^{wc} obtained above for the ordinary and a robust version of the Hoeffding test, respectively. We assess the advantages of our estimator over earlier work through numerical experiments.
3) We apply the Hoeffding test with our threshold estimator to two types of systems for the purpose of anomaly detection: (i) a communication network with flow data simulated by the software package SADIT [8]; and (ii) a real transportation network with traffic jam data reported by Waze, a smartphone GPS navigation application. To the best of our knowledge, the latter is a novel application of anomaly detection.

A preliminary conference version of this work appeared in [9]. The present paper includes detailed technical arguments, derives results for the robust version of the Hoeffding test, expands the numerical comparisons with earlier work, and develops the traffic jam anomaly detection application.

The rest of this paper is organized as follows. In Section II we review related work. We formulate the threshold estimation problem in Section III and derive theoretical results in Section IV. Section V contains experimental results. Concluding remarks are in Section VI and a number of proofs appear in the Appendix.

Notational conventions: All vectors are column vectors. For economy of space, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is its dimension. We use prime to denote the transpose of a matrix or vector. Denote by $\mathbb{N}_+$ the set of all nonnegative integers. $\|\mathbf{x}\|$ denotes the ℓ_2-norm of a vector $\mathbf{x}$, $\lfloor x \rfloor$ the integer part of a positive number x, $|\mathcal{A}|$ the cardinality of a set $\mathcal{A}$, $\log$ the natural logarithm, $\mathbb{P}(A)$ the probability of an event A, $\mathbb{E}[X]$ the expectation of a random variable X, and $\text{Cov}(X_1, X_2)$ the covariance between two random variables X_1 and X_2. We use $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ to denote a Gaussian distribution with zero mean

* Research partially supported by the NSF under grants CNS-1645681, CCF-1527292, and IIS-1237022, by the ARO under grants W911NF-11-1-0227 and W911NF-12-1-0390, and by a grant from the Boston Area Research Initiative (BARI).

† Division of Systems Engineering, Boston University, Boston, MA 02446, email: jzh@bu.edu.

‡ Division of Systems Engineering, Dept. of Electrical and Computer Engineering, and Dept. of Biomedical Engineering, Boston University. 8 St. Mary's St., Boston, MA 02215, email: yannisp@bu.edu, url: http://sites.bu.edu/paschalidis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

and covariance matrix $\mathbf{\Sigma}$. $X_1 \simeq X_2$ indicates that the two random variables X_1 and X_2 have approximately the same distribution. $\mathbb{1}\{\cdot\}$ denotes the indicator function and $\xrightarrow[n\to\infty]{d}$ (resp., $\xrightarrow[n\to\infty]{w.p.1}$) denotes convergence *in distribution* (resp., *with probability one*) as n approaches infinity.

II. Related Work

Modeling network traffic as stationary in time, [1] applies two methods: one assumes the traffic to be an i.i.d. sequence and the other assumes observations of system activity follow a finite-state Markov chain. Both methods are extended in [4] to the case where system activity is time-varying. When implementing the Hoeffding test, however, both [1] and [4] use the large deviations estimator η^{sv} to calculate the detection threshold in a finite sample-size setting, thus not being able to control the false alarm rate well enough.

To derive a more accurate threshold estimator, [5], [10] use a procedure commonly used by statisticians: deriving results based on *Weak Convergence (WC)* of the test statistic in order to approximate the error probabilities of the Hoeffding test. Under i.i.d. assumptions, [5] (see also [10], [11]) proposes an alternative estimator for η (let us denote it by η^{wc}), which is typically more accurate than η^{sv}, especially when not that many samples are available.

There has also been work on obtaining a tighter approximation of η by refining Sanov's theorem [12]. However, such refinements of large deviation results are typically faced with computational difficulty; for instance, as noted in [10], using the results of [12] requires the computation of a surface integral.

Several alternative anomaly detection approaches have been proposed, using for instance change detection methods [13]. We refer the reader for a comprehensive review of alternative methods to [13] and [1].

III. Problem Formulation

To model the statistical properties of a general system, we introduce a few more notational conventions and some definitions. Let $\mathbf{\Xi} = \{\xi_i;\ i = 1, \ldots, N\}$ be a finite alphabet containing N symbols $\xi_1, \ldots, \xi_N$, and $\mathbf{Y} = \{Y_l;\ l = 0, 1, 2, \ldots\}$ a time series of observations. Define the *null hypothesis* $\mathcal{H}$ as: $\mathbf{Y}$ is drawn according to a Markov chain with state set $\mathbf{\Xi}$ and transition matrix $\mathbf{Q} = [q_{ij}]_{i,\,j=1}^{N}$. To further characterize the stochastic properties of $\mathbf{Y}$, we define the *empirical Probability Law (PL)* by

$$\Gamma_n(\theta_{ij}) = \frac{1}{n}\sum_{l=1}^{n} \mathbb{1}\{Z_l = \theta_{ij}\}, \tag{1}$$

where $Z_l = (Y_{l-1},\ Y_l),\ l = 1, \ldots, n$, $\theta_{ij} = (\xi_i,\ \xi_j) \in \mathbf{\Xi} \times \mathbf{\Xi}$, $i,\ j = 1, \ldots, N$. Denote the transformed alphabet $\mathbf{\Theta} = \{\theta_{ij};\ i,\ j = 1, \ldots, N\} = \{\tilde{\theta}_k;\ k = 1, \ldots, N^2\}$ and note $\mathbf{\Theta} = \mathbf{\Xi} \times \mathbf{\Xi}$ with $\tilde{\theta}_1 = \theta_{11}, \ldots, \tilde{\theta}_N = \theta_{1N}, \ldots, \tilde{\theta}_{(N-1)N+1} = \theta_{N1}, \ldots, \tilde{\theta}_{N^2} = \theta_{NN}$. Let also the set of PLs on $\mathbf{\Theta}$ be $\mathcal{P}(\mathbf{\Theta})$.

The transformed observations $\mathbf{Z} = \{Z_l;\ l = 1, 2, \ldots\}$ form a Markov chain evolving on $\mathbf{\Theta}$; denote its transition matrix by $\mathbf{P} = [p_{ij}]_{i,\,j=1}^{N^2}$ and the stationary distribution by

$$\boldsymbol{\pi} = (\pi_{ij};\ i,\, j = 1, \ldots, N) = (\tilde{\pi}_k;\ k = 1, \ldots, N^2), \tag{2}$$

where π_{ij} denotes the probability of seeing θ_{ij}, and $\tilde{\pi}_1 = \pi_{11}, \ldots, \tilde{\pi}_N = \pi_{1N}, \ldots, \tilde{\pi}_{(N-1)N+1} = \pi_{N1}, \ldots, \tilde{\pi}_{N^2} = \pi_{NN}$. We have [7]

$$p(\theta_{ij}\,|\theta_{kl}\,) = \mathbb{1}\{i = l\}q_{ij}, \quad k,\, l,\, i,\, j = 1, \ldots, N, \tag{3}$$

which enables us to obtain $\mathbf{P}$ directly from $\mathbf{Q}$; see Remark 2 for an example. We can now restate the *null hypothesis* $\mathcal{H}$ as: the Markov chain $\mathbf{Z} = \{Z_l;\ l = 1, 2, \ldots\}$ is drawn from PL $\boldsymbol{\pi}$.

To quantify the distance between the empirical PL $\mathbf{\Gamma}_n$ and the actual PL $\boldsymbol{\pi}$, one considers the *relative entropy* (or *divergence*) between $\mathbf{\Gamma}_n$ and $\boldsymbol{\pi}$:

$$D(\mathbf{\Gamma}_n \| \boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Gamma_n(\theta_{ij}) \log \frac{\Gamma_n(\theta_{ij}) / \big(\sum_{t=1}^{N} \Gamma_n(\theta_{it})\big)}{\pi_{ij} / \big(\sum_{t=1}^{N} \pi_{it}\big)}, \tag{4}$$

and the *empirical measure*:

$$\mathbf{U}_n = \sqrt{n}(\mathbf{\Gamma}_n - \boldsymbol{\pi}), \tag{5}$$

where $\boldsymbol{\pi}$ is defined in (2) and $\mathbf{\Gamma}_n$ is the vector

$$\mathbf{\Gamma}_n = (\Gamma_n(\theta_{11}), \ldots, \Gamma_n(\theta_{1N}), \ldots, \Gamma_n(\theta_{N1}), \ldots, \Gamma_n(\theta_{NN})).$$

Let now $\mathcal{H}_n$ be the output of a test that decides to accept or to reject the *null hypothesis* $\mathcal{H}$ based on the first n observations in the sequence $\mathbf{Z}$. Under Markovian assumptions (Assumption 1 in Section IV), the Hoeffding test [7] is given by

$$\mathcal{H}_n \text{ rejects } \mathcal{H} \text{ if and only if } D(\mathbf{\Gamma}_n \| \boldsymbol{\pi}) > \eta, \tag{6}$$

where $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$ (cf. (4)) is the *test statistic* and η is a *threshold*.

It is known that the Hoeffding test (6) satisfies asymptotic Newman-Pearson optimality [1], [4], in the sense that it maximizes the exponential decay rate of the *misdetection probability* over all tests with a *false positive probability* with exponential decay rate larger than η. Thus, an appropriate threshold η should enable the test to have a small false positive rate while maintaining a satisfactorily high detection rate.

The theoretical *false positive rate* [5] of the test (6) is given by

$$\beta = \mathbb{P}_{\mathcal{H}}(D(\mathbf{\Gamma}_n \| \boldsymbol{\pi}) > \eta), \tag{7}$$

where the subscript $\mathcal{H}$ indicates that the probability is taken under the null hypothesis.

Given a tolerable (target) β, by conducting an ROC (Receiver Operating Characteristic) analysis for the Hoeffding test using labeled training data, we could "tune" η such that the corresponding *discrete test*[1] [14] has a small false alarm rate and a high detection rate. In particular, we could select an η corresponding to a point close to the northwest corner of the ROC graph. However, such tuning is too expensive and depends heavily on the quality and quantity of the training data. We can also, in principle, obtain the corresponding η

[1] A *discrete test* corresponds to a fixed value for η in (6).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

in (7) by directly simulating the samples of the test statistic $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$, thus deriving an empirical Cumulative Distribution Function (CDF) and using its $(1-\beta)$-quantile. However, we will note in Remark 5 that this is also computationally too expensive when applied through a so-called "windowing" technique for purposes of anomaly detection. Thus, we seek to estimate η without directly simulating the statistic. To that end, existing work uses Sanov's theorem [7] to derive an estimator for η. Specifically, for large enough n, by replacing the right hand side in (7) with an exponential we can obtain a minimal η that suffices to bring the false positive rate below β [1], [4]. Such an η is given by

$$\eta_{n,\beta}^{\text{sv}} \approx -(1/n)\log(\beta), \tag{8}$$

where we use the n, β subscript to denote the dependence of this estimator on β and n and the label sv indicates that it is obtained from Sanov's theorem. We note that the estimator (8) does not contain any direct distributional information of the statistic $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$; this might be one of the causes leading to inaccurate estimation of $\eta_{n,\beta}$, especially when the sample size n is relatively small in practice. To see this more clearly, one can consider an extreme scenario where $N = 4$, $\beta = 10^{-1000}$, and $n = 50$ (this is a reasonably small value; comparable to $N^2 = 16$). Then by (8), $\eta_{n,\beta}^{\text{sv}}$ would be way larger than necessary, tending to yield a test with zero false alarm rate but also zero detection rate for a typical test set. The issue arises because we use an asymptotic large deviations result for a relatively modest value of n. Our primary goal in this paper is to derive an alternative threshold estimator, which would hopefully be more accurate than $\eta_{n,\beta}^{\text{sv}}$ for modest values of n, in terms of a certain metric that we will introduce in Section V.

IV. Theoretical Results

We introduce the following assumption.

Assumption 1 *$\mathbf{Z} = \{Z_l;\ l = 1, 2, \ldots\}$ is an aperiodic, irreducible, and positive recurrent Markov chain ([15]) evolving on $\mathbf{\Theta}$ with transition matrix $\mathbf{P}$, stationary distribution $\boldsymbol{\pi}$, and with the same $\boldsymbol{\pi}$ as its initial distribution.*

Remark 1 Since $\mathbf{\Theta}$ is a finite set, $\mathbf{Z}$ is uniformly ergodic [15] under Assumption 1. Assuming $\boldsymbol{\pi}$ as the initial distribution is done for notational simplicity; our results apply for any feasible initial distribution. Note also that, under Assumption 1, $\boldsymbol{\pi}$ must have full support over $\mathbf{\Theta}$; i.e., each entry in $\boldsymbol{\pi}$ is strictly positive.

Lemma 1 *Suppose Assumption 1 holds. Then*

$$\frac{\pi_{ij}}{\sum_{t=1}^{N}\pi_{it}} = \frac{\pi_{ij}}{\sum_{t=1}^{N}\pi_{ti}} = q_{ij}, \quad i, j = 1, \ldots, N. \tag{9}$$

Proof: See Appendix A. ∎

Remark 2 Under Assumption 1, Remark 1 and Lemma 1 imply that all entries of $\mathbf{Q}$ are strictly positive, indicating that any two states of the original chain $\mathbf{Y}$ are connected. This is a stringent condition; yet, in practice, if some π_{ij} in (9) is zero, we can replace it with a small $\varepsilon > 0$, and then normalize the modified vector $\boldsymbol{\pi}$, thus ensuring that Assumption 1 is satisfied.

Another reason why we set the zero entries in $\boldsymbol{\pi}$ to $\varepsilon > 0$ is for convenience of computing the original transition matrix $\mathbf{Q}$, hence $\mathbf{P}$, via (9) and (3). If we simply eliminate the corresponding states in $\mathbf{Z}$, then it is possible that the number of the remaining states is not the square of some integer N; this would prevent us from easily recovering $\mathbf{P}$ from $\boldsymbol{\pi}$. Consider the following example: Assuming

$$\mathbf{Q} = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0 & 0.2 & 0.8 \\ 0.6 & 0.15 & 0.25 \end{bmatrix},$$

then by (3) we have

$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.15 & 0.25 \\ 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.15 & 0.25 \\ 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.15 & 0.25 \end{bmatrix},$$

and, by direct calculation, we obtain $\boldsymbol{\pi} = (0.03, 0.07, 0.23, 0, 0.05, 0.14, 0.3, 0.07, 0.11)$. Note that only 8 entries in $\boldsymbol{\pi}$ are non-zero and 8 is not the square of some integer N. Thus, if we eliminate the state corresponding to the zero entry in $\boldsymbol{\pi}$, it will be hard to recover $\mathbf{Q}$, hence $\mathbf{P}$.

A. Weak Convergence of Empirical Measure

Let us first establish CLT results for one-dimensional empirical measures

$$U_{n,k} = \sqrt{n}(\Gamma_n(\tilde{\theta}_k) - \tilde{\pi}_k), \quad k = 1, \ldots, N^2. \tag{10}$$

For $k \in \{1, \ldots, N^2\}$ define

$$f_k(Z) = \mathbb{1}\{Z = \tilde{\theta}_k\}. \tag{11}$$

Lemma 2 *Suppose Assumption 1 holds. Then a Central Limit Theorem (CLT) holds for $U_{n,k}$; that is, $U_{n,k} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\ \sigma_k^2)$ with $\sigma_k^2 = \text{Cov}(f_k(Z_1),\ f_k(Z_1)) + 2\sum_{m=1}^{\infty}\text{Cov}(f_k(Z_1),\ f_k(Z_{1+m})) < \infty$.*

Proof: See Appendix B. ∎

Now we state the CLT [16, Thm. 3.1] for the multidimensional empirical measure $\mathbf{U}_n = (U_{n,k};\ k = 1, \ldots, N^2)$ as Lemma 3. Several different proofs for this result are available in [16] and the references therein. For completeness, we provide a proof that leverages the results from [15], in terms of extending Lemma 2.

Lemma 3 *([16]) Suppose Assumption 1 holds. Then a multidimensional CLT holds for $\mathbf{U}_n$; that is,*

$$\mathbf{U}_n \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}), \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

with $\mathbf{\Lambda} = [\Lambda_{ij}]_{i,\,j=1}^{N^2}$ *being an* $N^2 \times N^2$ *covariance matrix given by*

$$\Lambda_{ij} = \tilde{\pi}_i(\mathbf{I}_{ij} - \tilde{\pi}_j) + \sum_{m=1}^{\infty} [\tilde{\pi}_i(\mathbf{P}_{ij}^m - \tilde{\pi}_j) + \tilde{\pi}_j(\mathbf{P}_{ji}^m - \tilde{\pi}_i)], \tag{13}$$

where $\mathbf{I}_{ij}$ *denotes the* $(i,\,j)$*-th entry of the identity matrix, and* $\mathbf{P}_{ij}^m$ *(resp.,* $\mathbf{P}_{ji}^m$*) is the* $(i,\,j)$*-th (resp.,* $(j,\,i)$*-th) entry of the matrix* $\mathbf{P}^m$ *(the* m*-th power of* $\mathbf{P}$*),* $i,\,j = 1, \ldots, N^2$*.*

Proof: See Appendix C. ∎

B. Weak Convergence of Test Statistic

In this section, and to derive weak convergence results for the test statistic $D(\boldsymbol{\nu}\|\boldsymbol{\pi})$, we will leverage a method commonly-used by statisticians in terms of combining a Taylor's series expansion for the test statistic and the CLT result for the empirical measure [11]. Recently, under i.i.d. assumptions, such a weak convergence analysis for certain test statistics has been conducted in [10], [5].

To this end, for $\boldsymbol{\nu} \in \mathcal{P}(\boldsymbol{\Theta})$ we consider

$$h(\boldsymbol{\nu}) = D(\boldsymbol{\nu}\|\boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \nu_{ij} \log \frac{\frac{\nu_{ij}}{\sum_{t=1}^{N}\nu_{it}}}{\frac{\pi_{ij}}{\sum_{t=1}^{N}\pi_{it}}}. \tag{14}$$

Let $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$ with $\mathbf{\Lambda}$ given by (13). Now, we are in a position to derive weak convergence results for our test statistic $D(\boldsymbol{\nu}\|\boldsymbol{\pi})$.

Theorem 1 *Suppose Assumption 1 holds. Then we have the following weak convergence results:*

$$D(\mathbf{\Gamma}_n\|\boldsymbol{\pi}) \xrightarrow[n\to\infty]{d} \frac{1}{2n}\mathbf{U}'\nabla^2 h(\boldsymbol{\pi})\mathbf{U}, \tag{15}$$

$$D(\mathbf{\Gamma}_n\|\boldsymbol{\pi}) \xrightarrow[n\to\infty]{d} \frac{1}{2n}\sum_{k=1}^{N^2}\rho_k\chi_{1k}^2, \tag{16}$$

where $\nabla^2 h(\boldsymbol{\pi})$ *is the Hessian of* $h(\boldsymbol{\nu})$ *evaluated at* $\boldsymbol{\nu} = \boldsymbol{\pi}$*,* $\rho_k,\, k = 1, \ldots, N^2$*, are the eigenvalues of the matrix* $\nabla^2 h(\boldsymbol{\pi})\mathbf{\Lambda}$*, and* $\chi_{1k}^2,\, k = 1, \ldots, N^2$*, are* N^2 *independent* χ^2 *random variables with one degree of freedom.*

Proof: Let us first compute the gradient of $h(\boldsymbol{\nu})$. Expanding the logarithm and after some algebra which leads to cancellations of gradient terms with respect to ν_{ij} in $\sum_{t=1}^{N}\nu_{it}$, for all $i, j = 1, \ldots, N$, we obtain

$$\frac{\partial h(\boldsymbol{\nu})}{\partial \nu_{ij}} = \log\nu_{ij} - \log\left(\sum_{t=1}^{N}\nu_{it}\right) - \log\pi_{ij} + \log\left(\sum_{t=1}^{N}\pi_{it}\right), \tag{17}$$

which implies

$$\nabla h(\boldsymbol{\pi}) = 0. \tag{18}$$

Further, from (17), we compute the Hessian $\nabla^2 h(\boldsymbol{\nu})$ by

$$\frac{\partial^2 h(\boldsymbol{\nu})}{\partial\nu_{ij}\partial\nu_{kl}} = \begin{cases} 0, \text{ if } k \neq i, \\ \dfrac{1}{\nu_{ij}} - \dfrac{1}{\sum_{t=1}^{N}\nu_{it}}, \text{ if } k = i \text{ and } l = j, \\ -\dfrac{1}{\sum_{t=1}^{N}\nu_{it}}, \text{ if } k = i \text{ and } l \neq j. \end{cases} \tag{19}$$

Evaluating all the terms in (19) at $\boldsymbol{\nu} = \boldsymbol{\pi}$ yields $\nabla^2 h(\boldsymbol{\pi})$, which will play a crucial role in approximating $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi})$. It is seen that $\nabla^2 h(\boldsymbol{\nu})$ is continuous in a neighborhood of $\boldsymbol{\pi}$, and we can utilize the second-order Taylor's series expansion of $h(\boldsymbol{\nu})$ centered at $\boldsymbol{\pi}$ to express $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi}) = h(\mathbf{\Gamma}_n) - h(\boldsymbol{\pi})$. Specifically, by (18) and (5) we have

$$\begin{aligned} 2nD\left(\mathbf{\Gamma}_n\|\boldsymbol{\pi}\right) &= 2n\left(h\left(\mathbf{\Gamma}_n\right) - h\left(\boldsymbol{\pi}\right)\right) \\ &= n\left(\mathbf{\Gamma}_n - \boldsymbol{\pi}\right)'\nabla^2 h(\tilde{\mathbf{\Gamma}}_n)\left(\mathbf{\Gamma}_n - \boldsymbol{\pi}\right) \\ &= \mathbf{U}_n'\nabla^2 h(\tilde{\mathbf{\Gamma}}_n)\mathbf{U}_n, \end{aligned} \tag{20}$$

where $\tilde{\mathbf{\Gamma}}_n = \xi_n\mathbf{\Gamma}_n + (1-\xi_n)\,\boldsymbol{\pi}$ is determined with some $\xi_n \in [0, 1]$. From the ergodicity of the chain $\mathbf{Z}$ it follows $\mathbf{\Gamma}_n \xrightarrow[n\to\infty]{w.p.1} \boldsymbol{\pi}$, leading to $\tilde{\mathbf{\Gamma}}_n \xrightarrow[n\to\infty]{w.p.1} \boldsymbol{\pi}$. By the continuity of $\nabla^2 h(\boldsymbol{\nu})$ we obtain

$$\nabla^2 h(\tilde{\mathbf{\Gamma}}_n) \xrightarrow[n\to\infty]{w.p.1} \nabla^2 h(\boldsymbol{\pi}). \tag{21}$$

Applying Slutsky's theorem [17], by (12), (20), and (21) we attain

$$D(\mathbf{\Gamma}_n\|\boldsymbol{\pi}) = \frac{1}{2n}\mathbf{U}_n'\nabla^2 h(\tilde{\mathbf{\Gamma}}_n)\mathbf{U}_n \xrightarrow[n\to\infty]{d} \frac{1}{2n}\mathbf{U}'\nabla^2 h(\boldsymbol{\pi})\mathbf{U}.$$

Finally, by means of a linear transformation [18] on the quadratic form $\mathbf{U}'\nabla^2 h(\boldsymbol{\pi})\mathbf{U}$, we derive the following alternative asymptotic result:

$$D(\mathbf{\Gamma}_n\|\boldsymbol{\pi}) = \frac{1}{2n}\mathbf{U}_n'\nabla^2 h(\tilde{\mathbf{\Gamma}}_n)\mathbf{U}_n \xrightarrow[n\to\infty]{d} \frac{1}{2n}\sum_{k=1}^{N^2}\rho_k\chi_{1k}^2,$$

where $\rho_k,\, k = 1, \ldots, N^2$, are the eigenvalues of the matrix $\nabla^2 h(\boldsymbol{\pi})\mathbf{\Lambda}$, and $\chi_{1k}^2,\, k = 1, \ldots, N^2$, are N^2 independent χ^2 random variables with one degree of freedom. ∎

C. Threshold Approximation

We use an empirical Cumulative Distribution Function (CDF) to approximate the actual CDF of $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi})$. In particular, it is seen from (15) that $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi}) \simeq (1/(2n))\mathbf{U}'\nabla^2 h(\boldsymbol{\pi})\mathbf{U}$ for large n. Thus, to derive an empirical CDF of $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi})$, we can generate a set of Gaussian sample vectors independently according to $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$ and then plug each such sample vector into the right-hand side of (15) (i.e., replace $\mathbf{U}$), thus, obtaining a set of sample scalars, as a reliable proxy for samples of $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi})$.

Once we obtain an empirical CDF of $D(\mathbf{\Gamma}_n\|\boldsymbol{\pi})$, say, denoted $F_{\text{em}}(\cdot; n)$, then, by (7), we can estimate $\eta_{n,\beta}$ as

$$\eta_{n,\beta}^{\text{wc}} \approx F_{\text{em}}^{-1}(1-\beta; n), \tag{22}$$

where $F_{\text{em}}^{-1}(\cdot; n)$ is the inverse of $F_{\text{em}}(\cdot; n)$. Note that the $\eta_{n,\beta}^{\text{wc}}$ derived by (22) depends on the entries of the PL $\boldsymbol{\pi}$. In practice, if $\boldsymbol{\pi}$ is not directly available, we can replace it by the empirical PL evaluated over a long past sample path. For such cases, we summarize the procedures of estimating the threshold based on our weak convergence analysis as Algorithm 1, where $\hat{\boldsymbol{\pi}}$ is a good estimate for $\boldsymbol{\pi}$. We note that the length n_0 of the past sample path should be sufficiently large (e.g., $n_0 \geq 500N^2$) so as to guarantee the validity of taking $\boldsymbol{\pi}$ to be $\hat{\boldsymbol{\pi}}$. In addition, the small positive number ε (e.g., $\varepsilon \leq 10^{-6}$) introduced in Step

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

1 is to avoid division by zero, thus ensuring the numerical stability of the algorithm. If, on the other hand, the actual PL $\boldsymbol{\pi}$ is known, then we can still apply Algorithm 1 by replacing the $\hat{\boldsymbol{\pi}}$ therein with $\boldsymbol{\pi}$.

Similar to (22), we can derive another weak convergence-based threshold estimator $\bar{\eta}^{\text{wc}}_{n,\beta}$ from (16). However, an easy way of calculating $\bar{\eta}^{\text{wc}}_{n,\beta}$ (also summarized in Algorithm 1) still cannot avoid simulations; it is hard to conclude any advantage of $\bar{\eta}^{\text{wc}}_{n,\beta}$ over $\eta^{\text{wc}}_{n,\beta}$. As a matter of fact, calculating the eigenvalues of $\nabla^2 h(\boldsymbol{\pi})\boldsymbol{\Lambda}$ makes the calculation of $\bar{\eta}^{\text{wc}}_{n,\beta}$ numerically not as stable, compared to the calculation of $\eta^{\text{wc}}_{n,\beta}$ via Algorithm 1. Other methods for numerically obtaining $\bar{\eta}^{\text{wc}}_{n,\beta}$ can be found, e.g., in [19] and the references therein. Another fact we should point out is that, in [20, p. 30], a slightly different statistic is considered and therefore an even simpler asymptotic distribution can be derived correspondingly. Moreover, some other papers, e.g., [21], [22], also considered similar but different statistics.

We will illustrate by extensive experiments that our weak convergence analysis can empirically produce more accurate estimation of the threshold than Sanov's theorem for moderate values of n; the price we have to pay, however, is a relatively long but still acceptable computation time.

Remark 3 In Algorithm 1, due to acceptable numerical errors, the originally estimated $\hat{\boldsymbol{\Lambda}}$ (Step 5) could be neither symmetric nor positive semi-definite. Symmetry is imposed by Step 6. Further, to ensure positive semi-definiteness we can diagonalize $\hat{\boldsymbol{\Lambda}}$ as

$$\hat{\boldsymbol{\Lambda}} = \mathbf{O}^{-1}\text{diag}(\lambda_1, \ldots, \lambda_{N^2})\mathbf{O}, \tag{23}$$

where $\mathbf{O}$ is an orthogonal matrix and $\text{diag}(\boldsymbol{\lambda})$ a diagonal matrix with the elements of $\boldsymbol{\lambda}$ in the main diagonal. Due to numerical errors, we might encounter cases where some λ_i are either negative or too small; we can replace them with small positive numbers and recalculate the right-hand side of (23), thus obtaining an updated positive-definite $\hat{\boldsymbol{\Lambda}}$. For implementation details, the reader is referred to [23].

D. A Robust Hoeffding Test

Many actual systems exhibit time-varying behavior. In this section, we extend our methodology to accommodate such systems and use a set of PLs (instead of a single PL $\boldsymbol{\pi}$) to model past system activity.

Let the *null hypothesis* $\mathcal{H}$ be defined as: $\mathbf{Z} = \{Z_l;\ l = 1, 2, \ldots\}$ is drawn according to the set of PLs $\boldsymbol{\Pi} = \{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(L)}\} \subset \mathcal{P}(\boldsymbol{\Theta})$, i.e., $\mathbf{Z}$ is drawn from one of the PLs in $\boldsymbol{\Pi}$ but we do not know from which one. Consider a robust version of the Hoeffding test [4], [5], [24] under Markovian assumptions:

$$\mathcal{H}_n \text{ rejects } \mathcal{H} \text{ if and only if } \inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n\|\boldsymbol{\pi}) > \eta. \tag{24}$$

Essentially, the test selects the most likely PL from $\boldsymbol{\Pi}$ and uses that to make a decision as in (6). Asymptotic Newman-Pearson optimality of this test is shown in [4].

Algorithm 1 Threshold estimation for the ordinary Hoeffding test under Markovian assumptions based on weak convergence analysis.

Input: The sample size n, the target false positive rate β, the alphabet $\boldsymbol{\Theta} = \{\tilde{\theta}_k;\ k = 1, \ldots, N^2\}$, a sample path of the chain $\mathbf{Z}$, denoted $\mathbf{Z}^{(0)} = \{Z_1^{(0)}, \ldots, Z_{n_0}^{(0)}\}$, where n_0 is the length, and the Boolean parameter χ^2_{enab}.

1: Estimate $\tilde{\pi}_k$ by

$$\hat{\tilde{\pi}}_k = \max\left\{\frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{1}\{Z_i^{(0)} = \tilde{\theta}_k\},\ \varepsilon\right\},\ k = 1, \ldots, N^2,$$

where $\varepsilon > 0$ is a small number.

2: Estimate $\boldsymbol{\pi}$ as $\hat{\boldsymbol{\pi}} = (\hat{\tilde{\pi}}_k/\hat{s};\ k = 1, \ldots, N^2)$, where $\hat{s} = \sum_{j=1}^{N^2} \hat{\tilde{\pi}}_j$ is a normalizing constant.

3: Estimate $\nabla^2 h(\boldsymbol{\pi})$ as $\nabla^2 h(\hat{\boldsymbol{\pi}})$, by plugging $\hat{\boldsymbol{\pi}}$ into (19) (i.e., using $\hat{\boldsymbol{\pi}}$ to replace $\boldsymbol{\nu}$).

4: Estimate $\mathbf{P}$ as $\hat{\mathbf{P}}$, via (cf. (3) and Lemma 1)

$$\hat{p}(\theta_{ij}|\theta_{kl}) = \mathbb{1}\{i = l\}\hat{q}_{ij}, \quad k, l, i, j = 1, \ldots, N,$$

where $\hat{q}_{ij} = \hat{\pi}_{ij}/(\sum_{t=1}^{N} \hat{\pi}_{it})$.

5: Estimate $\boldsymbol{\Lambda}$ as $\hat{\boldsymbol{\Lambda}}$, using (by (13) in Lemma 3)

$$\hat{\Lambda}_{ij} = \hat{\tilde{\pi}}_i(\mathbf{I}_{ij} - \hat{\tilde{\pi}}_j) + \sum_{m=1}^{m_0}\left[\hat{\tilde{\pi}}_i(\hat{\mathbf{P}}^m_{ij} - \hat{\tilde{\pi}}_j) + \hat{\tilde{\pi}}_j(\hat{\mathbf{P}}^m_{ji} - \hat{\tilde{\pi}}_i)\right],$$

where m_0 is a sufficiently large integer.

6: Update $\hat{\boldsymbol{\Lambda}}$ by setting $(\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\Lambda}}')/2$ to $\hat{\boldsymbol{\Lambda}}$.

7: **if** $\chi^2_{\text{enab}} = \text{FALSE}$ **then**

8: Generate T Gaussian sample vectors $\hat{\mathbf{U}}^{(t)}$, $t = 1, \ldots, T$, according to $\mathcal{N}(\mathbf{0}, \hat{\boldsymbol{\Lambda}})$.

9: Estimate T samples of $D(\boldsymbol{\Gamma}_n\|\boldsymbol{\pi})$ as $(1/(2n))\hat{\mathbf{U}}^{(t)\prime}\nabla^2 h(\hat{\boldsymbol{\pi}})\hat{\mathbf{U}}^{(t)}$, $t = 1, \ldots, T$ (cf. (15)).

10: Based on the T samples obtained in the last step, estimate an empirical CDF of $D(\boldsymbol{\Gamma}_n\|\boldsymbol{\pi})$, denoted $F_{\text{em}}(\cdot; n)$.

11: Obtain an estimated value for $\eta_{n,\beta}$ by calculating $\eta^{\text{wc}}_{n,\beta}$ via (22).

12: **else if** $\chi^2_{\text{enab}} = \text{TRUE}$ **then**

13: Calculate the eigenvalues $\hat{\rho}_k$, $k = 1, \ldots, N^2$, of the matrix $\nabla^2 h(\hat{\boldsymbol{\pi}})\hat{\boldsymbol{\Lambda}}$.

14: Generate T samples of $(1/(2n))\sum_{k=1}^{N^2} \hat{\rho}_k \chi^2_{1k}$ (cf. (16)).

15: Based on the T samples obtained in the last step, estimate an empirical CDF of $D(\boldsymbol{\Gamma}_n\|\boldsymbol{\pi})$, denoted $\bar{F}_{\text{em}}(\cdot; n)$.

16: Obtain an estimated value for $\eta_{n,\beta}$ by calculating $\bar{\eta}^{\text{wc}}_{n,\beta}$ via (22) with $F_{\text{em}}(\cdot; n)$ replaced by $\bar{F}_{\text{em}}(\cdot; n)$.

17: **end if**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

For $l = 1, \ldots, L$, let $\mathbf{P}^{(l)}$ denote the transition matrix corresponding to $\boldsymbol{\pi}^{(l)}$ and, similar to (2), we write

$$\boldsymbol{\pi}^{(l)} = (\pi_{ij}^{(l)}; i, j = 1, \ldots, N) = (\tilde{\pi}_k^{(l)}; k = 1, \ldots, N^2).$$

Assume $\mathbf{Z}$ is drawn from PL $\boldsymbol{\pi}^{(l)}$ which satisfies Assumption 1. Let $\mathbf{U}_n^{(l)} = \sqrt{n}(\boldsymbol{\Gamma}_n - \boldsymbol{\pi}^{(l)})$. By Lemma 3, we have

$$\mathbf{U}_n^{(l)} \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{(l)}), \tag{25}$$

where $\boldsymbol{\Lambda}^{(l)} = \left[\Lambda_{ij}^{(l)}\right]_{i,\,j=1}^{N^2}$ is given by

$$\begin{aligned}\Lambda_{ij}^{(l)} \;=\; & \tilde{\pi}_i^{(l)}(\mathbf{I}_{ij} - \tilde{\pi}_j^{(l)}) \\ & + \sum_{m=1}^{\infty} [\tilde{\pi}_i^{(l)}(\mathbf{P}_{ij}^{(l)m} - \tilde{\pi}_j^{(l)}) + \tilde{\pi}_j^{(l)}(\mathbf{P}_{ji}^{(l)m} - \tilde{\pi}_i^{(l)})],\end{aligned}$$

with $\mathbf{P}_{ij}^{(l)m}$ being the (i,j)-th entry of the matrix $\mathbf{P}^{(l)m}$ (the m-th power of $\mathbf{P}^{(l)}$). Let $\mathbf{U}^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{(l)})$. Using (15) we obtain

$$D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi}^{(l)}) \simeq \frac{1}{2n} \mathbf{U}^{(l)\prime} \nabla^2 h(\boldsymbol{\pi}^{(l)}) \mathbf{U}^{(l)},$$

which leads to an approximation for the infimum term in (24):

$$\inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi}) \simeq \inf_{l\in\{1,\ldots,L\}} \frac{1}{2n} \mathbf{U}^{(l)\prime} \nabla^2 h(\boldsymbol{\pi}^{(l)}) \mathbf{U}^{(l)}. \tag{26}$$

By the right-hand side of (26), we can generate Gaussian samples to compute a reliable proxy for the samples of $\inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, thereby, obtaining an empirical CDF, denoted $F_{\text{em}}^{\text{rob}}(\cdot; n)$, of $\inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$. Thus, given a target false positive rate β, similar to (22), we can estimate the threshold $\eta_{n,\beta}$ as

$$\eta_{n,\beta}^{\text{wc}} \approx (F_{\text{em}}^{\text{rob}})^{-1}(1 - \beta; n), \tag{27}$$

where $(F_{\text{em}}^{\text{rob}})^{-1}(\cdot; n)$ denotes the inverse of $F_{\text{em}}^{\text{rob}}(\cdot; n)$. Similar to (16), we can also derive a χ^2-type asymptotic approximation to the distribution of $\inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, thus obtaining another weak convergence-based threshold estimator $\bar{\eta}_{n,\beta}^{\text{wc}}$; for economy of space, we omit the details. For the cases where the PLs are not directly available, we summarize the calculation of $\eta_{n,\beta}^{\text{wc}}$ for the robust Hoeffding test as Algorithm 2.

Algorithm 2 Threshold estimation for the robust Hoeffding test under Markovian assumptions based on weak convergence analysis.

Input: The sample size n, the target false positive rate β, the alphabet $\boldsymbol{\Theta} = \{\tilde{\theta}_k;\ k = 1, \ldots, N^2\}$, and a sample path of each PL $\boldsymbol{\pi}^{(l)}$, denoted $\mathbf{Z}^{(l0)} = \{Z_1^{(l0)}, \ldots, Z_{n_0}^{(l0)}\}$, where n_0 is the length, $l = 1, \ldots, L$.

1: **for** $l = 1, \ldots, L$ **do**

2: Estimate $\tilde{\pi}_k^{(l)}$, $k = 1, \ldots, N^2$, by

$$\hat{\tilde{\pi}}_k^{(l)} = \max\left\{\frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{1}\{Z_i^{(l0)} = \tilde{\theta}_k\}, \varepsilon\right\},$$

where $\varepsilon > 0$ is a small number.

3: Estimate $\boldsymbol{\pi}^{(l)}$ as $\hat{\boldsymbol{\pi}}^{(l)} = (\hat{\tilde{\pi}}_k^{(l)}/\hat{s}^{(l)};\ k = 1, \ldots, N^2)$, where $\hat{s}^{(l)} = \sum_{j=1}^{N^2} \hat{\tilde{\pi}}_j^{(l)}$ is normalizing constant.

4: Estimate $\nabla^2 h(\boldsymbol{\pi}^{(l)})$ as $\nabla^2 h(\hat{\boldsymbol{\pi}}^{(l)})$, by plugging $\hat{\boldsymbol{\pi}}^{(l)}$ into (19) (i.e., using $\hat{\boldsymbol{\pi}}^{(l)}$ to replace $\boldsymbol{\nu}$).

5: Estimate $\mathbf{P}^{(l)}$ as $\hat{\mathbf{P}}^{(l)}$, via (cf. (3) and Lemma 1)

$$\hat{p}^{(l)}(\theta_{ij}|\theta_{kl}) = \mathbb{1}\{i = l\}\hat{q}_{ij}^{(l)}, \quad k, l, i, j = 1, \ldots, N,$$

where $\hat{q}_{ij}^{(l)} = \hat{\pi}_{ij}^{(l)}/(\sum_{t=1}^{N} \hat{\pi}_{it}^{(l)})$.

6: Estimate $\boldsymbol{\Lambda}^{(l)}$ as $\hat{\boldsymbol{\Lambda}}^{(l)}$, using (by (13) in Lemma 3)

$$\begin{aligned}\hat{\boldsymbol{\Lambda}}_{ij}^{(l)} =& \hat{\tilde{\pi}}_i^{(l)}(\mathbf{I}_{ij} - \hat{\tilde{\pi}}_j^{(l)}) + \sum_{m=1}^{m_0} \Big[\hat{\tilde{\pi}}_i^{(l)}(\hat{\mathbf{P}}_{ij}^{(l)m} - \hat{\tilde{\pi}}_j^{(l)}) \\ & + \hat{\tilde{\pi}}_j^{(l)}(\hat{\mathbf{P}}_{ji}^{(l)m} - \hat{\tilde{\pi}}_i^{(l)})\Big],\end{aligned}$$

where m_0 is a sufficiently large integer.

7: Update $\hat{\boldsymbol{\Lambda}}^{(l)}$ by setting $(\hat{\boldsymbol{\Lambda}}^{(l)} + \hat{\boldsymbol{\Lambda}}^{(l)\prime})/2$ to $\hat{\boldsymbol{\Lambda}}^{(l)}$.

8: Generate T Gaussian sample vectors $\hat{\mathbf{U}}^{(lt)}$, $t = 1, \ldots, T$, according to $\mathcal{N}(\mathbf{0}, \hat{\boldsymbol{\Lambda}}^{(l)})$.

9: **end for**

10: Estimate T samples of $\inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$ as $\inf_{l\in\{1,\ldots,L\}} (1/2n)\hat{\mathbf{U}}^{(lt)\prime} \nabla^2 h(\boldsymbol{\pi}^{(l)}) \hat{\mathbf{U}}^{(lt)}$, $t = 1, \ldots, T$ (cf. (26)).

11: Based on the T samples obtained in the last step, estimate an empirical CDF of $\inf_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, denoted $F_{\text{em}}^{\text{rob}}(\cdot; n)$.

12: Obtain an estimated value for $\eta_{n,\beta}$ by calculating $\eta_{n,\beta}^{\text{wc}}$ via (27).

V. Experimental Results

In this section, we assess the accuracy of our threshold estimator and the performance of the anomaly detection procedure. We start with a numerical evaluation of the threshold's accuracy and then perform anomaly detection in two application settings using simulated and actual data.

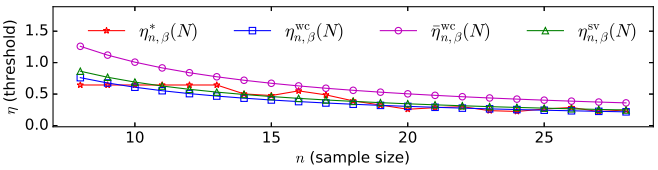
A. Numerical Results for Threshold Approximation

In this subsection, for simplicity we consider the ordinary (and not the robust) Hoeffding test. We have developed a software package TAM [23] to perform the experiments. We will use $\boldsymbol{\Theta} = \{1, 2, \ldots, N^2\}$ to indicate the states and assume the stationary distribution $\boldsymbol{\pi}$ to also be the initial distribution.

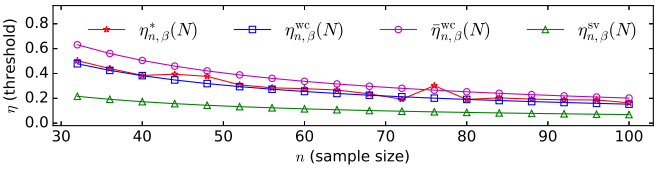
In the following numerical examples, we first randomly create a valid (i.e., such that Assumption 1 holds) $N \times N$ transition matrix $\mathbf{Q}$, giving rise to an $N^2 \times N^2$ transition matrix $\mathbf{P}$, and then generate T test sample paths of the chain $\mathbf{Z}$, each with length n, denoted $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \ldots, Z_n^{(t)}\}$, $t = 1, \ldots, T$. We use these samples to derive empirical CDF's. To simulate the case where the PL $\boldsymbol{\pi}$ is not directly available, we generate one more independent reference sample path $\mathbf{Z}^{(0)} = \{Z_1^{(0)}, \ldots, Z_{n_0}^{(0)}\}$ of length $n_0 \gg |\boldsymbol{\Theta}| = N^2$, thus enabling us to obtain a good estimate of $\boldsymbol{\pi}$. Note that we do not rely on the test sample paths to estimate the PL $\boldsymbol{\pi}$. The ground truth $\boldsymbol{\pi}$ is computed by taking any row of $\mathbf{P}^{m_0}$ for some sufficiently large m_0.

Having the ground truth PL $\boldsymbol{\pi}$ at hand, with the test sample paths $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \ldots, Z_n^{(t)}\}$, $t = 1, \ldots, T$, we can compute T samples of the scalar random variable $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, by (4). Using these samples, we obtain an empirical CDF of $D(\boldsymbol{\Gamma}_n \| \boldsymbol{\pi})$, denoted $F(\cdot; n)$, which can be treated as a

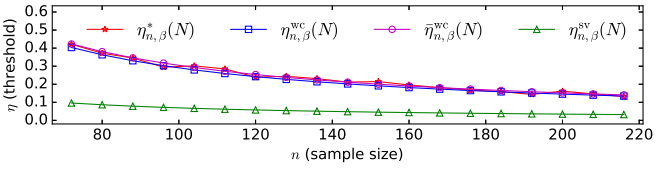
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing



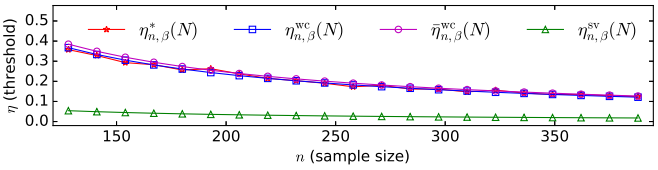
(a) $N = 2$.



(b) $N = 4$.



(c) $N = 6$.



(d) $N = 8$.

Fig. 1: Threshold versus sample size; scenarios corresponding to $\beta = 0.001$, $N = 2, 4, 6, 8$.

dependable proxy of the actual one. The threshold given by (22) with $F_{\rm em}(\cdot; n)$ replaced by $F(\cdot; n)$ is then taken as a reliable proxy of $\eta_{n,\beta}$. We denote this proxy by $\eta^*_{n,\beta}$. To emphasize the dependence on N, we write $\eta_{n,\beta}$ (resp., $\eta^*_{n,\beta}$) as $\eta_{n,\beta}(N)$ (resp., $\eta^*_{n,\beta}(N)$). Next, using the reference sample path $\mathbf{Z}^{(0)}$ and applying Algorithm 1, we obtain $\eta^{\rm wc}_{n,\beta}(N)$ and $\bar{\eta}^{\rm wc}_{n,\beta}(N)$.

Let the target false positive rate be $\beta = 0.001$. Consider four different scenarios where N is 2, 4, 6, and 8, respectively. Set $\varepsilon = 10^{-10}$, $T = 1000$, $m_0 = 1000$, and $n_0 = 1000N^2$. Here we note that, in all our experiments, an estimate $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ with $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\| \leq 10^{-6}$ can be obtained by executing Algorithm 1 with parameters $n_0 \geq 500N^2$ and $\varepsilon \leq 10^{-8}$. In Figures 1a through 1d, the red line plots $\eta^*_{n,\beta}(N)$, the blue line $\eta^{\rm wc}_{n,\beta}(N)$, the magenta line $\bar{\eta}^{\rm wc}_{n,\beta}(N)$, and the green line $\eta^{\rm sv}_{n,\beta}(N)$ (cf. (8)), all as a function of the sample size n. Setting sample sizes n reasonably small (n should at least be comparable to N^2), it can be seen that $\eta^{\rm wc}_{n,\beta}(N)$ and $\bar{\eta}^{\rm wc}_{n,\beta}(N)$ are more accurate than $\eta^{\rm sv}_{n,\beta}$, except for the case $N = 2$ where all estimators perform approximately equally well. In particular, as N increases, the estimation errors of $\eta^{\rm wc}_{n,\beta}(N)$ and $\bar{\eta}^{\rm wc}_{n,\beta}(N)$ are consistently close to zero, while the approximation error of $\eta^{\rm sv}_{n,\beta}$ increases significantly. Moreover, for the scenarios $N = 6, 8$, $\eta^{\rm wc}_{n,\beta}(N)$ and $\bar{\eta}^{\rm wc}_{n,\beta}(N)$ are very close.

Remark 4 In Figures 1a-1d, the red line representing the "actual" value $\eta^*_{n,\beta}$ is not smooth; this is because each time when varying the sample size n, we regenerate all the sample paths $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \ldots, Z_n^{(t)}\}$, $t = 1, \ldots, T$ from scratch. On the other hand, the blue (resp., magenta) line corresponding to $\eta^{\rm wc}_{n,\beta}$ (resp., $\bar{\eta}^{\rm wc}_{n,\beta}$) is smooth because we only need to generate the T Gaussian (resp., χ^2-type) sample vectors once. In our experiments, most of the running time is spent generating the sample paths $\mathbf{Z}^{(t)}$ and calculating $\eta^*_{n,\beta}$ therefrom. In practice, we will neither generate such samples nor calculate $\eta^*_{n,\beta}$, and only need to focus on obtaining $\eta^{\rm wc}_{n,\beta}$ or $\bar{\eta}^{\rm wc}_{n,\beta}$, which is computationally not expensive.

Remark 5 Theoretically speaking, we could use the "actual" threshold $\eta^*_{n,\beta}$ as obtained above, but it is of little practical value; the reason is that in statistical anomaly detection applications, we are typically faced with a long series of observations and want to use a so-called *windowing technique* (see Section V-C), which divides the observations into a sequence of detection windows with the same time length. The sample sizes n in different windows may not necessarily be equal, leading to different threshold settings when sliding the windows. If we use the simulated "actual" threshold, then, when varying the detection windows, we will need to regenerate the corresponding samples (for threshold estimation purposes) from scratch, which is computationally too expensive, especially when there are many detection windows. In contrast, to compute our estimator $\eta^{\rm wc}_{n,\beta}$ (resp., $\bar{\eta}^{\rm wc}_{n,\beta}$), we only need to generate one set of Gaussian (resp., χ^2-type) sample vectors (cf. Remark 4), which can be shared by all the detection windows, thus, saving a lot of computation time. To see this more clearly, let us denote by τ_1 the average running time for generating a set of samples with T ($T = 1000$ is empirically a good choice) Gaussian (resp., χ^2-type) vectors according to (15) (resp., (16)), and τ_2 the average running time for calculating a threshold via (22) given the corresponding sample vectors required to derive the empirical CDF. Clearly, we have $\tau_1 \gg \tau_2 > 0$. Assume we have W detection windows. Then, if we directly simulate the statistic so as to estimate the threshold for each and every detection window, the total running time would be $c_1 W\tau_1 + c_2 W\tau_2 = (c_1\tau_1 + c_2\tau_2)W$, where $c_1, c_2 > 0$ are two scaling constants satisfying $c_1\tau_1 \gg c_2\tau_2$. On the other hand, by simulating Gaussian (resp., χ^2-type) samples, the total running time required to estimate all the thresholds for the W detection windows would be $c_3\tau_1 + c_4\tau_2 W$, where $c_3, c_4 > 0$ are two scaling constants satisfying $c_4 \approx c_2$, leading to $0 < c_4\tau_2 \ll c_1\tau_1 + c_2\tau_2$. Thus, for large W we have $c_3\tau_1 + c_4\tau_2 W \ll (c_1\tau_1 + c_2\tau_2)W$.

To further investigate the performance of different classes of threshold estimators, we now take the randomness of the transition matrix $\mathbf{P}$ into account and define a simulation-based metric $d(\hat{\eta}, \eta^*; n, \beta, N, K)$ to quantify the average squared empirical estimation error, specified as follows:

$$d(\hat{\eta}, \eta^*; n, \beta, N, K) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\eta}^{(k)}_{n,\beta}(N) - \eta^{*(k)}_{n,\beta}(N)\right)^2. \tag{28}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

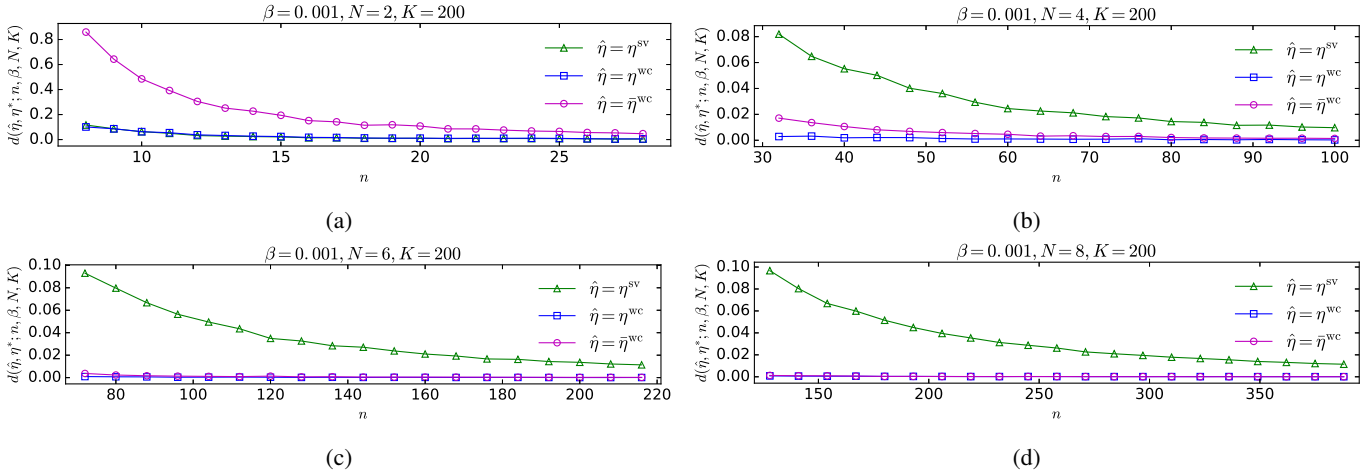


Fig. 2: Evaluation of average squared estimation errors for different types of threshold estimators.

Recall that N is a parameter representing the number of states in the original chain $\mathbf{Y}$. We denote by $\hat{\eta}$ the threshold estimator class (could be η^{sv}, η^{wc}, or $\bar{\eta}^{\text{wc}}$), and by η^* a proxy of the actual threshold class (derived by directly simulating the samples of the test statistic). Denote by K the number of independent repetitions of the calculation for $(\hat{\eta}_{n,\beta}^{(k)}(N) - \eta_{n,\beta}^{*(k)}(N))^2$, where $\hat{\eta}_{n,\beta}^{(k)}(N)$ (resp., $\eta_{n,\beta}^{*(k)}(N)$) denotes the class $\hat{\eta}$ (resp., η^*) instantiated under parameters n, β, N, and $k \in \{1, \ldots, K\}$.

Setting $\beta = 0.001$, $K = 200$, $N \in \{2, 4, 6, 8\}$, and $n \in \left\{\bar{n} = 2N^2 + i \times \left\lfloor 0.2N^2 + 1 \right\rfloor : \bar{n} < 6N^2 + 5, i \in \mathbb{N}_+\right\}$, we evaluate $d(\hat{\eta}, \eta^*; n, \beta, N, K)$. The results are shown in Figure 2. Several observations can be made from Figures 2a-2d: (i) Except for the case $N = 2$, both η^{wc} and $\bar{\eta}^{\text{wc}}$ outperform η^{sv}, that is, $d(\eta^{\text{wc}}, \eta^*; n, \beta, N, K) < d(\eta^{\text{sv}}, \eta^*; n, \beta, N, K)$ and $d(\bar{\eta}^{\text{wc}}, \eta^*; n, \beta, N, K) < d(\eta^{\text{sv}}, \eta^*; n, \beta, N, K)$. (ii) For the cases $N = 6$, 8, η^{wc} and $\bar{\eta}^{\text{wc}}$ perform almost equally well, with both $d(\eta^{\text{wc}}, \eta^*; n, \beta, N, K)$ and $d(\bar{\eta}^{\text{wc}}, \eta^*; n, \beta, N, K)$ being very close to zero and, for the cases $N = 2$, 4, η^{wc} outperforms $\bar{\eta}^{\text{wc}}$, i.e., $d(\eta^{\text{wc}}, \eta^*; n, \beta, N, K) < d(\bar{\eta}^{\text{wc}}, \eta^*; n, \beta, N, K)$. (iii) Only for the case $N = 2$, η^{sv} performs the best among the three estimators and, η^{wc} performs approximately equally well with η^{sv} in this case. More extensive comparison results can be derived using TAM [23]. We may empirically conclude that η^{wc} performs consistently the best among the three for almost all scenarios that we have considered and, on the other hand, η^{sv} performs unsatisfactorily when $N > 2$. Further, $\bar{\eta}^{\text{wc}}$ is numerically not as stable as η^{wc}, especially for the cases where $N \leq 4$.

B. ROC Analysis for the Hoeffding Test with Different Threshold Estimators

In this subsection, for simplicity and economy of space, we again only consider the ordinary (and not the robust) Hoeffding test. We note here that similar results can be derived for the robust Hoeffding test. The numerical experiments are conducted using the software package ROM [25].

Let $\mathbf{\Theta} = \{1, 2, \ldots, N^2\}$ containing N^2 states. For a given sample size n and a given target *False Positive Rate (FPR)* β, the three thresholds $\eta_{n,\beta}^{\text{wc}}$, $\bar{\eta}_{n,\beta}^{\text{wc}}$, and $\eta_{n,\beta}^{\text{sv}}$, respectively, give rise to three different *discrete tests* (denote them by "HTWC-1," "HTWC-2," and "HTSV," respectively). To compare their performances, we will conduct the Receiver Operating Characteristic (ROC) [14] analysis (detection rate vs. false alarm rate) using simulated data.

TABLE I: ROC points vs. target FPR ($N = 4$, $n = 50$).

target FPR β	HTWC-1		HTWC-2		HTSV	
	FPR	TPR	FPR	TPR	FPR	TPR
0.001	0.002	0.885	0.0	0.816	0.402	0.999
0.01	0.011	0.965	0.002	0.888	0.752	1.0
0.02	0.018	0.983	0.003	0.943	0.844	1.0
0.03	0.025	0.99	0.01	0.96	0.898	1.0
0.04	0.038	0.99	0.018	0.971	0.927	1.0
0.05	0.047	0.991	0.029	0.981	0.945	1.0

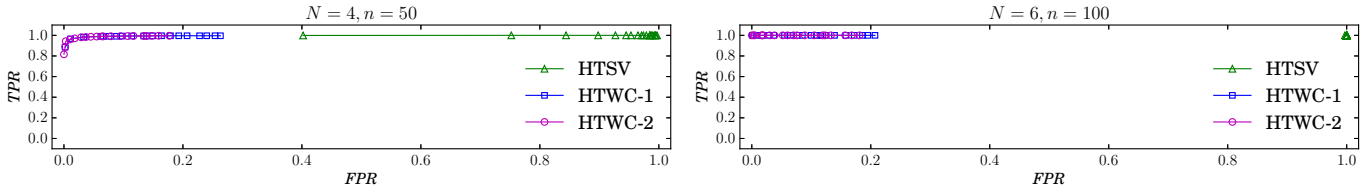
TABLE II: ROC points vs. target FPR ($N = 6$, $n = 100$).

target FPR β	HTWC-1		HTWC-2		HTSV	
	FPR	TPR	FPR	TPR	FPR	TPR
0.001	0.001	1.0	0.0	1.0	0.997	1.0
0.01	0.008	1.0	0.003	1.0	1.0	1.0
0.02	0.017	1.0	0.005	1.0	1.0	1.0
0.03	0.028	1.0	0.017	1.0	1.0	1.0
0.04	0.037	1.0	0.017	1.0	1.0	1.0
0.05	0.055	1.0	0.019	1.0	1.0	1.0

Similar to what we have done in Section V-A, we first randomly create a valid $N \times N$ transition matrix $\mathbf{Q}$, hence an $N^2 \times N^2$ transition matrix $\mathbf{P}$, and then generate T sample paths of the chain $\mathbf{Z}$, each with length n, denoted by $\mathbf{Z}^{(t)} = \{Z_1^{(t)}, \ldots, Z_n^{(t)}\}$, $t = 1, \ldots, T$. From $\mathbf{P}$ we derive the PL $\boldsymbol{\pi}$. Next, to simulate anomalies, we create another valid $N \times N$ transition matrix $\bar{\mathbf{Q}}$, hence an $N^2 \times N^2$ transition matrix $\bar{\mathbf{P}}$, and generate T sample paths of the corresponding chain $\bar{\mathbf{Z}}$, each with length n, denoted by $\bar{\mathbf{Z}}^{(t)} = \{\bar{Z}_1^{(t)}, \ldots, \bar{Z}_n^{(t)}\}$, $t = 1, \ldots, T$. Label each sample path of $\bar{\mathbf{Z}}$ (resp., $\mathbf{Z}$) with length n as "positive" (resp., "negative"). Then, $\{\mathbf{Z}^{(t)} : t \in \{1, \ldots, T\}\} \cup \{\bar{\mathbf{Z}}^{(t)} : t \in \{1, \ldots, T\}\}$ will be our test set, which contains T negative ($\mathbf{Z}^{(t)}$) and T positive ($\bar{\mathbf{Z}}^{(t)}$) sample paths.

Now, by executing Algorithm 1 without estimating $\boldsymbol{\pi}$ (since

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing



(a) (b)

Fig. 3: Results from ROC analysis of the ordinary Hoeffding test.

the ground truth is available), we obtain $\eta^{\text{wc}}_{n,\beta}$ and $\bar{\eta}^{\text{wc}}_{n,\beta}$. Also, by (8) we obtain $\eta^{\text{sv}}_{n,\beta}$. For each sample path in the test set, we compute $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$ by (4). Next, using $\eta^{\text{wc}}_{n,\beta}$ (resp., $\bar{\eta}^{\text{wc}}_{n,\beta}$, $\eta^{\text{sv}}_{n,\beta}$), we can apply HTWC-1 (resp., HTWC-2, HTSV) to detect each sample path as positive or negative. Then, we integrate these reports with the ground truth labels so as to calculate the *True Positive Rate (TPR)* and FPR, thereby, obtaining a point of the ROC space.

In our experiments, we take $T = 1000$. Figure 3a (resp., 3b) shows the ROC graphs of HTWC-1, HTWC-2, and HTSV for a scenario corresponding to $N = 4, n = 50$ (resp., $N = 6$, $n = 100$); different points on the graph are obtained by β taking values from a predesignated finite set $\{0.001\} \cup \{0.01, 0.02, \ldots, 0.19\}$. It is seen from Figure 3a (or Figure 3b) that all TPR values are very close to 1, which is good, but for most cases (each case corresponds to a specific "small" target FPR β) HTWC-1 and HTWC-2 have much closer FPR values to the target FPR value than HTSV, meaning HTWC-1 and HTWC-2 are able to control for false alarms better than HTSV. To see this more clearly, we show a few specific values of the (TPR, FPR) pair in Tables I and II. It is worth noting that in the $N = 6$ scenario, HTSV is almost a random guess for all the target FPR cases that are considered. More extensive experiments show that, as N increases, the performance of HTSV gets worse and worse; in particular, when $N \geq 6$, HTSV is very likely merely a random guess yielding an ROC point close to $(1, 1)$. During our experiments, another observation is that, for each fixed N and β, when n increases, all HTWC-1, HTWC-2, and HTSV perform better and better; this is because with larger sample sizes, all the three estimators $\eta^{\text{wc}}_{n,\beta}$, $\bar{\eta}^{\text{wc}}_{n,\beta}$, and $\eta^{\text{sv}}_{n,\beta}$ approximate the actual $\eta_{n,\beta}$ better. We therefore conclude that HTWC-1 (or HTWC-2) typically outperforms HTSV in the sense that the former has a better capability of controlling the false alarm rate (i.e., FPR) while maintaining a satisfactory detection rate (i.e., TPR).

Remark 6 A natural concern about the ROC analysis above might be the setting of the target FPR (β) values; one may ask: How about always setting β to a "very small" value, say, 10^{-10}, 10^{-100}, or even 10^{-1000}? We have actually already discussed this partly in Section III. Setting a too small β would typically lead to an unsatisfactory detection rate (TPR). In addition, note that $\eta^{\text{wc}}_{n,\beta}$ (or $\bar{\eta}^{\text{wc}}_{n,\beta}$) is numerically obtained from an empirical CDF, say, $G(x)$, of some scalar random variable; we have $G(x)$ nondecreasing, and $\lim_{x \to +\infty} G(x) = 1$, implying that finding an "accurate" x such that $G(x) = 1 - \beta$ would be hard for a too small $\beta \in (0, 1)$. An empirically "good" choice of β is 0.001 (see Tables I and II), which is what we use in our applications. Because HTWC-1 and HTWC-2 perform almost equally well in our experiments, but HTWC-1 is more stable and less computationally demanding, we will only apply HTWC-1 in the following.

C. Simulation Results for Network Anomaly Detection

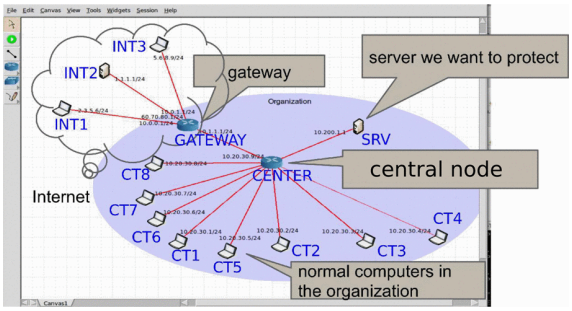


Fig. 4: Simulation setting (from [4]).

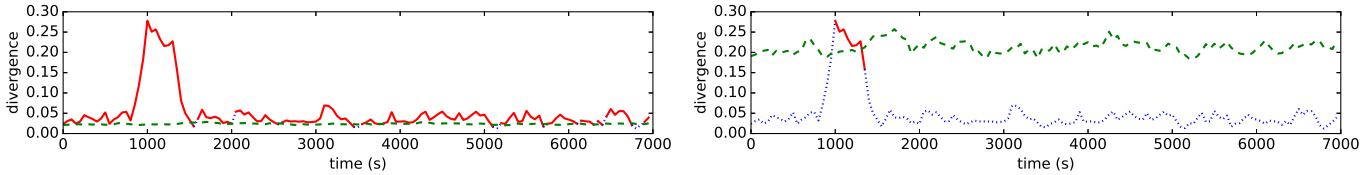
In this subsection we test our approach in a communication network traffic anomaly detection application. We will use the term *traffic* and *flow* interchangeably. We perform the simulations using the software package SADIT [8], which, based on the *fs*-simulator [26], is capable of efficiently generating flow-level network traffic datasets with annotated anomalies.

As shown in Figure 4, the simulated network consists of an internal network involving eight normal users (*CT1*-*CT8*), a server (*SRV*) that stores sensitive information, and three Internet nodes (*INT1*-*INT3*) that connect to the internal network via a gateway (*GATEWAY*).

As in [4, Sec. III.A], to characterize the statistical properties of the flow data, we use as features the flow duration and size (bits). We also cluster the source/destination IP addresses and use as features for each flow the assigned cluster ID and the distance of the flow's IP from the cluster center. For each feature, we quantize its values into discrete symbols so as to obtain a finite alphabet $\boldsymbol{\Xi}$, hence $\boldsymbol{\Theta}$, for our model. Based on the time stamps (the start times) of the flows, we divide the flow data into a series of detection windows, each of which contains a set of flow observations (see [4] for details).

To implement our anomaly detection approach, we first estimate a PL $\boldsymbol{\pi}$ (resp., a PL set $\boldsymbol{\Pi}$) from the stationary (resp., time-varying) normal traffic. Note that, for either case, the reference data should be anomaly-free ideally. However, in our experiments, for the stationary case we use as reference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing



(a) Sanov (b) WC

Fig. 5: Detection results for Scenario V-C-1 with $w_{\mathrm{d}} = 50$ s, $w_{\mathrm{s}} = 200$ s, $k = 2$, $n_1 = 1$, $n_2 = 2$, $n_3 = 2$; (a) threshold is estimated by use of Sanov's theorem; (b) threshold is estimated by use of the weak convergence result.

traffic the entire flow sequence with anomalies injected at some time interval; this makes sense because the size of a typical detection window is much smaller than that of the whole flow sequence and the fraction of anomalies is indeed very small, leading to an estimation for the PL with acceptable accuracy. On the other hand, for the time-varying case we generate the reference traffic without anomalies and the test traffic with anomalies separately, sharing all the parameter settings in the statistical model used in SADIT except the ones for introducing anomalies. Note that, estimating a PL for the stationary traffic is relatively easy, while, for the time-varying traffic, we need to make an effort to estimate several different PLs corresponding to certain periods of the day. We apply the two-step procedure proposed in [4]; that is, we first generate a relatively large PL set and then refine the candidate PLs therein by solving a *weighted set cover problem*. Note also that, if we already know the periodic system activity pattern, then we can directly estimate the PL set period by period; see another anomaly detection application in Section V-D for example.

Now, having the reference PL (resp., PL set) at hand, we persistently monitor the test traffic and report an anomaly instantly as long as the relative entropy $D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$ (resp., $\inf_{\boldsymbol{\pi} \in \mathbf{\Pi}} D(\mathbf{\Gamma}_n \| \boldsymbol{\pi})$) exceeds the threshold $\eta^{\mathrm{wc}}_{n,\beta}$ for the current detection window, where n is the number of flow samples within the window. It is worth pointing out that, for the current application, we will not seek to identify which flows belonging to an abnormal detection window contribute mostly to causing the anomaly, but, in some other applications, e.g., the one in Section V-D, we will do so.

In the following, we consider two scenarios – one for stationary traffic and the other for time-varying traffic.

1) Stationary Network Traffic – Scenario V-C-1: We mimic anomalies caused by a large file download [3, Sec. IV.A.2]. The simulation time is 7000 s. A user increases its mean flow size to 10 times the usual value between 1000 s and 1500 s. The interval between the starting points of two consecutive time windows is taken as $w_{\mathrm{d}} = 50$ s, the window-size is set to $w_{\mathrm{s}} = 200$ s, and the target false positive rate is set to $\beta = 0.001$. The number of user clusters is $k = 2$ and the quantization level for flow duration, flow size, and distance to cluster center is set to $n_1 = 1$, $n_2 = 2$, and $n_3 = 2$, respectively. Thus, the original chain has $N = 2 \times 1 \times 2 \times 2 = 8$ states, and we have $N^2 = 64$ states in the transformed chain.

The detection results are shown in Figures 5a and 5b, both of which depict the relative entropy (divergence) metric defined in (4). The green dashed line in Figure 5a is the threshold estimated using Sanov's theorem (i.e., $\eta^{\mathrm{sv}}_{n,\beta}$ given by (8), where n is the sample size in each specific detection window). The green dashed line in Figure 5b is the threshold given by our estimator (i.e., $\eta^{\mathrm{wc}}_{n,\beta}$ computed by Alg. 1). The interval during which the divergence curve is above the threshold line (the red segment) corresponds to the time instances reported as abnormal. Figure 5a shows that, if $\eta^{\mathrm{sv}}_{n,\beta}$ is used as the threshold, then the Hoeffding test reports too many false alarms, and, Figure 5b shows that, if, instead, we use $\eta^{\mathrm{wc}}_{n,\beta}$ as the threshold, then the Hoeffding test does not report any false alarm while successfully identifying the true anomalies between 1000 s and 1500 s.

2) Time-Varying Network Traffic – Scenario V-C-2: Consider the case where the network in Figure 4 is simulated with a day-night traffic pattern in which the flow size follows a log-normal distribution. We use precisely the same scenario as that in [4, Sec. IV.B.2]. The ground truth anomaly (consider an anomaly where node *CT2* increases its mean flow size by 30%) is injected beginning at 59 h and lasting for 80 minutes.

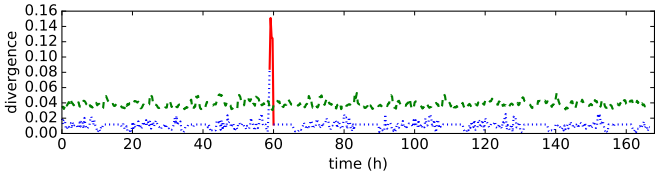


Fig. 6: Detection result for Scenario V-C-2 with $w_{\mathrm{d}} = 1000$ s, $w_{\mathrm{s}} = 1000$ s, $k = 1$, $n_1 = 1$, $n_2 = 4$, $n_3 = 1$.

Using the two-step procedure proposed in [4, Sec. III.C], we first obtain 32 rough PL candidates. Then, using the PL refinement algorithm given in [4, Sec. III.D] equipped with the cross-entropy threshold parameter $\lambda = 0.028$, which is determined by applying Alg. 2, we finally obtain 6 PLs, being active during *morning*, *afternoon*, *evening*, *night*, *dawn*, and *the transition time around sunrise*, respectively. Note that, since we have obtained the PL set in a different way, in the following, when applying Alg. 2 for each detection window, we can skip the first two steps (lines 2 and 3). In the subsequent detection procedure, the chief difference between our method and the one used in [4] is that we no longer set the threshold universally as a constant; instead, we calculate the threshold $\eta^{\mathrm{wc}}_{n,\beta}$ for each detection window using Alg. 2. Set $k = 1$, $n_1 = 1$, $n_2 = 4$, and $n_3 = 1$. Thus, the original chain has $N = 1 \times 1 \times 4 \times 1 = 4$ states, and we have $N^2 = 16$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

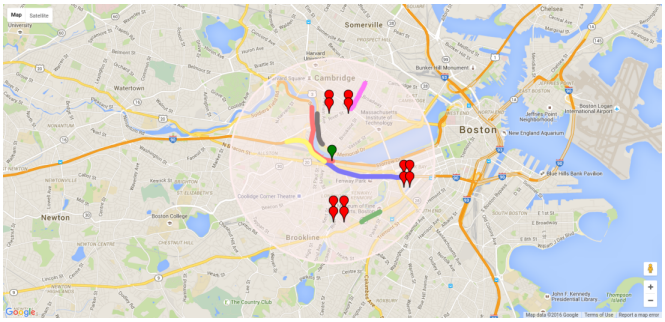


Fig. 7: Location cluster centers and detected abnormal jams for a circle area around Boston University.

states in the transformed chain for this case. Take $w_{\rm d} = 1000$ s, $w_{\rm s} = 1000$ s, and $\beta = 0.001$. We see from Figure 6 that the anomaly is successfully detected, without any false alarms.

D. Anomaly Detection for Waze Jams

1) Dataset Description: The Waze datasets under investigation are kindly provided to us by the Department of Innovation and Technology (DoIT) in the City of Boston. The datasets include three parts: the jam data $\mathcal{J}_1$ (traffic slowdown information generated by Waze based on users' location and speed; note that each jam consists of a set of points), the corresponding point data $\mathcal{J}_2$ (latitudes and longitudes of the points within jams), and the alert data $\mathcal{J}_3$ (traffic incidents reported by active users; we will call such a user a "Wazer"). For each part, we only list the features that we have used in our algorithms. In particular, each entry (jam) in $\mathcal{J}_1$ has the following fields: uuid (unique jam ID), start time, end time, speed (current average speed on jammed segments in meters per second), delay (delay caused by the jam compared to free flow speed, in seconds), and length (jam length in meters). The information for each entry in $\mathcal{J}_2$ includes a jam uuid and the locations (latitudes and longitudes) of the points within the jam. The fields of each entry in $\mathcal{J}_3$ include: uuid (unique system ID; this is different from the jam ID in $\mathcal{J}_1$), location (latitude and longitude per report), type (event type; e.g., accident, weather hazard, road closed, etc.), start time, and end time. It is seen that, by combing $\mathcal{J}_1$ and $\mathcal{J}_2$, we can denote each jam in $\mathcal{J}_1$ as

$$(i, \text{uuid}[i], \text{loc}[i], \text{speed}[i], \text{delay}[i], \text{length}[i], \text{startTime}[i]),$$

where i is the index, uuid is the unique jam ID, "loc" (resp., "startTime") is the abbreviation for location (resp., start time). Because we are only interested in detecting the abnormal jams in real-time, we will not use the jam end times.

2) Anomaly Description: Typically we can observe lots of jams in certain areas during rush hour, e.g., the AM/PM peaks, and most of them are "normal" except those with extremely atypical features (delay, length, etc.). On the other hand, if a jam was observed outside of rush hours or typical areas, then it would likely be "abnormal."

3) Description of the Experiments: Treating Waze jams as a counterpart of the network flows in Section V-C, we implement the robust Hoeffding test on the quantized jam data in the following experiments.

Consider an area around the Boston University (BU) bridge, whose location is specified by latitude and longitude $(42.351848, -71.110730)$ (see the green marker in Figure 7). Extract the jam data no farther than 3 kilometers from BU (within the circle in Figure 7). Note that it is possible for Waze to report several jams at the same time. To assign each jam a unique time stamp, we slightly perturb the start time of the jams that share the same time stamp in the raw data. Such slight adjustments would not alter the original data significantly.

Reference (resp., test) data are taken as jams reported on March 9, 2016 (resp., March 16, 2016). Both dates are Wednesdays, representing typical workdays. There are 3218 jams in the reference data, and 3882 jams in the test data. Note that we have historical data for a relatively long time period (compared to the test data within a detection window); including all the jams reported within the selected reference time period would not hurt the accuracy of the PLs (anomaly-free ideally) to be estimated.

The features that we use for anomaly detection are location, speed, delay, and length. The time stamp of a jam is taken as its start time. To quantize the location, we need to define the distance between two jams. For any valid index i, denote the complete location data of jam i by

$$\widehat{\text{loc}}[i] = \{(x_{i,1}, y_{i,1}), \ldots, (x_{i,i_n}, y_{i,i_n})\}, \tag{29}$$

where x's and y's denote the latitudes and longitudes, respectively, and i_n is the number of points in jam i (typically, i_n is greater than 4). Noting that most of the jams are approximately linear in shape, we simplify (29) by using the 4 vertices of the "smallest" rectangle that covers all the points in the jam and update (29) by

$$\begin{aligned}\text{loc}[i] = \{&(x_{i,\min}, y_{i,\min}), (x_{i,\min}, y_{i,\max}),\\ &(x_{i,\max}, y_{i,\min}), (x_{i,\max}, y_{i,\max})\},\end{aligned} \tag{30}$$

where $x_{i,\min} = \min\{x_{i,1}, \ldots, x_{i,i_n}\}$, $x_{i,\max} = \max\{x_{i,1}, \ldots, x_{i,i_n}\}$, $y_{i,\min} = \min\{y_{i,1}, \ldots, y_{i,i_n}\}$, and $y_{i,\max} = \max\{y_{i,1}, \ldots, y_{i,i_n}\}$. Note that $\text{loc}[i]$ in (30) only contains 4 points. Denote the point-to-point distance (in meters) yielded by Vincenty's formula [27] as $d_V(\cdot, \cdot)$. Then, for any pair of jams, say, indexed i and j, we define the distance between them as

$$\min\{d_V(\mathbf{z}_1, \mathbf{z}_2);\ \forall \mathbf{z}_1 \in \text{loc}[i],\ \mathbf{z}_2 \in \text{loc}[j]\}.$$

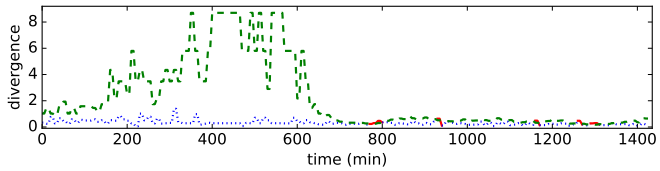
Using the distance defined above and setting the quantization level for "location" as 3, we apply the commonly used K-means clustering method [28], thus obtaining 3 cluster centers as depicted in Figure 7 (note that, by (30) each cluster center is represented by 4 red markers).

In all our experiments, we take the quantization level to be 1 for "speed," and set the target false alarm rate as $\beta = 0.001$. The window size is taken as $w_{\rm s} = 10$ minutes, and the distance between two consecutive windows is $w_{\rm d} = 5$ minutes. To estimate the PLs, we divide a whole day into 4 subintervals: 5:00-10:00 (AM), 10:00-15:00 (MD), 15:00-19:00 (PM), and

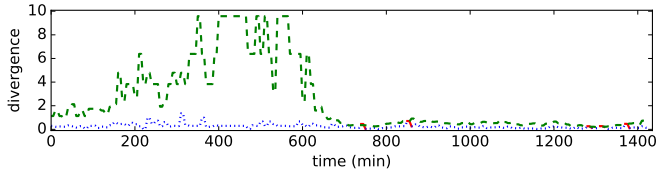
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

TABLE III: Key features of the detected abnormal jams.

index	start time	detected time	latitude	longitude	delay (in seconds)	length (in meters)	alert type
788	12:25:0.302	12:30:0.0	42.361951	-71.117963	232.0	**3568.0**	heavy traffic
1502	15:35:0.072	15:40:0.0	42.356275	-71.119852	**585.0**	844.0	heavy traffic
2412	19:25:0.365	19:30:0.0	42.342549	-71.085011	**643.0**	**3568.0**	heavy traffic
3005	21:25:0.238	21:30:0.0	42.349125	-71.10778	168.0	**1962.0**	weather hazard
3094	21:35:0.267	21:40:0.0	42.373336	-71.097731	**509.0**	897.0	road closed
3126	21:35:0.326	21:40:0.0	42.355048	-71.110335	**528.0**	1293.0	heavy traffic



(a) delay



(b) length

Fig. 8: Initial detection results for Waze jams.

19:00-5:00 (NT). So, for each scenario we end up with 4 PLs, corresponding to the AM peak, the middle day, the PM peak, and the night, respectively. To calculate the threshold $\eta^{\text{wc}}_{n,\beta}$ for each detection window, we use Alg. 2.

4) Detection Results: First, let the quantization level for "delay" be 2 and for "length" be 1. The original sample path has $N = 3{\times}1{\times}2{\times}1 = 6$ states. Thus, we have $N^2 = 36$ states in the transformed chain. We use relatively sparse quantization levels for "delay" and "length" to avoid unnecessary computational overhead in the quantization subroutine for the jam location data. After running our algorithm in the initial step, 910 out of 3882 jams are reported within abnormal detection windows, which correspond to the red segments in Figure 8a. We then perform a refinement procedure by selecting jams in these windows with non-typical individual features as follows. For each selected feature, we calculate the sample mean μ and sample standard deviation σ using the reference data. We then label as anomalous any jam with feature value exceeding $\mu + 3\sigma$. We first consider the delay feature. Using the 3σ-rule on delay, we obtain an anomaly list $\mathcal{L}_1$ containing 4 jams.

Second, let the quantization level for "delay" be 1 and for "length" be 2. Then, again, the original sample path has $N = 3 \times 1 \times 1 \times 2 = 6$ states, and we have $N^2 = 36$ states in the transformed chain. After rerunning the algorithm in the initial step, 590 out of 3882 jams are reported within abnormal detection windows, which correspond to the red segments in Figure 8b, and, after refining by use of the 3σ-rule on the feature "length", we end up with an anomaly list $\mathcal{L}_2$ containing 2 jams.

Finally, we take $\mathcal{L} = \mathcal{L}_1 \cup \mathcal{L}_2$ as our ultimate anomaly list, which contains 6 jams in total. By checking the time stamps and the alarm instances, we see that all of these 6 jams would be reported as abnormal by our method within 5 minutes from their start time; this is satisfactory in a real-time traffic jam anomaly detection application. Note that we can tune w_{d} and w_{s} such that the detection becomes even faster while maintaining good accuracy in identifying anomalies. Specifically, smaller w_{d} leads to faster detection while w_{s} should be reasonably big (the number of jams in a window should at least be comparable to N^2). By comparing the locations and time stamps, we map the jams in the final anomaly list to the alert data $\mathcal{J}_3$, and find that one of them was reported by Wazers as "road closed," another as "weather hazard," and all the others as "jam heavy traffic." In addition, all of them occurred during non-peak hours. We list the key features of these abnormal jams in Table III, where the atypical values of the features "delay" and "length" have been highlighted in bold red. It is worth pointing out that jam 2412 is reported as abnormal based on "delay," but its length (highlighted in bold black) is also above the threshold for refining the detection results based on "length." Note also that the latitude and longitude in each row of Table III represent the closest location of the Wazer who reported the alert for the corresponding jam (extracted from the alert data $\mathcal{J}_3$); the shapes of the actual jams have been visualized as colored bold curves in Figure 7. While in this application we do not have ground truth, it is reassuring that the jams we identify as anomalous have indeed been reported as non-typical by Wazers. Clearly, depending on how such a detection scheme will be used by a City's transportation department, our approach provides flexibility in setting thresholds to adjust the volume of reported anomalous jams. This volume will largely depend on the resources that City personnel have to further investigate anomalous jams (e.g., using cameras) and intervene.

Remark 7 If we directly apply the 3σ-rule on the whole test data without implementing the Hoeffding test to obtain a potential anomaly list first, then we would very likely end up with too many anomalies, which might include undesirable false alarms. Indeed, when we apply the 3σ-rule on the whole test data for "delay" (resp., "length"), we obtain 38 (resp., 62) "anomalies," which are much more than those in our final anomaly list (6 only). Thus, including the well-validated Hoeffding test in our method ensures a good control of false alarms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2017.2771722, IEEE Transactions on Signal Processing

VI. Conclusions and Future Work

We have established weak convergence results for the relative entropy in the Hoeffding test under Markovian assumptions, which enables us to obtain a tighter estimator (compared to the existing estimator based on Sanov's theorem) for the threshold needed by the test. We have demonstrated good performance of our estimator by applying the Hoeffding test in extensive numerical experiments for the purpose of statistical anomaly detection. The application scenarios involve not only simulated communication networks, but also real transportation networks. Our work contributes to enhancing cyber security and helping build smarter cities.

As for future work, it is of interest to establish theoretical comparison results concerning the tightness of the threshold estimators. The challenge in this direction arises from associating the finite sample-size setting with the asymptotic properties of the Central Limit Theorem and the large deviations results (Sanov's theorem). It is also of interest to conduct rigorous analysis relating the computation time of the proposed estimation approach to its accuracy. Also, it is possible to consider additional applications.

Appendix A
Proof of Lemma 1

Expanding the first N entries of $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, we obtain $q_{1i}\sum_{t=1}^{N}\pi_{t1} = \pi_{1i}$, $i = 1,\ldots,N$. Summing up both sides of these equations, it follows

$$\Big(\sum\nolimits_{i=1}^{N} q_{1i}\Big)\Big(\sum\nolimits_{t=1}^{N}\pi_{t1}\Big) = \sum\nolimits_{t=1}^{N}\pi_{1t}. \tag{A.1}$$

Noticing $\sum_{i=1}^{N} q_{1i} = 1$, (A.1) implies $\sum_{t=1}^{N}\pi_{t1} = \sum_{t=1}^{N}\pi_{1t}$, which, together with $q_{11}\sum_{t=1}^{N}\pi_{t1} = \pi_{11}$, yields

$$\frac{\pi_{11}}{\sum_{t=1}^{N}\pi_{1t}} = \frac{\pi_{11}}{\sum_{t=1}^{N}\pi_{t1}} = q_{11}.$$

Similarly, we can show (9) holds for all the other $(i,\, j)$'s.

Appendix B
Proof of Lemma 2

This can be established by applying [15, Corollary 1]. Noting $f_k(\cdot)$ is an indicator function, thus Borel measurable and bounded, and the chain $\mathbf{Z}$ is uniformly ergodic, we see that, $\exists B \in (0,\infty)$ s.t. $|f_k(Z)| \le B$, $\forall Z$, implying that $\mathbb{E}[|f_k(Z)|^3] \le B^3 < \infty$, and [15, (3)] holds with $M(\cdot)$ bounded, leading to $\mathbb{E}[M] < \infty$, and $\gamma(n) = t^n$ for some $t \in (0,1)$, indicating that $\sum_n (\gamma(n))^{1/3} = \sum_n t^{n/3} = \sum_n \left(t^{1/3}\right)^n < \infty$. Thus, all the conditions needed by [15, Corollary 1] are satisfied.

Appendix C
Proof of Lemma 3

We can directly extend Lemma 2 to the multidimensional case (see [29, Chap. 8]). In particular, under Assumption 1, (12) holds with $\boldsymbol{\Lambda}$ given by

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{(0)} + \sum\nolimits_{m=1}^{\infty}\boldsymbol{\Lambda}^{(m)}, \tag{C.1}$$

where $\boldsymbol{\Lambda}^{(0)}$ and $\boldsymbol{\Lambda}^{(m)}$ are specified, respectively, by

$$\begin{aligned}\boldsymbol{\Lambda}^{(0)} =& [\mathrm{Cov}(f_i(Z_1),\ f_j(Z_1))]_{i,\,j=1}^{N^2},\\ \boldsymbol{\Lambda}^{(m)} =& [\mathrm{Cov}(f_i(Z_1), f_j(Z_{1+m}))\\ &+ \mathrm{Cov}(f_j(Z_1), f_i(Z_{1+m}))]_{i,\,j=1}^{N^2}, \quad m = 1,2,\ldots.\end{aligned}$$

Let the subscript ij denote the (i,j) elements of the matrices $\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{(0)}, \boldsymbol{\Lambda}^{(m)}$. By the Markovian properties, after some direct algebra, for $i,\, j = 1,\ldots,N^2$ we obtain $\Lambda_{ij}^{(0)} = \tilde{\pi}_i(\mathbf{I}_{ij} - \tilde{\pi}_j)$ and

$$\Lambda_{ij}^{(m)} = \tilde{\pi}_i(\mathbf{P}_{ij}^{m} - \tilde{\pi}_j) + \tilde{\pi}_j(\mathbf{P}_{ji}^{m} - \tilde{\pi}_i), \quad m = 1,2,\ldots.$$

Acknowledgments

We thank Jing Wang for his contributions and help in developing the software package SADIT [8]. We also thank the DoIT of the City of Boston, Chris Osgood, Alex Chen, and Connor McKay for supplying the Waze data. We thank Athanasios Tsiligkaridis for his help in deriving Table III in Section V-D. We finally thank the anonymous reviewers for useful comments on preliminary versions of this paper.

References

[1] I. C. Paschalidis and G. Smaragdakis, "Spatio-temporal network anomaly detection by assessing deviations of empirical measures," *IEEE/ACM Trans. Networking*, vol. 17, no. 3, pp. 685–697, 2009.

[2] S. Meyn, A. Surana, Y. Lin, and S. Narayanan, "Anomaly detection using projective Markov models in a distributed sensor network," in *Proceedings of the 48th IEEE Conference on Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009.*, Dec 2009, pp. 4662–4669.

[3] J. Wang, D. Rossell, C. G. Cassandras, and I. C. Paschalidis, "Network anomaly detection: A survey and comparative analysis of stochastic and deterministic methods," in *Proceedings of the 52nd IEEE Conference on Decision and Control*, Florence, Italy, December 2013, pp. 182–187.

[4] J. Wang and I. C. Paschalidis, "Statistical traffic anomaly detection in time-varying communication networks," *IEEE Trans. Control of Network Sys.*, vol. 2, no. 2, pp. 100–111, 2015.

[5] J. Unnikrishnan and D. Huang, "Weak convergence analysis of asymptotically optimal hypothesis tests," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4285–4299, 2016.

[6] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, pp. 369–401, 1965.

[7] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Springer, 1998.

[8] J. Wang, J. Zhang, and I. C. Paschalidis, "Statitical Anomaly Detector of Internet Traffic (SADIT)," https://github.com/hbhzwj/SADIT, 2014.

[9] J. Zhang and I. C. Paschalidis, "An improved composite hypothesis test for Markov models with applications in network anomaly detection," in *Proceedings of the 54th IEEE Conference on Decision and Control*, Osaka, Japan, December 2015, pp. 3810–3815.

[10] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1587–1603, 2011.

[11] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[12] M. Iltis, "Sharp asymptotics of large deviations in $\mathbb{R}^d$," *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, 1995.

[13] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[14] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[15] G. L. Jones, "On the Markov chain central limit theorem," *Probability Surveys*, vol. 1, pp. 299–320, 2004.

[16] P. Billingsley, "Statistical methods in Markov chains," *The Annals of Mathematical Statistics*, pp. 12–40, 1961.

[17] ——, *Convergence of probability measures*. John Wiley & Sons, 2013.

[18] J. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, no. 3/4, pp. 419–426, 1961.
[19] H. Liu, Y. Tang, and H. H. Zhang, "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables," *Computational Statistics & Data Analysis*, vol. 53, no. 4, pp. 853–856, 2009.
[20] P. Billingsley, *Statistical inference for Markov processes*. University of Chicago Press, 1961.
[21] M. L. Menéndez, D. Morales, L. Pardo, and I. Vajda, "Testing in stationary models based on divergences of observed and theoretical frequencies," *Kybernetika*, vol. 33, no. 5, pp. 465–475, 1997.
[22] ——, "Inference about stationary distributions of Markov chains based on divergences with observed frequencies," *Kybernetika*, vol. 35, no. 3, pp. 265–280, 1999.
[23] J. Zhang, "Threshold Approximation for Hoeffding's Test under Markovian Assumption (TAM)," *https://github.com/jingzbu/TAHTMA*, 2015.
[24] C. Pandit and S. Meyn, "Worst-case large-deviation asymptotics with application to queueing and information theory," *Stochastic processes and their applications*, vol. 116, no. 5, pp. 724–756, 2006.
[25] J. Zhang, "ROC analysis for Hoeffding test under Markovian assumptions (ROM)," *https://github.com/jingzbu/ROCHM*, 2017.
[26] J. Sommers, R. Bowden, B. Eriksson, P. Barford, M. Roughan, and N. Duffield, "Efficient network-wide flow record generation," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 2363–2371.
[27] Wikipedia, "Vincenty's formulae," *https://en.wikipedia.org/wiki/Vincenty%27s_formulae*.
[28] ——, "k-means clustering," *https://en.wikipedia.org/wiki/K-means_clustering*.
[29] C. Geyer, "Stat 5101 (Geyer) Course Notes," *http://www.stat.umn.edu/geyer/5101/notes/n2.pdf*, 2001.

Jing Zhang obtained his M.S. degree in complex systems and control from the Institute of Systems Science, Chinese Academy of Sciences, Beijing, China, in June 2013. Since September 2013 he has been with the Division of Systems Engineering, Boston University, Boston, MA, USA, where he is pursuing the Ph.D. degree in systems engineering. His research interests include optimization, statistics, and machine learning, with applications in communication networks and transportation systems.

Mr. Zhang is a recipient of the Boston Area Research Initiative (BARI) Research Seed Grant Award (Spring 2017). He placed second in the 2017 Net Impact & Toyota Next Generation Mobility Challenge. He is currently a reviewer for several journals, including IEEE Transactions on Automatic Control, IEEE/ACM Transactions on Networking, and IEEE Transactions on Automation Science and Engineering.

Ioannis Ch. Paschalidis (M'96–SM'06–F'14) received the M.S. and Ph.D. degrees both in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1993 and 1996, respectively. In September 1996 he joined Boston University where he has been ever since. He is a Professor at Boston University with appointments in the Department of Electrical and Computer Engineering, the Division of Systems Engineering, and the Department of Biomedical Engineering. He is the Director of the Center for Information and Systems Engineering (CISE). He has held visiting appointments with MIT and Columbia University, New York, NY, USA. His current research interests lie in the fields of systems and control, networking, applied probability, optimization, operations research, computational biology, and medical informatics.

Dr. Paschalidis is a recipient of the NSF CAREER award (2000), several best paper and best algorithmic performance awards, and a 2014 IBM/IEEE Smarter Planet Challenge Award. He was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the U.S. National Academy of Engineering and the 2014 U.S. National Academies Keck Futures Initiative (NAFKI) Conference. He is the inaugural Editor-in-Chief of the IEEE Transactions on Control of Network Systems.